Software Manual
A mathematical model for universal semantics

Weinan E[1,2*], Yajun Zhou[2*]

[1]Department of Mathematics & Program in Applied and Computational Mathematics,
Princeton University, Princeton, NJ 08544, USA
[2]Beijing Institute of Big Data Research, Beijing 100871, P. R. China

[*]Corresponding authors. E-mail: weinan@math.princeton.edu (W.E), yajun.zhou.1982@pku.edu.cn (Y.Z.)

March 2020

Contents

This Software Manual accompanies our article entitled "A mathematical model for universal semantics" [1], supplying linguistic and algorithmic details. Mathematical proofs behind our model are elaborated in [1, Appendices A and B], so they will not be repeated here.

In Part I, we describe the scope and limitations of our current computational methods for automated word translation and text comprehension. Numerical tests on corpora are also briefly summarized in tables and figures.

In Part II, we define some terminologies and notations in linguistics, in the context of our current software implementation. We also list all the procedures required to cleanse our text corpora.

In Part III, we present detailed algorithms for treating 14 languages out of 5 representative language families, including 12 European languages (Danish, German, Dutch, Spanish, French, Latin, Polish and Russian from the Indo–European language family; Finnish and Hungarian from the Uralic language family; Basque from the Vasconic language family) and 2 Asian languages (Korean from the Koreanic language family; Turkish from the Turkic language family). We illustrate these algorithms with automated word translations from English masterpieces to other languages, and with question answering on the WikiQA dataset.

Part I
Scope and Limitations

1 What our software does

1.1 Markov analysis of word patterns and topic extraction

As in [1, §1], we are interested in word patterns, each of which is a collection of morphologically related content words. We refer our readers to Part III for the working definition of content words (which varies from language to language), and how we recognize and group content words that have related morphologies.

In our numerical studies of text documents, we count waiting times between two word patterns W_i and W_j by the *effective fragment length* L_{ij} (defined in [1, Fig. 1]). As we have mentioned in the caption to [1, Fig. 1], a text fragment qualifies as an *long-range transition* from W_i and W_j, if

- it is flanked by a word in W_i to the left and a word in W_j to the right;
- it does not mention any word in the set W_i;
- it is longer than the longest word in the set $\mathsf{W}_i \cup \mathsf{W}_j$.

We always ignore close encounters of words (that is, those failing the last of the three criteria above), while collecting data for effective fragment lengths over long-range transitions.

With n_{ij} samples of effective fragment lengths L_{ij} in long-range transitions from W_i to W_j,[1] we model a text by a Markov matrix $\mathbf{P} = (p_{ij})_{1\le i,j\le N}$, whose entries are given by [1, (5)]

$$p_{ij} := \frac{n_{ij} e^{-\langle \log L_{ij}\rangle}}{\sum_{k=1}^{N} n_{ik} e^{-\langle \log L_{ik}\rangle}}. \tag{1.1}$$

Numerical evidence suggests that this empirical Markov matrix $\mathbf{P} = (p_{ij})_{1\le i,j\le N}$ is a fair approximation to an *ergodic* matrix $\mathbf{P}^* = (p^*_{ij})_{1\le i,j\le N}$, whose equilibrium state $\boldsymbol{\pi}^* = (\pi^*_1, \ldots, \pi^*_N)$ matches word count statistics [1, Fig. 4a], and honors the *detailed balance* condition $\pi^*_i p^*_{ij} = \pi^*_j p^*_{ji}$ [1, Fig. 4b].

For an ergodic Markov chain satisfying the detailed balance condition, one can show that the probability distribution of recurrence times $\tau = L_{ii}$ goes like [1, Appendix B.3]

$$\mathbb{P}(\tau > t) \sim \sum_m c_m e^{-k_m t}, \quad \left(\text{where } c_m, k_m > 0, \text{ and } \sum_m c_m = 1\right), \tag{1.2}$$

a functional form that frequently crops up in dynamic studies of biological macromolecules [7]. This weighted mixture of multiple exponential decay laws imposes an inequality constraint on the recurrence time τ [1, (3)]:

$$\langle \log \tau\rangle - \log\langle \tau\rangle + \gamma_0 = \sum_m c_m \log\frac{1}{k_m} - \log\sum_m \frac{c_m}{k_m} \le 0, \tag{1.3}$$

[1] It is worth noting that we construct an empirical Markov matrix by an *in situ* analysis of a text, without digesting a document (or small parts of it) as a scrambled bag of words, a procedure implemented in conventional algorithms [2, 3, 4, 5, 6]. The notation p_{ij} in the semantic model of Turney–Pantel [4] is unrelated to ours.

where $\gamma_0 := \lim_{n\to\infty}\left(-\log n + \sum_{m=1}^{n}\frac{1}{m}\right)$ is the Euler–Mascheroni constant. This inequality explains the systematic data trend in [1, Fig. 2e], where most data points reside on one side of the line of Poissonian banality $\langle\log\tau\rangle - \log\langle\tau\rangle + \gamma_0 = 0$ (blue line in [1, Fig. 2e]).

In [1, Appendix A], we have shown that the statistic

$$\delta_i := \log\langle L_{ii}\rangle - \langle\log L_{ii}\rangle - \gamma_0 + \frac{1}{2n_{ii}} \tag{1.4}$$

satisfies

$$|\delta_i| < \frac{2}{\sqrt{n_{ii}}}\sqrt{\frac{\pi^2}{6} - 1 - \frac{1}{2n_{ii}}} \tag{1.5}$$

with probability 95%, when there are n_{ii} independent samples of exponentially distributed (hence banal) L_{ii}. In our current work, we automatically extract topics (Figs. S3–S16 in Part III), by identifying all the word patterns that violate the inequality above. Being consistent with the weighted exponential mixture model in (1.2), almost all of our automatically identified topics in Figs. S3–S16 satisfy

$$\delta_i \geq \frac{2}{\sqrt{n_{ii}}}\sqrt{\frac{\pi^2}{6} - 1 - \frac{1}{2n_{ii}}} > 0. \tag{1.6}$$

If a concept (representable by a word pattern) qualifies as a topical (resp. non-topical) pattern in a certain document, according to the algorithm above, then this concept usually remains topical (resp. non-topical) if we examine the long-range statistic δ_i in a parallel document written in a different language [1, Fig. 3b,b′]. In other words, topicality/non-topicality is an invariant (*i.e.* a language-independent property) under translations.

Generalizing this further, we hypothesize that temporal structures are nearly universal across different languages, on all the scales longer than next-to-nearest neighboring words. In fact, later in §1.2, we will demonstrate that the statistics of effective fragment lengths L_{ij} almost represent a complete set of semantic invariants, quantifying the semantic similarity between W_i and W_j in a language-independent way.

1.2 Semantic fingerprints from invariant Markov spectra

What gives meanings to things? If we threw this ontological question to paleolithic cavemen or preschool children, we would probably expect tangible answers that define objects by physical and physiological stimuli. Putatively speaking, cavemen would conceptualize "fire" by associating it to "brightness" (visual impression), "warmth" (feeling at a distance) and "pain" (reaction upon touch); children would recognize an "apple" as something that is "round" (in shape), "red" (in color) and "sweet" (in taste). While this reductionist view of semantics does not generalize well to abstract concepts ("love", "pride", ...), fictional characters ("Elizabeth Bennet", "Harry Potter", ...) or non-existent places ("Meryton", "Hogwarts", ...), we still find it an attractive idea to generate semantic tags by association—perhaps a universal way to clarify the meaning of a word is to link it to a small set of other words (even if they do not represent bodily stimuli).[2]

Such an ontological picture motivates us to define the meaning of words in terms of a language-independent numerical fingerprint, using Markov matrices that characterize association and connectivity. As we will soon see, certain invariance properties of Markov matrices account for the translatability of concepts across a wide variety of languages. Put differently, in our Markov language model, the semantic context of an individual word pattern is (up to numerical errors) independent of the language in which it is expressed—the mental states underlying a text thus remain (approximately) invariant under translation.

Through numerical experiments ([1, Fig. 2e] and Fig. S1a,a′,a″), we find that the spectrum of a Markov matrix (collection of eigenvalues, counting multiplicity) is language-independent.

In [1, §2.2.3], we have designed a thought experiment (estimation of brainstorming rates from words in language A to words in language B) involving two monolingual human subjects, which has led us to an equation [1, (6)]

$$\mathbf{P}_{\mathrm{A}}\mathbf{T}_{\mathrm{A}\to\mathrm{B}} = \mathbf{T}_{\mathrm{A}\to\mathrm{B}}\mathbf{P}_{\mathrm{B}}. \tag{1.7}$$

Here, $\mathbf{P}_{\mathrm{A}}$ and $\mathbf{P}_{\mathrm{B}}$ are, respectively, the Markov matrices characterizing the mental language of Alice (knowing only language A) and Bob (knowing only language B), while $\mathbf{T}_{\mathrm{A}\to\mathrm{B}}$ is the common dictionary matrix shared by both people. If we have an invertible Markov matrix $\mathbf{T}_{\mathrm{A}\to\mathrm{B}}$ (when translations are lossless), then $\mathbf{P}_{\mathrm{A}}$ and $\mathbf{P}_{\mathrm{B}}$ share the same characteristic polynomial $\det(\lambda\mathbf{I} - \mathbf{P}_{\mathrm{A}}) = \det(\lambda\mathbf{I} - \mathbf{P}_{\mathrm{B}})$, hence the same spectrum. We note that the dictionary matrix $\mathbf{T}_{\mathrm{A}\to\mathrm{B}}$ here does not need to be

[2] It is worth mentioning that Jane Austen has written several novels, besides *Pride and Prejudice*, with themes of love and marriage, despite her very lack of personal experience in these matters—she was able to create vivid characters out of her pure imagination, while tagging them with appropriate contexts.

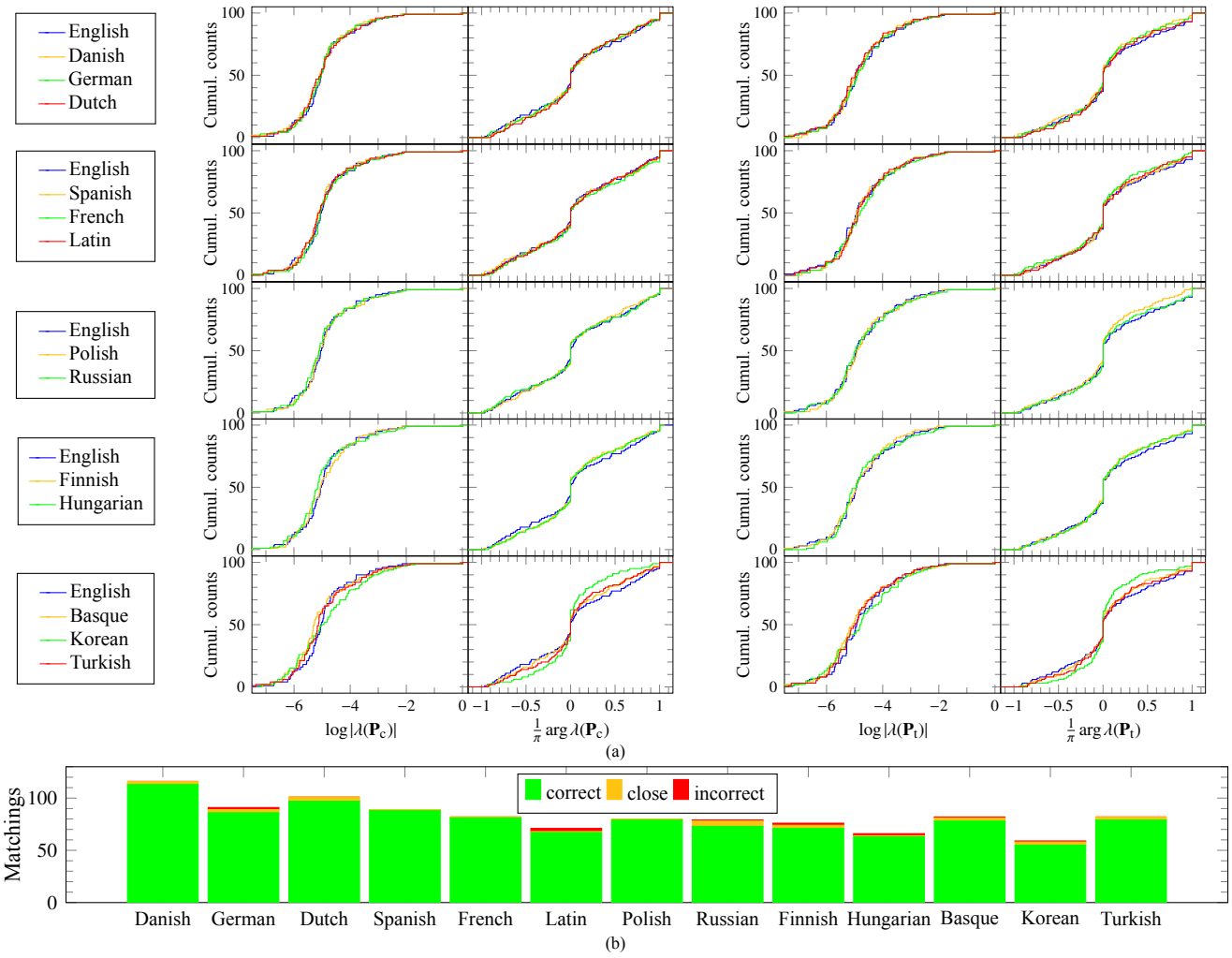


Fig. S1. Language-independence of Markov spectra. (a) Approximate spectral invariance in the Markov matrices $\mathbf{P}_c$ (resp. $\mathbf{P}_t$) for top 100 content (resp. topic) word clusters from the English original and 13 different translations of *Pride and Prejudice*. (See §9.2 for difficulties in analyzing Korean texts.) (b) Evaluations of automated semantic alignments ([1, Fig. 6] and Figs. S4–S7, S9–S16) between topics in the English version of *Pride and Prejudice* and those in the 13 different translations.

a permutation matrix[3] to qualify for spectral invariance. An invertible $\mathbf{T}_{A\to B}$ may assume non-permutation forms (which partially accommodates to polysemy of words in realistic languages), such as $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$. Furthermore, the inverse matrix $\mathbf{T}_{A\to B}^{-1}$ does not necessarily coincide with the dictionary matrix $\mathbf{T}_{B\to A}$. This is because $\mathbf{P}_A(\mathbf{A}\mathbf{T}_{A\to B}) = (\mathbf{A}\mathbf{T}_{A\to B})\mathbf{P}_B$ and $\mathbf{P}_A(\mathbf{T}_{A\to B}\mathbf{B}) = (\mathbf{T}_{A\to B}\mathbf{B})\mathbf{P}_B$ hold for any matrices $\mathbf{A}$ and $\mathbf{B}$ satisfying $\mathbf{A}\mathbf{P}_A = \mathbf{P}_A\mathbf{A}$ and $\mathbf{B}\mathbf{P}_B = \mathbf{P}_B\mathbf{B}$.

In [1, Fig. 4c], we have demonstrated approximate spectral invariance, by localizing Markov matrices on the top 100 content word patterns, in the English, French, Russian and Finnish versions of *Pride and Prejudice*. Such invariance properties remain unscathed when we consider translations of the same novel into other languages, and when we localize on the top 100 topical patterns instead (Fig. S1a). The numerical agreement is understandable from the language-independence of semantic content. In fact, we have pushed this spectral invariance down to an even more local scale in [1, Fig. 5b]: after translations, there are only small perturbations in the dominant eigenvalues for Markov matrices localized on *semantic cliques* surrounding specific concepts.

Here, to define a semantic clique $\mathscr{S}_i$ surrounding a specific word pattern W_i, we need to specify numerical semantic fields through an affinity score $\alpha_{ij}(\ell)$. If we have a Markov process on a semantic web that honors detailed balance, then for each fixed word pattern W_i, we can determine [8] the distribution of hitting times $L_{ij}, 1 \le j \le N$ from that of the return times L_{ii}. Further assuming that the distribution of $\langle \log L_{ij} \rangle$ is nearly Gaussian, we can use the following approximation [1, (7)]:

$$\mathbb{P}(\langle \log L_{ij} \rangle > \ell) \approx \alpha_{ij}(\ell) := \sqrt{\frac{n_{ij}}{2\pi\beta_i}} \int_\ell^\infty e^{-\frac{n_{ij}(x-\ell_i)^2}{2\beta_i}} \, \mathrm{d}x, \tag{1.8}$$

[3] A permutation matrix is a Markov matrix whose entries are either 0 or 1. When the dictionary matrix $\mathbf{T}_{A\to B}$ is a permutation matrix, translation becomes a bijection of words between two languages. This picture, of course, is much too restrictive.

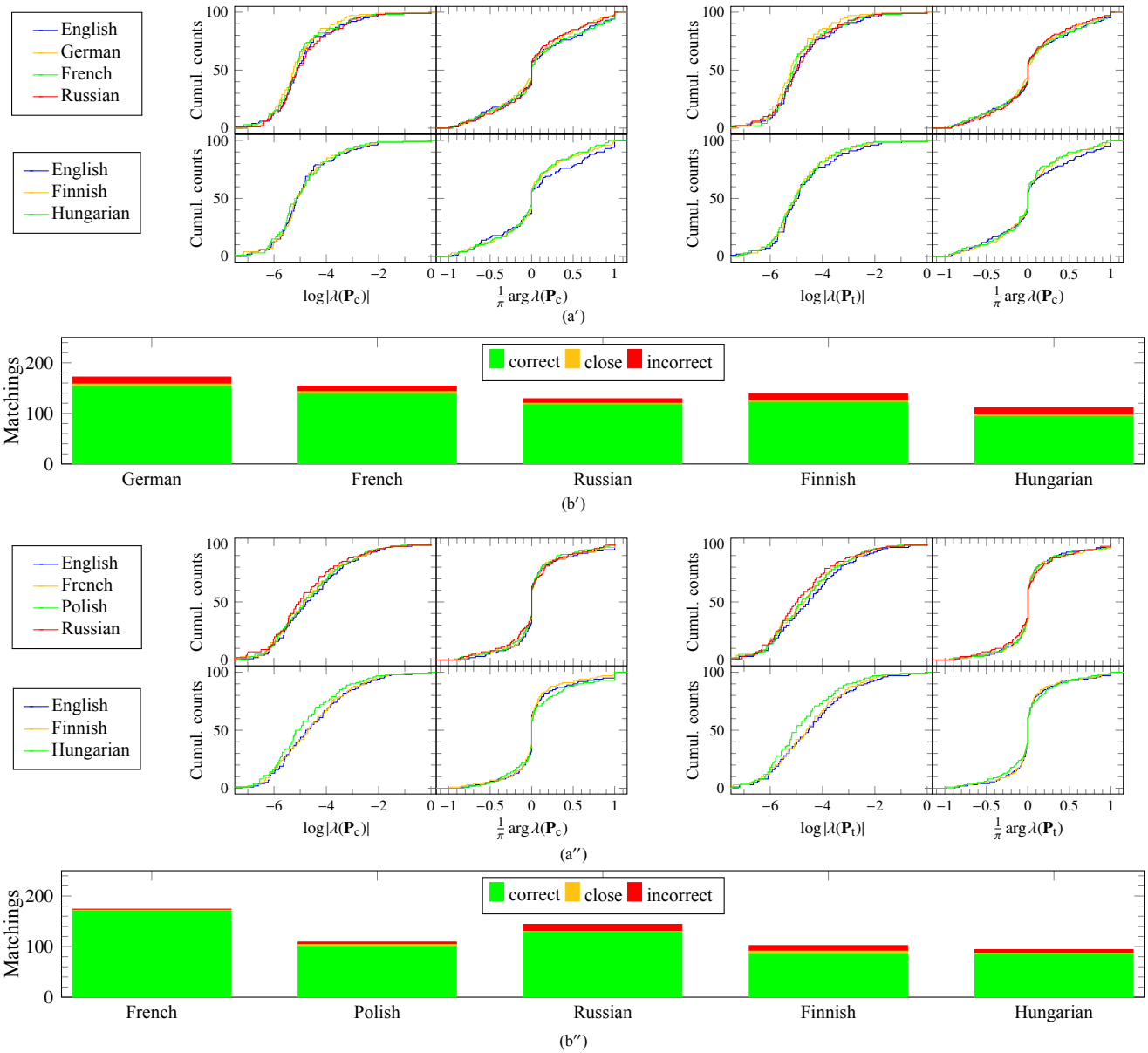


Fig. S1. Language-independence of Markov spectra. (*Continued*) Similar service on multiple versions of (a′)–(b′) *Jane Eyre* and (a″)–(b″) *Origin of Species*.

where [1, (8) and Theorem 4 in Appendix B.4]

$$\ell_i := \frac{\langle L_{ii} \log L_{ii} \rangle}{\langle L_{ii} \rangle} - 1, \quad \beta_i := \frac{\langle L_{ii} (\ell_i - \log L_{ii})^2 \rangle}{\langle L_{ii} \rangle}. \tag{1.9}$$

Our semantic clique $\mathscr{S}_i$ [1, Fig. 5a,b] admits W_j as its member if and only if

$$\min\{\alpha_{ij}(\langle \log L_{ij} \rangle), \alpha_{ji}(\langle \log L_{ji} \rangle)\} > \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1} e^{-\frac{x^2}{2}} \mathrm{d}\, x. \tag{1.10}$$

Here, we give some caveats on the spectral invariance for localized Markov models of semantic cliques, and further clarify the meaning of "dominant eigenvalues". The criterion (1.10) for membership in semantic cliques and the 95% confidence interval for topicality [see (1.5)] both involve arbitrary cut-offs. Whenever there are close calls, our algorithm may include certain topics in a particular semantic clique $\mathscr{S}_i^{\mathrm{A}}$ in language A, but exclude them from a parallel clique $\mathscr{S}_i^{\mathrm{B}}$ in language B. Thus, the size of semantic cliques might not be stable across languages (see dashed lines in [1, Fig. 5b]), due to the presence of marginal topics. To combat this instability, we have proposed to use the entropy production rate[4] to compress dimensions (see

[4]Heuristically speaking, the entropy production rate $\eta(\mathbf{P}) = -\sum_{i,j} \pi_i p_{ij} \log p_{ij}$ [9, (4.27)] of a Markov matrix $\mathbf{P}$ represents the weighted average (assigning

solid lines in [1, Fig. 5b]). For an $N\times N$ Markov matrix $\mathbf{P}$ with strictly positive entries, the entropy production rate $\eta(\mathbf{P})$ satisfies a sharp inequality[5] $\eta(\mathbf{P}) \le \log N$ [10, Theorem 14.1], so $\lfloor e^{\eta(\mathbf{P})} \rfloor$ represents the effective dimension of our localized Markov matrix $\mathbf{P}$ for a semantic clique. This effective dimension is less susceptible to marginal topics than the number of members in a numerically constructed semantic clique (= the size of matrix $\mathbf{P}$), hence we use it to decide how many dominant eigenvalues to keep before vectorizing a topic from the recurrence eigenvalues [1, Fig. 5b] in its surrounding semantic clique. The vector that arranges these recurrence eigenvalues (corresponding to decay modes in (1.2)) in descending order is a *language-independent semantic fingerprint* for a topic.

1.3 Automated word translation based on semantic fingerprints

Our machine translation experiments solve bipartite matching problems involving the Ružička similarities [11, 12]

$$s_R(\mathbf{a},\mathbf{b}) := \frac{\|\mathbf{a}\wedge\mathbf{b}\|_1}{\|\mathbf{a}\vee\mathbf{b}\|_1} \equiv \frac{\sum_i \min(a_i,b_i)}{\sum_i \max(a_i,b_i)} \tag{1.11}$$

between vectorized topics.[6] The Ružička dissimilarity function $d_R(\mathbf{a},\mathbf{b}) = 1 - s_R(\mathbf{a},\mathbf{b})$ measures the distance between two non-zero vectors with non-negative components, and satisfies the three axioms for a distance metric:

1. (Discernibility) $d_R(\mathbf{a},\mathbf{b}) = 0 \Longleftrightarrow \mathbf{a} = \mathbf{b}$;
2. (Symmetry) $d_R(\mathbf{a},\mathbf{b}) = d_R(\mathbf{b},\mathbf{a})$;
3. (Subadditivity) $d_R(\mathbf{a},\mathbf{c}) \le d_R(\mathbf{a},\mathbf{b}) + d_R(\mathbf{b},\mathbf{c})$.

To verify subadditivity, one can use Gilbert's bounds [13]

$$1 - \frac{\|\mathbf{v}_1\wedge\mathbf{v}_2\wedge\mathbf{v}_3\|_1}{\|\mathbf{v}_1\vee\mathbf{v}_2\vee\mathbf{v}_3\|_1} \ge d_R(\mathbf{v}_i,\mathbf{v}_j) = 1 - \frac{\|\mathbf{v}_i\wedge\mathbf{v}_j\|_1}{\|\mathbf{v}_i\vee\mathbf{v}_j\|_1} \ge \frac{\|\mathbf{v}_i\vee\mathbf{v}_j\|_1 - \|\mathbf{v}_i\wedge\mathbf{v}_j\|_1}{\|\mathbf{v}_1\vee\mathbf{v}_2\vee\mathbf{v}_3\|_1} \tag{1.12}$$

for $i,j\in\{1,2,3\}$, together with the fact that $\|\mathbf{a}\vee\mathbf{b}\vee\mathbf{c}\|_1 - \|\mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c}\|_1 \le (\|\mathbf{a}\vee\mathbf{b}\|_1 - \|\mathbf{a}\wedge\mathbf{b}\|_1) + (\|\mathbf{b}\vee\mathbf{c}\|_1 - \|\mathbf{b}\wedge\mathbf{c}\|_1)$. Since the Ružička dissimilarity satisfies the axioms for a distance metric, our bipartite matching model for machine translation is a discrete version of the optimal transport in variational calculus.

During the ballpark screening stage of our machine translation algorithm, we divide our target document into K chapters and vectorize each topic pattern W_i by its chapter-wise counts $\mathbf{b}_i \in \mathbb{R}^K$. To pass a ballpark screening test, a pair of topics $\mathsf{W}_i^{\mathrm{A}}$ and $\mathsf{W}_j^{\mathrm{B}}$ from parallel versions (in languages A and B) of the same document have to satisfy

$$d_R(\mathbf{b}_i^{\mathrm{A}},\mathbf{b}_j^{\mathrm{B}}) \le \varepsilon\sqrt{K} \quad \text{and} \quad d_R(\mathbf{b}_i^{\mathrm{A}},\mathbf{b}_j^{\mathrm{B}}) \le \sqrt{\frac{\|\mathbf{b}_i^{\mathrm{A}}\wedge\mathbf{b}_j^{\mathrm{B}}\|_0}{\|\mathbf{b}_i^{\mathrm{A}}\vee\mathbf{b}_j^{\mathrm{B}}\|_1}}. \tag{1.13}$$

Here, the coefficient $\varepsilon = 0.07$ in the first inequality above is chosen empirically (see control experiments in Figs. S5, S8, S10, S11, S12, S13), so that the right-hand side extrapolates to a reasonable error margin[7] for total word counts (when $K = 1$). The second inequality in (1.13) draws upon a Poisson approximation to word count fluctuations, with $\|\mathbf{b}_i^{\mathrm{A}}\wedge\mathbf{b}_j^{\mathrm{B}}\|_0$ being the number of corresponding chapters that contain both $\mathsf{W}_i^{\mathrm{A}}$ and $\mathsf{W}_j^{\mathrm{B}}$. Through numerical experiments, we find that the screening criteria in (1.13) lead to acceptable recall and precision of machine translation later afterwards.

The vector $\mathbf{b}_i \in \mathbb{R}^K$ captures only the temporal behaviour of W_i on the chapter scale. To paint a fuller kinetic picture of W_i above the discourse level, we vectorize it by $\mathbf{v}_i$, the list of dominant recurrence eigenvalues in its semantic clique, and use $s_R(\mathbf{v}_i^{\mathrm{A}},\mathbf{v}_j^{\mathrm{B}})$ in our final similarity score for bipartite matching (see Fig. S1b,b′,b″, Tables S2–S5 for summaries of performance; see [1, Fig. 6] and Figs. S4–S16 for detailed bipartite matching of semantic fingerprints $\mathbf{v}_i$). In the current work, we do impose an additional censorship $s_R(\mathbf{v}_i^{\mathrm{A}},\mathbf{v}_j^{\mathrm{B}}) \ge 0.7$ before throwing similarity scores to the Hungarian algorithm for bipartite matching. This threshold value 0.7 is somewhat arbitrary. Relaxing this arbitrary censorship (*i.e.* reducing the threshold to 0) seems to have minimal consequences for translations between English and its close relative (see a control experiment for English–French in Table S6 and Fig. S8c), but seems to boost the recall rate significantly for translations between English and a non-Indo-European language (see a control experiment for English–Korean in Table S6 and Fig. S15c). Perhaps more work is needed to understand whether/how the threshold here adapts to the specific language pairs.

probability mass π_i to the ith Markov state) of Boltzmann's partition entropies $-\sum_j p_{ij}\log p_{ij}$ [10, §8.2]. For a formal discussion of entropy production and information compressibility in Markov chains, see [9, Chaps. 4–5] and [10, Chaps. 3, 8 and 14].

[5] The equality is attained when $\mathbf{P} = (p_{ij} = \frac{1}{N})_{1\le i,j\le N}$. This is the case of the least compressible Markov chain, whose maximal effective dimension is equal to N.

[6] Here as in [1, (4)], $\mathbf{a}\wedge\mathbf{b} = (\min\{a_1,b_1\},\ldots,\min\{a_N,b_N\})$ and $\mathbf{a}\vee\mathbf{b} = (\max\{a_1,b_1\},\ldots,\max\{a_N,b_N\})$ for $\mathbf{a} = (a_1,\ldots,a_N), \mathbf{b} = (b_1,\ldots,b_N)$; $\|\mathbf{u}\|_1 = \sum_{i=1}^N |u_i|$ for $\mathbf{u} = (u_1,\ldots,u_N)$.

[7] Possible error sources are imperfections in translation and word clustering. See discussions below.

In Figs. S3–S16, we run our experiments on multilingual corpora (Table S1) involving 9 members (Danish, Dutch, English, French, German, Latin, Polish, Russian and Spanish) in the Indo-European language family, 2 members (Finnish and Hungarian) in the Uralic language family, and 3 more agglutinative languages (Basque, Korean, Turkish) that are neither Indo-European nor Uralic, which exhibit diverse word orders[8] for subject (S), verb (V) and object (O):

SVO: Danish, English, French, Spanish;[9]
$SV_1O(V_2)$: Dutch, German;[10]
SOV: Basque, Latin, Korean, Turkish;[11]
Free word order: Finnish, Hungarian, Polish, Russian.[12]

The partial success of these experiments demonstrates that our Markov model (of topics on the long-range length scale) is translatable, regardless of a language's genealogical affiliation and syntactical structure.

Table S1. **Provenances of multilingual texts used in our experiments and discussions**

Language	**Author**/Adaptor	Title	Publisher	Electronic availability
			Indo-European	
English	**Jane Austen**	*Pride and Prejudice*	T. Egerton (Whitehall, UK, 1813)	`ExampleData` in *Mathematica* (essentially identical to the 1813 version, with modernized spellings)
	Deborah Moggach	*Pride and Prejudice*	Internet Movie Scripts Database (2005)	*`https://www.imsdb.com/scripts/Pride-and-Prejudice.html`
Danish	Vibeke Houstrup	*Stolthed og fordom*	Lindhardt og Ringhof Forlag (Copenhagen, Denmark, 2015)	`https://play.google.com/store/books/details?id=dd0-CgAAQBAJ&rdid=book-dd0-CgAAQBAJ` (Google Play)
German	Werner Beyer	*Stolz und Vorurteil*	Fischer Klassik Plus (Frankfurt am Main, Germany, 2011)	`https://www.amazon.de/Stolz-Vorurteil-Roman-Fischer-Klassik-ebook/dp/B006QH3IIC/` (Amazon)
	Christian Grawe & Ursula Grawe	*Stolz und Vorurteil*	Philipp Reclam jun. GmbH & Co. KG (Stuttgart, Germany, 2015)	`https://www.amazon.de/Stolz-Vorurteil-Roman-Reclam-Taschenbuch-ebook/dp/B01BHWFUX6/` (Amazon)
	Karin von Schwab	*Stolz und Vorurteil*	Null Papier (Düsseldorf, Germany, 2018)	`https://www.amazon.de/Stolz-Vorurteil-Vollständige-Ausgabe-Klassiker-ebook/dp/B0075O6EGS/` (Amazon)
Dutch	Margret Stevens & Annelies Roeleveld	*Trots en vooroordeel*	Athenaeum-Polak & Van Gennep (Amsterdam, The Netherlands, 2014), 2nd edn.	`https://www.amazon.de/Trots-vooroordeel-Dutch-Jane-Austen-ebook/dp/B00NYOLO1W/` (Amazon)
Spanish	Ana M. Rodríguez	*Orgullo y prejuicio*	Penguin Clásicos (Barcelona, Spain, 2016)	`https://www.amazon.com/Orgullo-prejuicio-mejores-clásicos-Spanish-ebook/dp/B00XJQQ9BG/` (Amazon)
	José de Urríes y Azara	*Orgullo y prejuicio*	Biblioteca Luna (Madrid, Spain, 2017)	`https://www.amazon.com/Orgullo-Prejuicio-Spanish-Jane-Austen-ebook/dp/B075FGRTPP/` (Amazon)
	Marciano Guerrero	*Primeras Impresiones u Orgullo y Prejuicio*	CreateSpace Independent Publishing Platform (2018)	`https://www.amazon.com/Primeras-Impresiones-Orgullo-Prejuicio-Illustrated-ebook/dp/B07CMJ8S4V/` (Amazon)
French	Valentine Leconte & Charlotte Presoir	*Les cinq filles de Mrs Bennet*	Librairie Plon (Paris, France, 1932)	*`https://fr.wikisource.org/wiki/Les_Cinq_Filles_de_Mrs_Bennet/Texte_entier` (Wikisource)
Latin	Thomas M. Cotton	*Superbia Et Odium*	Phaselus Publishing (Rhosilli, UK, 2015)	`http://www.lulu.com/shop/tom-cotton/superbia-et-odium/ebook/product-22288563.html` (Lulu)
Polish	Anna Przedpełska-Trzeciakowska	*Duma i uprzedzenie*	Prószyński Media (Warsaw, Poland, 2012)	*`http://www.rulit.me/books/duma-i-uprzedzenie-read-270058-1.html` (Electronic Library RuLit)
Russian	И. С. Маршак	*Гордость и предубеждение*	Художественная литература (Москва, СССР [Moscow, USSR], 1988)	*`http://lib.ru/INOOLD/OSTEN/gord.txt` (Maxim Mashkov's Library)
			Uralic	
Finnish	O. A. Joutsen	*Ylpeys ja ennakkoluulo*	Werner Söderström Osakeyhtiö (Porvoo, Finland, 1922)	*`http://www.gutenberg.org/cache/epub/45186/pg45186.txt` (Project Gutenberg)
Hungarian	Szenczi Miklós	*Büszkeség és balítélet*	Európa Könyvkiadó (Budapest, Hungary, 1958)	*`http://mek.oszk.hu/00300/00317/` (Hungarian Electronic Library)
			Vasconic	
Basque	Ana Isabel Morales	*Harrotasuna eta aurre-juzguak*	Elkar argitaletxea (Donostia, Spain, 2013)	`https://play.google.com/store/books/details/Austen_Jane_Harrotasuna_eta_aurrejuzguak?id=cMrNDQAAQBAJ` (Google Play)
			Koreanic	
Korean	원유경(元裕卿)	오만과 편견(傲慢과 偏見)	주식회사 열린책들(株式會社 열린冊들) (대한민국 경기도 파주시(大韓民國 京畿道 坡州市) [Paju, Gyeonggi-do, Republic of Korea], 2018)	`https://play.google.com/store/books/details/%EC%A0%9C%EC%9D%B8_%EC%98%A4%EC%8A%A4%ED%8B%B4_%EC%98%A4%EB%A7%8C%EA%B3%BC_%ED%8E%B8%EA%B2%AC?id=umMaCgAAQBAJ` (Google Play)
			Turkic	
Turkish	Işıl Önder	*Gurur ve Önyargı*	İletişim Yayınları (Istanbul, Turkey, 2019)	`https://play.google.com/store/books/details/Jane_Austen_Gurur_ve_%C3%96nyarg%C4%B1?id=EfCEDwAAQBAJ` (Google Play)
			Indo-European	
English	**Charlotte Brontë**	*Jane Eyre*	Service & Paton (London, UK, 1897)	*`http://www.gutenberg.org/cache/epub/1260/pg1260.txt` (Project Gutenberg)
	Moira Buffini	*Jane Eyre*	Internet Movie Scripts Database (2008)	*`https://www.imsdb.com/scripts/Jane-Eyre.html`
German	Maria von Borch	*Jane Eyre*	Jazzybee Verlag (Altenmünster, Germany, 2012)	`https://www.amazon.com/gp/product/B009XCD1D4/` (Amazon)
French	Noëmi Lesbazeilles Souvestre	*Jane Eyre ou Les moires d'une institutrice*	Librarie Hachette (Paris, France, 1847)	*`http://www.gutenberg.org/cache/epub/16235/pg16235.txt` (Project Gutenberg)
Russian	В. Станевич	*Джен Эйр*	Издательство «Правда» (Москва, СССР [Moscow, USSR], 1988)	*`http://www.lib.ru/INOOLD/BRONTE/janeair.txt` (Maxim Mashkov's Library)
			Uralic	
Finnish	Tyyne Haapanen-Tallgren	*Kotiopettajattaren romaani*	Werner Söderström Osakeyhtiö (Porvoo, Finland, 1921)	*`http://www.gutenberg.org/cache/epub/47275/pg47275.txt` (Project Gutenberg)
Hungarian	Ruzitska Mária	*Jane Eyre*	Európa Könyvkiadó (Budapest, Hungary, 1972)	*`http://mek.oszk.hu/05600/05680/` (Hungarian Electronic Library)
			Indo-European	
English	**Charles Darwin**	*Origin of Species*	John Murray (London, UK, 1873), 6th edn.	*`http://www.gutenberg.org/cache/epub/2009/pg2009.txt` (Project Gutenberg)
French	Edmond Barbier	*L'origine des espèces*	Schleicher Frères éditeurs (Paris, France, 1906)	*`http://www.gutenberg.org/cache/epub/14158/pg14158.txt` (Project Gutenberg)
Polish	Szymon Dickstein & Józef Nusbaum	*O powstawaniu gatunków drogą doboru naturalnego*	Przegląd Tygodniowy (Warsaw, Poland, 1884)	*`https://wolnelektury.pl/katalog/lektura/darwin-o-powstawaniu-gatunkow/` (Polish Free Reading)
Russian	К. А. Тимирязев & М. А. Мензбир & А. П. Павлов & П. А. Петровский	*Происхождение видов*	Наука (Санкт-Петербург, СССР [Saint Petersburg, USSR], 1991)	*`http://darwin-online.org.uk/content/frameset?itemID=F763b&viewtype=text&pageseq=1` (Darwin Online)
			Uralic	
Finnish	Aarno Rafael Koskimies	*Lajien synty*	Arvi A. Karisto (Hämeenlinna, Finland, 1913)	*`http://www.gutenberg.org/cache/epub/52187/pg52187.txt` (Project Gutenberg)
Hungarian	Kampis György	*A fajok eredete*	Neumann Kht. (Budapest, Hungary, 2004)	*`http://mek.oszk.hu/05000/05011/` (Hungarian Electronic Library)

Notes:
(1) Translators' names are given in native script and native order (surnames go first for Hungarian and Korean translators). Names of collaborating translators are separated by an & sign, irrespective of native practices. Books' titles and publishers' names are printed in their original languages. Publishers' addresses are translated into English if originally written in the Latin script. Otherwise, English translations (in brackets) accompany the original addresses in non-Latin script.
(2) Freely available electronic resources are marked with an asterisk. Descriptions of websites are translated into English, wherever applicable.
(3) The German (resp. Spanish) version of *Pride and Prejudice* used in the preparation of Table S2 is translated by Christian Grawe & Ursula Grawe (resp. Ana M. Rodríguez). The same will also apply to Fig. S5 (resp. Fig. S7).

[8] We do not consider questions or dependent clauses, whose word orders might differ from those of the declarative sentences as independent clauses.

[9] Both French and Spanish switch to SOV order when the object is a pronoun. Since we ignore pronouns (which belong to function words, as opposed to content words) in our Markov model, we can regard both French and Spanish as fully SVO languages.

[10] Here, the (optionally present) second verb (V_2) is in a non-finite form, such as an infinitive or a participle. In addition, Danish, Dutch and German are V2 (verb-second, not to be confused with the aforementioned V_2 notation) languages, placing the conjugated verb in the second position of a sentence, even when its preceding constituent is not the subject.

[11] In principle, one can freely switch word orders in Latin without affecting the meaning of a sentence: "Darcy loves Elizabeth" can be translated by all the six possible permutations of the words *Darceius*, *amat* and *Elizabetham*. In practice, the SOV order (such as *Darceius Elizabetham amat*) is dominant. The situation of Basque is similar to that of Latin. Korean and Turkish are both strictly SOV.

[12] Like Latin, these languages mark overtly subjects and objects through declensions, so word order is not important. It is arguable whether these languages should be classified as (predominantly) SVO, just as Latin is classified as SOV (by default).

Table S2. **Statistical dictionary generated from bipartite matching of semantic fingerprints, using Jane Austen's *Pride and Prejudice*, and its translation into an Indo-European language**

(en) English	(da) Danish	(de) German	(nl) Dutch	(es) Spanish	(fr) French	(la) Latin	(pl) Polish	(ru) Russian
absence			afwezigheid		absence			
admire	beundrede	Bewunderung						
agreeable	behagelig							
asked				preguntó				
aunt	tanten	Tante	tante	tía	tante	matertera	ciotka	тетушка
ball	bal	Ball	bal		bal		balu	бал
beauty	smukke			hermosa				красоты
believe	tror		geloven					
Bennet	Bennet	Bennet	Bennet	Bennet	Bennet	Bennet	Bennet	Беннет
Bingley	Bingley	Bingley	Bingley	Bingley	Bingley	Bingleius	Bingley	Бингли
book	bog	Buch	boek					книги
Bourgh		de			Bourgh	Burghana	Bourgh	де
breakfast						ientaculum		
Brighton	Brighton	Brighton	Brighton	Brighton	Brighton	Britonam	Brighton	Брайтон
brother	morbror	Bruder	broer	hermano	frère	fratris	brata	брат
carriage	vognen	Kutsche	rijtuig	coche	voiture		powóz	экипаж
case	tilfælde							
Catherine	lady	Lady	lady	Catherine	Catherine	Caterina	Katarzyna	леди
character			karakter					характер
Charlotte	Charlotte	Charlotte	Charlotte	Charlotte	Charlotte	Carlotta	Charlotta	Шарлотта
children	born	Kind	kinderen		enfants			детей
choose	valgte							
civility				cortesía				любезно
Collins	Collins	Collins	Collins	Collins	Collins	Collina	Collins	Коллинз
colonel	oberst	Oberst	kolonel	coronel	colonel	praefectus	pułkownik	полковник
comfort	trøst	Trost						
compliment	kompliment		compliment	cumplido				
congratulations					félicitations	gratulor		
consolation			troost	consuelo				
cousin		Cousine			cousin	consobrina	kuzyna	кузена
cried	udbrød		rep	exclamó	écria			
dance	danse	tanzen	dansen				tańce	танцевать
Darcy	Darcy	Darcy	Darcy	Darcy	Darcy	Darceius	Darcy	Дарси
daughter	datter	Tochter	dochter	hija		filia	córki	дочери
day	dag	Nachricht	dag	día		die	dnia	день
de	Bourgh	Bourgh	Bourgh	Bourgh			de	Бур
dear	kære		lieve		chère	carissima		дорогая
deceived				engaño				
Derbyshire	Derbyshire	Derbyshire	Derbyshire		Derbyshire	Derventiano		Дербишире
deserve	fortjent							
dine				comer		coenam		
dinner	middag	Essen	dineren		dîner		obiad	обед
disappointment	skuffelse	Enttäuschung						
dislike								неприязнь
distance					distance			
door	døren	Tür		puerta			drzwi	
elegant			elegante				wytworną	
Elizabeth	Elizabeth	Elizabeth	Elizabeth	Lizzy	Elizabeth	Elizabetha	Elżbieta	Элизабет
entered		eintritt				ingressus		
estate	gods	Erbe					majątek	
evening	aftenen	Abend	avond	velada	soir	vesperem	wieczór	вечер
eyes			ogen	ojos	yeux	oculos		
father	far	Vater	vader	padre	père	pater	ojciec	отец
felt			gevoelens		sentiments	sententias	uczucia	
Fitzwilliam	Fitzwilliam		Fitzwilliam	Fitzwilliam	Fitzwilliam	Filgulielmus	Fitzwilliam	
five	fem	fünf	vijf	cinco			pięć	пять
forget						oblivisci		
Forster	Forster	Forster	Forster	Forster	Forster	Forster	Forster	Форстер
fortnight	uger							
fortune	formue	Vermögen		fortuna	fortune			
four	fire	vier	vier	cuatro			cztery	четыре
Gardiner	Gardiner	Gardiner	Gardiner	Gardiner	Gardiner		Gardiner	Гардинер
gentlemen	herre		heer	caballero				
girls	pige	Mädchen	meisjes	muchachas			dziewczęta	
glad			blij					
good	godt	gut		mejor	mieux			
hand			hand					
handsome		hübsch	knap					
happy		Glück	geluk	felicidad			szczęścia	счастье
honour	ære	Ehre	eer	honor		honore		
house	huset	Haus	huis			domum		
Hurst				Hurst			Hurst	
husband	ægtemand			marido	mari			мужа
ignorant	uvidende			ignoro	ignore			
impossible	umuligt	unmöglich			impossible			
intended	hensigt							
invitation		Einladung	uitnodiging	invitación	invitation		zaproszenie	
Jane	Jane	Jane	Jane	Jane	Jane	Ioanna	Jane	Джейн
journey							podróży	
Kitty	Kitty	Kitty	Kitty	Kitty	Kitty	Kitteia	Kitty	Китти
know		weiß	weet				wiem	
lady					lady		lady	

(en) English	(da) Danish	(de) German	(nl) Dutch	(es) Spanish	(fr) French	(la) Latin	(pl) Polish	(ru) Russian
laugh		lachen	lachen		rire			
letter	brev	Brief	brief	carta	lettre	litteras	list	письмо
library	biblioteket	Bibliothek	bibliotheek	biblioteca		bibliotheca	biblioteki	библиотеку
Lizzy	Lizzy	Lizzy	Lizzy		Lizzy	Lisula	Lizzy	Лиззи
London		London	Londen	Londres		Londinii		Лондон
long	ønske	lange	lang			visum		
looked		blick		miró				
Lucas	Lucas	Lucas	Lucas	Lucas	Lucas	Lucas	Lucas	Лукас
Lydia	Lydia	Lydia	Lydia	Lydia	Lydia	Lydia	Lidia	Лидия
madam								сударыня
man	mand	Mann	man	hombre		homo		молодой
marriage	gift				mariage	matrimonium	małżeństwo	замуж
Mary	Mary	Mary	Mary	Mary	Mary		Mary	Мэри
means		fragen						
Meryton	Meryton	Meryton		Meryton	Meryton	Merytonae	Meryton	Меритоне
miles	miles					passuum		
Miss	Miss	Miss	juffrouw	Miss	Miss	hera	panna	мисс
money	penge	Geld		dinero		pecunia	pieniędzy	денег
mother	mor	Mutter	moeder	madre	mère	matrem	matka	матери
motive	bevæggrund				motif			
Mr	Mr	Mr	mijnheer		Mr			Мистер
Mrs	Mrs	Mrs	mevrouw	Mrs	Mrs	Mra		Миссис
neighborhood	nabolaget						sąsiedztwie	
nephew	nevø	Neffen			neveu			
Netherfield	Netherfield	Netherfield	Netherfield	Netherfield	Netherfield	Infrapratensem	Netherfield	Незерфилде
niece	niece			sobrino	nièce		siostrzenica	
officers	officererne	Offiziere	officieren	oficiales	officiers	subpraefectos	oficerów	офицеров
oh	åh		o	oh				
pain	smerte							
park	parken	Park	park	parque	parc		park	парк
Pemberley	Pemberley	Pemberley	Pemberley	Pemberley	Pemberley	Pemberleianam	Pemberley	Пемберли
people			mensen				ludzi	
Phillips	Philips	Philips	Philips	Philips	Philips	Philipa	Philips	Филипс
plan							plan	
pleased						honestiores		
politeness					politesse			
poor	stakkels	arme	arme	pobre	pauvre			
pounds	pund	Pfund	pond				funtów	тысяч
pride	stolthed	Stolz	trots	orgullo			duma	гордость
promised				prometido	promesse			
read	læste	las	las	leer			czytać	
refuse				rechazar	refus	recusare		
regiment	regimentet	Regiment	regiment	regimiento	régiment			полк
relations	fortalte							
respect				respeto			szacunku	
room		Zimmer	kamer				pokoju	комнату
Rosings	Rosings	Rosings	Rosings	Rosings	Rosings	Rosariam	Rosings	Розингсе
said			zei					сказала
sat	sad							сидела
see			zien					
seemed			leek					
send	sendt			enviado	envoyer			
serious	alvorligt	ernst	serieus					
servant		Diener	bediende	criado		ministris	służącego	
Sir	sir	Sir	William	sir	sir	Senior	sir	сэр
society					société			
sorry	ked		spijt					
speak	tale		spreken			loqui		
surprise	overrasket	überrascht		sorprendida		attonita		
table	bordet	Tisch	tafel		table		stole	столом
talking	talte					colloqui		
tell			vertellen					
ten	ti		tien	diez	dix			десять
thank					remerciements			
thousand	tusinde		tienduizend	mil	mille	librarum	tysięcy	фунтов
town	London			capital	Londres		Londynu	
turned	vendte							
two	to		twee		deux			
uncle	morbror	Onkel	oom	tío	oncle	avunculus	wuja	дядя
vain	forfængelighed			vanidad		vano		тщеславие
visit	besøg	Besuch	bezoek		visite		wizyty	
walk						ambulare		
want	mangel							
week	uge		weken	semana	semaine	septimanis	tygodnie	недели
Wickham	Wickham	Wickham	Wickham	Wickham	Wickham	Vicham	Wickham	Уикхем
wife	hustru					uxor	żony	
William	William	William	sir	William	William	Gulielmus	William	Уильям
window	vinduet	Fenster	raam	ventana		fenestram	okna	окна
woman	kvinde		vrouw	mujer			kobieta	женщина
write	skrive	schreiben	schrijven	escribir	écrire	scripsit	pisać	написать
years	år	Jahr	jaar	años	ans	annos		лет
young	unge	jungen	jongedames	joven	jeune	nātu		

Notes:
(1) See Table S1 for provenances of texts.
(2) This table is sorted alphabetically according to the English entries.
(3) For clarity, only the most frequent word within a word pattern is tabulated. See Figs. S4–S11 for details of bipartite matching.
(4) Color encoding follows [1, Fig. 5c]: green = correct match; amber = close match; red = incorrect match.
(5) Column separator demarcates different subgroups (Germanic, Romance & Latin, Slavic) of the Indo-European language family. A Germanic compound noun (such as Danish *morbror*) may appear as the translation of more than one word.

Table S3. **Statistical dictionary generated from bipartite matching of semantic fingerprints, using Jane Austen's *Pride and Prejudice*, and its translation into a non-Indo-European language**

(en) English	(fi) Finnish	(hu) Hungarian	(eu) Basque	(ko) Korean	(tr) Turkish
admire					hayranlık
arrival				도착(到着)한	
attachment			atxikimendu		
aunt	täti		izeko		
ball		bál		무도회(舞蹈會)에	balo
believe	uskoa				inanıyorum
Bennet	Bennet	Bennet	Bennet	베넷(Bennet)	Bennet
Bingley	Bingley		Bingley	빙리(Bingley)	Bingley
book			liburu		
Bourgh	Bourghin		Bourgh		Bourgh
breakfast			gosari		
Brighton		Brightonba	Brightonera	브라이튼(Brighton)에	Brighton'a
brother	veljensä	bátyja	neba		
carriage	vaunut		kotxea		araba
Catherine	lady	Catherine	lady	캐서린(Catherine)	Lady
character		jellemét	izaera		karakteri
Charlotte	Charlotte	Charlotte	Charlotte	샬럿(Charlotte)이	Charlotte
children	lapseni	gyermekem			çocuklarının
civility		udvariasan			nezaketle
Collins	Collins	Collins	Collins		Collins
colonel	eversti	ezredes	koronelak		
come		jött	etorri	오는	
compliment					ilfifat
congratulations	onnitteli			축하(祝賀)해	tebrik
consolation			kontsolabide		teselli
country			herri		
cousin	serkkunsa		lehengusuak	사촌(四寸)	kuzeninin
cried		kiáltott			
dance				춤을	dans
Darcy	Darcy			다시(Darcy)	
daughter	tyttärensä			만	
day	päivänä	kedves	egun	시간(時間)을	gün
de	de	Bourgh	de	버그(Bourgh)	
dear			ene		sevgili
Derbyshire		Derbyshire-ben	Derbyshiren	더비셔(Derbyshire)에	
different					farklı
dine			bazkaltzeko		
dinner		vacsorára	bazkari		
disappointment		csalódott		실망(失望)을	kırıklığına
door			atea	문(門)을	
elegant			dotorezia	우아(優雅)한	
Elizabeth	Elizabeth				
evening		este			akşam
eyes				미소(微笑)를	göz
face	kasvot				
father	isä	apja	aita	아버지가	babasının
felt		érzett			hisleri
fine					
Fitzwilliam		Fitzwilliam	Fitzwilliam	피츠윌리엄(Fitzwilliam)	
five	viisi		bost		
forget			ahaztu		unutmaya
Forster	Forster		Forster	포스터(Forster)	Forster
four		négy	lau		dört
Gardiner	Gardiner	Gardiner	Gardiner	가디너(Gardiner)	Gardiner
gentlemen			zaldun		beyler
girls			neska		
good			ona	좋은	
hand		kezét			
happy	onnea	boldog		행복(幸福)한	
honour	kunnia		ohorea		
house		ház			
Hurst			Hurst		
husband		férjhez	senarra	남편(男便)이	kocasının
impossible	mahdoton				imkânsız
intended					niyetli
interest			interesgarriago		
introduction				소개(紹介)해	
invitation	kutsua		gonbidatu	초대(招待)를	davet
Jane	Jane	Jane	Janek	제인(Jane)이	Jane
Kitty	Kitty	Kitty	Kitty	키티(Kitty)가	Kitty

(en) English	(fi) Finnish	(hu) Hungarian	(eu) Basque	(ko) Korean	(tr) Turkish
know			jakin		
laugh	nauraa				
letter	kirjeen	levelet			mektubu
library	kirjastoon	könyvtárszobába	liburutegira	서재(書齋)로	kütüphaneye
like			antzun		
Lizzy	Lizzy	Lizzy	Lizzy	리지(Lizzy)야	Lizzy
London		Londonban	Londresen		Londra'ya
long	päästä	akarta			
Lucas	Lucas	Lucas	Lucas	루커스(Lucas)	Lucas
Lydia	Lydia	Lydia	Lydia		Lydia
man	mies	ember			
marriage			ezkontza	결혼(結婚)을	evlenme
Mary	Mary	Mary	Maryk	메리(Mary)는	Mary
master			ugazaba		
means		teheti			
Meryton	Merytonissa	Merytonban	Merytonen		Meryton'a
Miss	nti	Miss	andereñoa	양(孃)이	Miss
money	rahoja	pénzt			para
mother	äiti	anyja		어머니가	annesi
Mr	hra				Mr
Mrs	rva	Mrs		부인(夫人)은	Mrs
neighborhood				이웃	
nephew	sisarenpoikansa	unokaöccse			
Netherfield	Netherfieldin	Netherfieldi	Netherfielden	네더필드(Netherfield)에서	Netherfield'da
niece		unokahúga			yeğeni
officers				장교(將校)들	subay
oh	ah		o	아	ah
pain					acı
park	puiston	park	parke		
Pemberley	Pemberleyn	Pemberleyben	Pemberley	펨벌리(Pemberley)에	Pemberley'e
Phillips	Philips	Philips	Philips	필립스(Philips)	Phillips
poor		szegény			
pounds			libra		pound
pretty				예쁜	
pride	ylpeä		harro		gururu
promised	lupasi		agindu		
read	lukea	olvasott			okumaya
refuse	hyljännyt			거절(拒絶)하는	reddetti
relations	kertoi				akrabaları
respect					saygı
room			gelan		
Rosings	Rosingsissa	Rosingsi	Rosingsen	로징스(Rosings)에서	Rosings'teki
said					dedi
sat	istui		jesarrita		
see			ikusi		gördü
send			bidali		
serious			serio	진지(眞摯)하게	ciddi
servant	palvelijan			하인(下人)이	hizmetçi
sir	sir	sir	William	경(卿)이	William
suffer		szenvednie			
surprise		meglepte			
table		asztalhoz	mahai		masaya
ten	kymmenen		hamar		
thank	kiitollinen				teşekkür
thousand	puntaa			파운드(pound)를	
town	Lontooseen				
turned	hymyillen				
two			bi		
uncle	enon	nagybátyja		숙부(叔父)와	dayısı
vain		hiába		허영심(虛榮心)을	
visit			bisita	방문(訪問)을	ziyaret
way	ajatus		aldatu		
week	viikon		aste		hafta
Wickham	Wickham	Wickham		위컴(Wickham)	Wickham
wife	vaimonsa			아내가	karısı
William	William	William	sir	윌리엄(William)	sir
window	akkunasta	ablakhoz	leiho		
woman		nő			
write	kirjoittaa	írt	idatzi		yaz
years	vuotta	éves	urte		yıl
young	nuori		gazte	젊은	genç

Notes:
(1) See Table S1 for provenances of texts.
(2) This table is sorted alphabetically according to the English entries.
(3) For clarity, only the most frequent word within a word pattern is tabulated. See Figs. S12–S16 for details of bipartite matching.
(4) Color encoding follows [1, Fig. 5c]: green = correct match; amber = close match; red = incorrect match.
(5) Column separator demarcates different language families.
(6) Korean stems that derive from Chinese or English are accompanied by their etymologies in parentheses.

Table S4. **Statistical dictionary generated from bipartite matching of semantic fingerprints, using Charlotte Brontë's *Jane Eyre*, and its translation into a European language**

(en) English	(de) German	(fr) French	(ru) Russian	(fi) Finnish	(hu) Hungarian
Abbot		Abbot	Эббот	Abbot	Abbot
accompany	begleiten				
Adele	Adele	Adèle	Адель	Adèle	Adèle
admiration	Bewunderung	admiration			
afternoon	Nachmittag			iltapäivällä	délután
angel	Engel	ange			angyalok
answer				totta	
arms		bras		käsivartensa	
aunt	Tante	tante	тетя	täti	néni
bed	Bette				ágy
Bessie	Bessie	Bessie		Bessie	Bessie
birds	Vögel	oiseaux	птицы		madarak
black	schwarzen	noir	черные	musta	fekete
blood	Blut	sang	кровь		
blue	blauen	bleu			kék
book	Buch	livres	книгу	kirjan	könyvet
bread	Brot				kenyeret
breakfast	Frühstück	déjeuner	завтрака	aamiaista	
Brocklehurst	Brocklehurst	Brocklehurst	Брокльхерст	Brocklehurst	Brocklehurst
brother	Bruder	frère	брат	veljensä	bátyja
Burns		Burns	Бернс	Burns	
called	Licht	appelait		tunnin	magas
candle			свечу	kynttilä	gyertyát
chair	Stuhl		кресло	tuolit	
change		effet			
child	Kind	enfant	ребенок	lapsi	gyermek
Christian	Christen	Chrétien			
church	Kirche	église	церковь	kirkon	
class	Klasse	classe			
clergyman				pappi	
cloak	Mantel	manteau			
cloud	Wolken	nuages		pilvien	
cold	holen	retirer	холодно	kylmä	hideg
come	kam	venait			
cousins		cousines	кузины	serkkuni	
curls	Locken	boucles		kihara	
curtain		rideau	занавески		függönyt
danger	Gefahr	danger		vaarassa	
dark	dunkel	lumière			sötét
daughter			дочь	tytär	
day	Zeit	grande	день	päivänä	nap
dead	tot	mort	смерти		
dear		chère			
deep	tief			syvä	mély
Dent	Eshton		Дэнт	Dent	
Diana	Diana	Diana	Диана	Diana	
dinner		dîner	обеда		
discovered	entdeckt				
dog	Hund	chien	собака	koira	
door	Thür		дверь	oven	ajtót
doubt	Zweifel				
drawing-room	Salon			seurusteluhuoneeseen	
dream	Traum	rêve			álmodtám
dressed	Kleider				ruhát
eat					enni
Edward	Edward	Edouard		Edward	Edward
eight			восемь	kahdeksan	
Eliza	Eliza		Элиза	Eliza	Eliza
England		Angleterre	Англии		
Eshton	Dent			Eshton	
evening		soir	вечером		este
event			событие		
existence		existence			
eyes	Fuhr	seule	глаза	silmät	arca
Eyre	Eyre	Eyre	Эйр	Eyre	Eyre
face	Gesicht			kasvot	
Fairfax	Fairfax	Fairfax	Фэйрфакс	Fairfax	Fairfax
family	Familie	famille	семьи	perheen	
father	Vater	père	отец	isäni	apám
fear	Furcht	crainte		pelkää	
feet	füßen			kulki	
fell	fast	tomber			
felt				puhua	
field		champs			
figure	Speisezimmer				
fine			прекрасно		
fingers	Finger	doigts			
fire	Feuer	assis	представить		tüz
first	bringen	sortir	удалось	palasi	esendes
five	fünf	cinq	пять		
flesh	fleisch				
flowers	Blumen	fleurs			virágok
forget	vergessen	oublié	забыла	unohtanut	
forgive		pardon			
four	vier	quatre		neljä	
French	Französischen	Français	Французски		Francia
friends	Freunde	amis			barátságos
gallery	Galerie	corridor		gallerian	

(en) English	(de) German	(fr) French	(ru) Russian	(fi) Finnish	(hu) Hungarian
garden	Garten	jardin	сад		kertbe
Gateshead	Gateshead	Gateshead	Гейтсхэда	Gatesheadista	
Georgiana	Georgina	Éliza	Джорджиана	Georgiana	Georgiana
girl	Mädchen			tyttö	
go	gehen	aller			menjen
god	gott	dieu		jumala	isten
gold					arany
good		mieux			
governess	Gouvernante		гувернантка	kotiopettajatarnevelönö	
grace		grace			
ground					becsukta
hair	Haar	cheveux	волосы	tukka	
half	halb				helyes
hall	Hall				
hand	Worte	main	руку	mennä	kezét
Hannah	Hannah	Anna			Hannah
happy			счастлива	onnellinen	
hate				vihaa	gyülölettel
head	Kopf			paha	fejét
health	gesund				
heard	hörte	entendre			hallottam
heart	Herz	coeur	чувства	sydämeni	szivem
Helen	Helen		Элен		
hills	Hügel		холмы	kukkulan	
home				kotiin	
hope		espère	желание	levollinen	
horse	Pferd	cheval	лошадь	hevonen	
hour	Stunde	heure	час		óra
house	Hause	maison		talon	ház
human	Menschliche				
hungry	Hunger				
husband	Gatten	mari			
Ingram			Ингрэм		Ingram
Jane	Jane	Jane	Джен	Jane	Jane
John	John	John	Джон	John	
kiss		embrasser	поцеловала		megcsókolam
know			знаю		
lady	Damen				hölgy
laugh	lachen	rire		naurua	
Leah	Leah	Leah		Lea	Leah
left				jättää	
letter	Brief	lettre	письмо	kirjeen	levelet
library	Bibliothek			kirjastoon	könyvtárba
life	Leben		жизнь	elää	
light	Schloß	air	свете	valo	Bosszú
lips					ajkä
listen		écouter			
little	kleine	petite		pieni	
long	lange	longtemps		aikaa	
love			любовь	rakkautta	
low	drei	bas			
Lowood	Lowood	Lowood	Ловуде	Lowoodin	Lowood
Lynn			Эштон	Lynn	Eshton
ma'am			сударыня		
man		homme			
married			жениться	naimisiin	
Mary	Mary		Мери	Mary	Mary
master		maître	хозяин		
miles	Meilen			peninkulman	mérföldnyire
Millcote	Millcote	Millcote		Millcotesta	Millcote
mind	gedanken	plaisir	просто	toi	nyugodtan
Miss	Miß	mlle	мисе	Miss	kisasszony
money			денег	rahaa	
moon	Mond	lune			
moor	Moor				
morning	morgen		утром	aamuna	reggel
Morton	Morton			Mortonin	
mother				äitinsä	
mouth	Mund	bouche			
Mr			Мистер		Mr
Mrs	Mrs		Миссис		Mrs
name	Namen	nom		nimen	
nature	Art	nature	грудию		
nerves	Nerven		нервы		
new	neuen		новую	uuden	új
night			ночь		éjszaka
old	recht	vieux		vanha	
Oliver	Oliver				
opened					nyitott
paper		papier	бумаги		
parlour					
part					nappali
passed		servir	прошло	pitkä	szokott
passion	Leidenschaft				adott
Pilot	Pilot	Pilote		Pilot	Vezér
place			место	paikka	éreztem
pleasure	Freude				
Poole	Poole	Poole	Пул	Poole	Poole

(en) English	(de) German	(fr) French	(ru) Russian	(fi) Finnish	(hu) Hungarian
poor		pauvre	бедный		szegény
pounds	Pfund		фунтов		
prayers					rukoukset
pride	Stolz		гордость	ylpeä	büszke
principle		principes			
promise	versprechen	promesse			
prove	beweisen				
rain		pluie			
ran					nüzza
read	lesen	comprendre	читатель	työtä	olvasni
reed	Reed	reed	рид	reed	
relations		parents			
remember				muistin	
returned	Ruhe	silence		seisoi	visszajött
right			чисто		
Rivers	Rivers	Rivers	Риверс	Rivers	Rivers
road		route			
Rochester	Rochester		Рочестер	Rochester	Rochester
roof	Dache				
room	Zimmer	chambre	комнате		
rose			поднялась	nousi	világon
sacrifice	Opfer	sacrifice			
sat	begann	figuré	разговор	istui	engedje
Scatcherd	Scatcherd	Scatcherd	Скетчерд	Scatcherd	Scatcherd
school			школу	kouluun	
schoolroom	Schülzimmer		класс	kouluhuoneessa	
secret		secret			
sense				käytöksensä	
sent		envoyer		anna	
servant				näkönen	
side		côté			
silk		soie		silkkiin	
sincere		sincèrement			
sing	singen	chanter			
Sir	Sir	mr	сэр	Sir	
sisters	Schwestern	soeurs	сестры	sisarensa	húga
sleep		dormir		nukkua	
smile	lächeln	sourire	улыбка	hymyili	
Sophie	Sophie	Sophie		Sophie	Sophie
soul		âme		sieluni	
speak	Kraft	oh		mahdotonta	hagytam
spirit			убежище		
St		Saint			
stay		rester			
stopped			остановился	katosi	megállt
storey			этаж		
strange	seltsam		странное		
suffer			страдания		
sun	Sonne		солнце		
sweet	süße				gyönyörü
table	Tische	table	стол	pöydän	asztalon
talk				uskon	
tea	Thee	thé			teát
teachers	Lehrerin			opettajat	
tears	Thränen	larmes	слезы		
temple	Temple	temple	гемпль		
thing					tossz
Thornfield	Thornfield	Thornfield	Торнфильд	Thornfield	Thornfield
thousand			тысяч		
three		trois		kolme	három
told				viisi	
town	Stadt				
treated	Behandlung				
trees	Bäume	arbres			
twenty		vingt	двадцать		
two	zwei	deux	два	kaksi	
uncle	Onkel	oncle	дядя	setänne	nagybátyám
veil	Schleier		вуаль		fátyol
voice	Stimme	voix	голос	voimakas	remélem
wait	warten				
wake		éveiller			
wall			стены		fal
want		besoin			
water	Wasser	eau	воды		
way	bleiben	quitter			
week			недели	viikon	
whispered	flüsterte			kuiskasi	
white			Белым		fehér
wife				vaimonsa	
wind	Wind	vent	ветер		
window	Fenster	fenêtre	окна		ablakon
winter		hiver		talvi	
wish	wünschte	désirais	придется	toivon	igazán
woman			огонь	nainen	nő
world			дух		
written	schreiben	écrit	написала	kirjoittaa	
years	Jahre	ans	лет	vuotta	évvel
young	jungen		молодой	nuori	

Table S5. **Statistical dictionary generated from bipartite matching of semantic fingerprints, using Charles Darwin's *Origin of Species*, and its translation into a European language**

(en) English	(fr) French	(pl) Polish	(ru) Russian	(fi) Finnish	(hu) Hungarian	(en) English	(fr) French	(pl) Polish	(ru) Russian	(fi) Finnish	(hu) Hungarian	(en) English	(fr) French	(pl) Polish	(ru) Russian	(fi) Finnish	(hu) Hungarian
accumulated	accumulation					extent	reste					order	ordres				
acquired	acquis					external	extérieures					organisation					szervezet
adapted	adaptation		адаптации			extinct				sukupuut-toon		Owen			Оуэн	Owen	Owen
adjoining					szomszédos							parallelism	parallélisme	równoległość			párhuzam
adult				láysinke-hittyneen		extraordinaire	extraordinaire					parts					részek
affinities	affinités					extremely			вследствие	käy		period			периода		
Africa	Afrique	Afryki	Африки	Afrikan	Afrika	eye			глаза			perish	périssent				
air		powietrzu				F	F	F	F			pigeons	pigeons		голубей		galamb
allied		spokrewnione				families	familles		семейства		család	pistil	pistil				
America	Amérique	Ameryki		Amerikan	Amerikai	faunas	faunes	fauny	фауны			place	place				
analogous		analogiczne				favourable	favorables					plants	plantes	roślin	растений		
animals	animaux					feathers	plumes					pollen	pollen	pyłek	пыльцы	siitepölyä	
answer			ответ			females	femelles				nőstény	principle	vertu	zasady	принцип	alkuperää	
ants			рабочих		dolgozók	fertility	fécondité	płodność				probably		prawdopo-dobnie			
archipelago	archipel	archipelagu	архипелага			fishes		ryb		kalojen							
arctic	arctiques	strefy		lauhkean	sarkvidéki	flight				lentoon		process	conséquent		процесс		
asked	demander					floated	flottantes	arktyczne				Professor	professeur		проф	professori	professzor
attributed			невероятно			flowers	fleurs	kwiaty			virágok	progenitor	ancêtre				őstől
Australia		Australii		Australian	Ausztrália	follow				seuraavat	biztos	purpose				tarkotukseen	
authors			авторы			formerly			настоящее			quadrupeds			четвероногих		
base			тонких			forms		formy				range				levenemisalue	
bats	chauves	nietoperza	летучих	lepakon	denevérek	fossil	fossiles		ископаемые			reciprocal		obustronnym	реципрокных	siitoselimistö	
beak	bec					found		znajdujemy	найти			remains			открыты		
beautiful		piękne				fresh	douce		пресноводные	suolattoman	édesvízi	remote					nyílt
beds				üledék		fruit	fruits	owoce				reptiles	reptiles				
bees		pszczoły	пчелы	mehiläiset	méhek	function	fonctions	funkcji	функции	tehtävänä		resemble				muistuttavat	
beetles		chrząszcze	жуков		bogarak	future			будущем			reversion	bleu		реверсии		
beings	êtres					Gartner	Gärtner	Międzygatunk	Гертнер	Gärtner	Gärtner	rock-pigeon	biset			kalliokyyh-	szirti
believe	sait					genera	genres									kysestä	
belonging			принадле-жащие			generally				yleensä		rudimentary	rudimentaires		рудиментар-ные	surkastuneet	
						generations		pokoleń	поколений								
						geographical	géographiques	geograficzne	географиче-ских		földrajzi	rule	règle				
birds	oiseaux	ptaków	птиц		madarak							said				tärkeä	
bones				valaan		geological	géologiques	geologicznych		geologian	geológiai	sea			моря	meren	
botanists	botanistes		ботаниками			germinated	germèrent		прораста-нию		meleg	secondary	secondaires				
breeds	races											sediment	sédiments		разрушения		süllyedés
Britain		Brytanii				glacial	glaciaire	lodowcowej	ледникового	jääkauden	jégkorszak	seedlings				taimia	
called		nazwać			nevezetesen	gradations			градации			seeds	graines	nasion		siemeniä	magot
Cambrian	Cumbrienne	Kambryjskim		Kambrisen	Kambriumi	grafted	greffe	między-mianowych	прививки	vastavuoro-isten	reciprok	selection	sélection	doboru	отбора	valinnan	kiválasztás
Candolle	Candolle	Candolle	Декандоль	Candolle	Candolle							separated			я		
carefully			тщательно	vaihtelee		groups	groupes	grupy	группы	ryhmän	csoportok	sexual	sexes	płci			
case			случаях			habits	habitudes					sheep	moutons				juhot
cattle			скота			hand				helposti		short	courte				rövid
cause	cause		причины	uskoa	oka	head			головы			side					oldalán
caves			пещерах	luolissa	barlangi	hemisphere		półkuli		pallonpuo-liskon		similar					különféle
cells	cellules	komórek	ячейки									simple				yksinkertai-sesti	egyszerű
central	centre		центров			hermaphrodites	hermaphrodites										
certain			основании			homologous		homologiczne	гомологичны	homologisia	homológ	simultane-ously	simultané-ment				
change				muutoksia		Hooker	Hooker	Hooker	Хукер	Hooker							
chapter	chapitre	rozdziale				horns	cornes	rogów				Sir	sir	sir	сэр	ulottuu	sir
characters	caractères	cechy				horse	cheval		лошади	hevosen		six	six				
cirripedes	cirripèdes	wąsonogich		siimajal-kaisten	kacslábú	hybrids	hybrides	mieszańców	гибридов		hibridek	slaves	esclaves	wosku	воска		
						identical		identyczne				slowly	lentement		медленно		
class	classe					ignorance						south		południowej	южной	etelä	déli
climate	climat	klimatu	климата		éghajlat	illegitimate	illégitimes	nieprawowitych				special	spéciale			kysymys	
climbing	grimpantes		боку	kiipeile		immigration	immigration		иммиграции			species	espèces			lajien	
cold			холодных	kylmyyden	hideg	implies	implique					specific	spécifiques				
colour			окраска			important	importance					spheres		plastra			
compared				verran		impossible		zwanych			lehetetlen	stamens	étamines	pręcików	тычинок	heteistä	
competition			конкурен-цию	kilpailuun		India		Indiach			Indiában	sterility	stérilité		стерильности		terméket-lenség
						infinitely					végtelen						
complex	complexes			monimut-kaiset	bonyolult	inhabitants			обитателей			stigma	stigmate		рыльце		
conditions	conditions		условиях			inherited	hérédité					strata			осадков		
continent				mantereen	kontinens	injurious			вредным			striped	raies				csíkos
correlated	corrélation	współzależ-ności	корреляции			innumerable	innombrables					struggle			борьба		küzdelem
						insects	insectes	owadów	насекомых	hyönteisten	rovarok	subject	sujet		вопрос		
						instincts	instincts	instynkt	инстинкты		ösztönök	subsidence	affaissement	obniżania			
correspond-ing	correspond-ant					intermediate	intermédiaires		промежуточ-ных			suddenly			внезапно	äkkiä	hirtelen
												sufficient				lyhyesti	
creation	création	stworzenia				intervals	intervalles		промежутки			supplant					
crossed		krzyżowaniu			kereszteződés	islands	îles	wyspy	острова	saarilla	szigetek	supposed	exterminées				
crustaceans				äyriäisten		jaws	mâchoire					system	supposer				
cuckoo	coucou	kukułki	кукушки			judge	juger					tail	système				
de		de		de		kind		sposobem				teeth	queue	ogona	хвоста		
depth					mély	kingdom	règne					temperate	zęby				
described	décrit	opisuje				larvae	larves	larwy	личинки	toukka	lárvák	ten	tempérées		умеренных	vyöhykkeen	mérsékelt
developed	développe-ment					laws			законы		lárvények	tendency	tendance			kymmenen	
						level	niveau					terrestrial					
deviations	déviations					lines	lignes		линии			tertiary	tertiaires	trzeciorzędo-wych	третичных	maapallolla	
diagram		schemacie	диаграмме			links			звеньев								
different	différentes				különböző	long	longtemps					theory	théorie	teorii	теории	teoriamme	elmélete
difficulty	difficulté		трудности	vaikeata		look			славе	tarvoin	igaz	thickness	épaisseur				vastag
direct	direction					lower	inférieure					third		trzecie			
discovered	découvrir				felfedezése	Lyell	Lyell	Lyell	Лайелль	Lyell	Lyell	thousand			тысяч		
dispersal	dispersion					Madeira	Madère	Madery	Мадейре	Madeiran		three	trois	trzy		kolme	
distinct	distinctes					males	mâles					trees	arbres	drzewa	деревьев		
disuse		nieużywania	неупотре-бления	puutteen		mammals	mammifères	ssaków	млекопи-тающих	imettävä-isten	emlősök	true			редко		
diversity	transmettre					man	homme		человека	ihmisen	ember	tumbler	culbutant	młynka	турмана	kuperkeikka-kyyhkysen	bukóga-lamb
division	division					masters		robotnic			sejtjét	two			двух		
dogs	chiens		собак			members				jäsenet		type			типа		
domestic	domestiques					metamorphosed	métamorphose					unconscious	inconsciente		бессозна-тельного	itsetiedotto-masti	szándék-talan
dominant			доминирую-щие			migration		wędrówki	миграции								
						miles			миль	penikulman		understand		zrozumieć	понять		
doubt	doute	wątpienia	сомнения	epäilemättä	hiszem	Mivart	Mivart	Mivart	Китового	Mivart	Mivart	unknown	inconnues			tuntematon	ismeretlen
Dr	docteur			tohtori		modified			модифи-кации			upper			верхней		
dry	sommets			kuivia								useless				hyödyttömiä	
duck	canard					mongrels		szczepienia	воспроизво-дительной	risteytetty		varieties		odmiany			
duration			кембрий-ской		ösmaradvá-nyok				уродства			view		przyjąć			megma-gyarázni
economy			экономии			monstrosities			горах								
effects		wpływ				mountains	montagnes				hegyi	visited		odwiedzają			
eggs	oeufs		яйца			movement					alsó	wall			стенки		
element	élément					Mr					úr	water	eau				
embryo		zarodka	зародыша			Muller		Müller	Мюллер	Müller	Müller	wax	cire				
enemies			врагов	vihollisten		nearly				seikka		whale	baleine	wieloryba			bálna
equator			экватором	päivänta-saajan		nectar		nektaru	нектара			white		białą			
						nest		gniazda	гнезда			widely				läheistä	
						neuters	neutres			suvuttomat	ivartalan	wild			диких		
Europe	Europe	Europy		Euroopan	Európai	new	nouvelles	nowych	новых	uusia	új	wings	ailes	skrzydła	крылья		szárnya
evidence				merkillistä		north		północnej	северной	pohjois		years	années		лет	vuotta	évben
expect	attendre		ожидать			oceanic			океани-ческих			young	jeunes	młode			
experiments	expérience											Zealand	Zélande	Zelandii	Зеландии		

Table S6. **Statistical dictionaries compiled from *Pride and Prejudice* and its translations into French and Korean, with different thresholds for Ružička similarities**

(en) English	(fr) French	
$s_R(\mathbf{v}_i^{\mathrm{en}}, \mathbf{v}_j^{\mathrm{fr}}) \geq 0.7$		≥ 0
absence	absence	absence
aunt	tante	tante
ball	bal	bal
Bennet	Bennet	Bennet
Bingley	Bingley	Bingley
Bourgh	Bourgh	Bourgh
Brighton	Brighton	Brighton
brother	frère	frère
carriage	voiture	voiture
Catherine	Catherine	Catherine
Charlotte	Charlotte	Charlotte
children	enfants	enfants
Collins	Collins	Collins
colonel	colonel	colonel
congratulations	félicitations	félicitations
cousin	cousin	cousin
cried	écria	écria
dance		danser
Darcy	Darcy	Darcy
dear	chère	chère
Derbyshire	Derbyshire	Derbyshire
dinner	dîner	dîner
distance	distance	distance
Elizabeth	Elizabeth	Elizabeth
evening	soir	soir
eyes	yeux	yeux
father	père	père
felt	sentiments	sentiments
Fitzwilliam	Fitzwilliam	Fitzwilliam
Forster	Forster	Forster
fortune	fortune	fortune
Gardiner	Gardiner	Gardiner
good	mieux	mieux
Hurst		Hurst
husband	mari	mari
ignorant	ignore	ignore
impossible	impossible	impossible
invitation	invitation	invitation
Jane	Jane	Jane
Kitty	Kitty	Kitty
lady	lady	lady
laugh	rire	rire
letter	lettre	lettre
Lizzy	Lizzy	Lizzy

(en) English	(fr) French	
$s_R(\mathbf{v}_i^{\mathrm{en}}, \mathbf{v}_j^{\mathrm{fr}}) \geq 0.7$		≥ 0
Lucas	Lucas	Lucas
Lydia	Lydia	Lydia
marriage	mariage	mariage
Mary	Mary	Mary
master		maître
Meryton	Meryton	Meryton
Miss	Miss	Miss
mother	mère	mère
motive	motif	motif
Mr	Mr	Mr
Mrs	Mrs	Mrs
nephew	neveu	neveu
Netherfield	Netherfield	Netherfield
niece	nièce	nièce
officers	officiers	officiers
park	parc	parc
Pemberley	Pemberley	Pemberley
Phillips	Philips	Philips
politeness	politesse	politesse
poor	pauvre	pauvre
promised	promesse	promesse
refuse	refus	refus
regiment	régiment	régiment
Rosings	Rosings	Rosings
said		dit
send	envoyer	envoyer
Sir	Sir	sir
society	société	société
stranger		étrange
table	table	table
ten	dix	dix
thank	remerciements	remerciements
thousand	mille	mille
town	Londres	Londres
two	deux	deux
uncle	oncle	oncle
visit	visite	visite
way		sujet
week	semaine	semaine
Wickham	Wickham	Wickham
William	William	William
write	écrire	écrire
years	ans	ans
young	jeune	jeune

(en) English	(ko) Korean	
$s_R(\mathbf{v}_i^{\mathrm{en}}, \mathbf{v}_j^{\mathrm{ko}}) \geq 0.7$		≥ 0
amiable		상냥한
arrival	도착(到着)한	도착(到着)한
aunt		숙모(叔母)가
ball	무도회(舞蹈會)에	무도회(舞蹈會)에
Bennet	베넷(Bennet)	베넷(Bennet)
Bingley	빙리(Bingley)	빙리(Bingley)
book		책(冊)을
Bourgh		드(de)
Brighton	브라이튼(Brighton)에	브라이튼(Brighton)에
brother		오빠가
carriage		마차(馬車)를
Catherine	캐서린(Catherine)	캐서린(Catherine)
Charlotte	샬럿(Charlotte)이	샬럿(Charlotte)이
Collins		콜린스(Collins)
colonel		대령(大領)이
come	오는	오는
congratulations	축하(祝賀)해	축하(祝賀)해
cousin	사촌(四寸)	사촌(四寸)
dance	춤을	춤을
Darcy	다시(Darcy)	다시(Darcy)
daughter	딸	딸
day	시간(時間)을	시간(時間)을
de	버그(Bourgh)	버그(Bourgh)
Derbyshire	더비셔(Derbyshire)에	더비셔(Derbyshire)에
disappointment	실망(失望)을	실망(失望)을
door	문(門)을	문(門)을
elegant	우아(優雅)한	우아(優雅)한
estate		상속(相續)에
eyes	미소(微笑)를	미소(微笑)를
father	아버지가	아버지가
fear		두려워서
Fitzwilliam	피츠윌리엄(Fitzwilliam)	피츠윌리엄(Fitzwilliam)
five		다섯
forget		잊지
Forster	포스터(Forster)	포스터(Forster)
Gardiner	가디너(Gardiner)	가디너(Gardiner)
gentlemen		신사(紳士)들이
good	좋은	좋은
happy	행복(幸福)한	행복(幸福)한
Hurst		허스트(Hurst)
husband	남편(男便)이	남편(男便)이
inquiries		물었다
introduction	소개(紹介)해	소개(紹介)해
invitation	초대(招待)를	초대(招待)를

(en) English	(ko) Korean	
$s_R(\mathbf{v}_i^{\mathrm{en}}, \mathbf{v}_j^{\mathrm{ko}}) \geq 0.7$		≥ 0
Jane	제인(Jane)이	제인(Jane)이
Kitty	키티(Kitty)가	키티(Kitty)가
laugh		웃음이
library	서재(書齋)로	서재(書齋)로
Lizzy	리지(Lizzy)야	리지(Lizzy)야
long		그고 싶어
love		사랑하는
Lucas	루커스(Lucas)	루커스(Lucas)
marriage	결혼(結婚)을	결혼(結婚)을
Mary	메리(Mary)는	메리(Mary)는
Meryton		메리(Mary)턴에
Miss	양(孃)이	양(孃)이
money		돈을
mother	어머니가	어머니가
Mr		씨(氏)가
Mrs	부인(夫人)은	부인(夫人)은
neighborhood	이웃	이웃
Netherfield	네더필드(Netherfield)	네더필드(Netherfield)
	에서	에서
officers	장교(將校)들	장교(將校)들
oh	아	아
Pemberley	펨벌리(Pemberley)에	펨벌리(Pemberley)에
Phillips	필립스(Philips)	필립스(Philips)
pounds		천(千)
pretty	예쁜	예쁜
promised		약속(約束)을
read		읽어
refuse	거절(拒絶)하는	거절(拒絶)하는
regiment		연대(聯隊)가
resolved		결심(決心)을
room		방(房)을
Rosings	로징스(Rosings)에서	로징스(Rosings)에서
sat		앉아
serious	진지(眞摯)하게	진지(眞摯)하게
servant	하인(下人)이	하인(下人)이
Sir	경(卿)이	경(卿)이
thousand	파운드(pound)를	파운드(pound)를
uncle	숙부(叔父)와	숙부(叔父)와
vain	허영심(虛榮心)을	허영심(虛榮心)을
visit	방문(訪問)을	방문(訪問)을
Wickham	위컴(Wickham)	위컴(Wickham)
wife	아내가	아내가
William	윌리엄(William)	윌리엄(William)
young	젊은	젊은

1.4 Automated question answering based on Markov semantics

When we are given a document (of moderate length) and a natural language question as input, we rate and rank the sentences within the document by their relevance to the question, containing topical patterns $\mathscr{Q} = \{\mathsf{W}_{q_1}, \ldots, \mathsf{W}_{q_K}\}$. We expand the query into $\mathscr{Q} \cup \mathscr{Q}'$, a union of semantic cliques: $\mathscr{Q} \cup \mathscr{Q}' = \bigcup_{k=1}^{K} \mathscr{S}_{q_k}$. We build a localized Markov matrix $\mathbf{P} = (p_{ij})_{1 \leq i,j \leq N}$ on $\mathscr{Q} \cup \mathscr{Q}'$. We further use the Brin–Page damping [15] to form an ergodic Markov matrix $\widetilde{\mathbf{P}} = (\widetilde{p}_{ij})_{1 \leq i,j \leq N}$, where $\widetilde{p}_{ij} = 0.85 p_{ij} + \frac{0.15}{N}$.

If our question Q contains words from $\mathsf{W}_{Q_1}, \ldots, \mathsf{W}_{Q_m} \in \mathscr{Q}$ and a candidate answer A contains words $\mathsf{W}_{A_1}, \ldots, \mathsf{W}_{A_n} \in \mathscr{Q} \cup \mathscr{Q}'$ (counting multiplicities, but excluding function words and patterns with fewer than 3 occurrences in the reference document), then we assign the following entropy production score [1, (11) and Footnote 4]

$$\mathscr{F}[Q, A] := -\sum_{i=1}^{m} \sum_{j=1}^{n} \widetilde{\pi}_{Q_i} \widetilde{p}_{Q_i A_j} \log \widetilde{p}_{Q_i A_j} \tag{1.14}$$

to this question-answer pair.

In [1, Fig. 7a,b], we test our Markov language model on the WikiQA dataset. The distribution of average precisions (AP) and reciprocal ranks (RR) for individual questions in the WikiQA dataset is summarized in Fig. S2. One may wish to check Tables S11–S12 in §4 and see how our algorithm performs on each of the 1242 questions.

2 What our software does not

2.1 Non-universal statistics on short time scales

As recapitulated in §1.1, we reject text fragments shorter than or equal to the lengths of the flanking words. This is because we find such short-range features highly sensitive to grammatical rules that are specific to certain languages, thus lacking universality.

In [1, Footnote 2], we have already alluded to the highly versatile reduplications (or lack of) in different languages: German *liebe Studentinnen und Studenten* "dear (female and male) students", Malay *orang-orang* "people" (vs. *orang* "person"), and Hawaiian *wiki wiki* "very quick" (vs. *wiki* "quick"). We choose to ignore such short-range recurrence statistics in our current study, which are typically not transferrable across languages and cultures.

The numerical instability of short-range textual features can be further illustrated by a comparison of Turkish and English morphemes (adapted from [17, p. 61]):

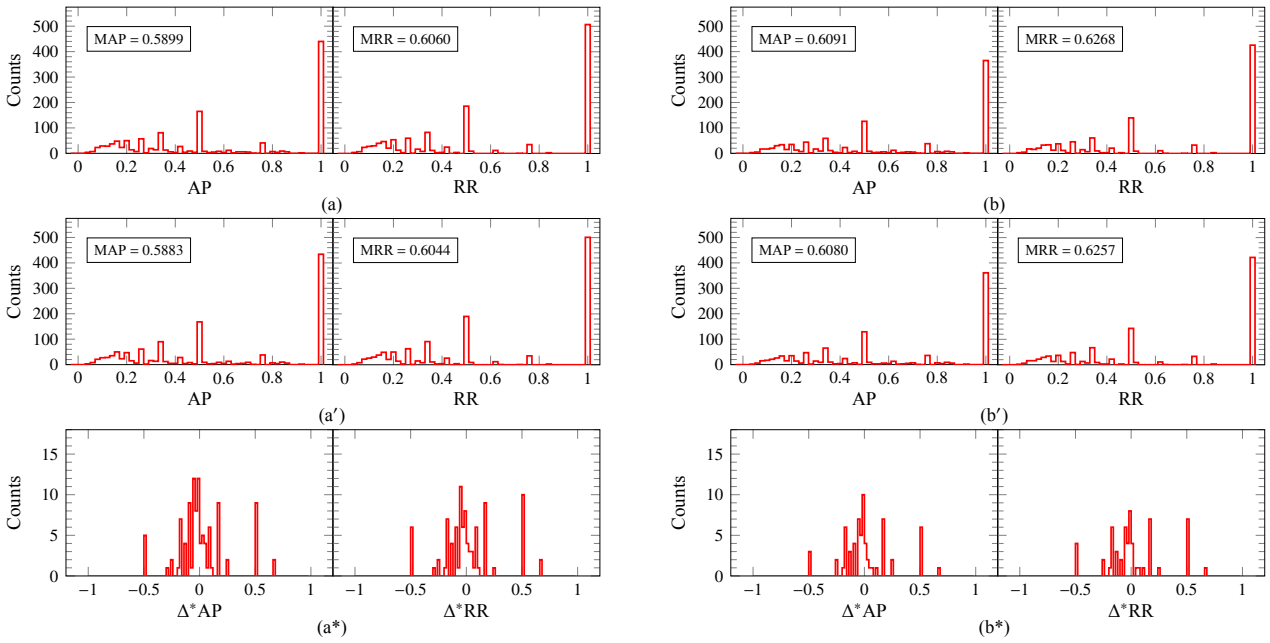


Fig. S2. Score distributions in WikiQA tests. (a) Test results on 1242 answerable questions the WikiQA dataset: histograms for average precisions (AP) and reciprocal ranks (RR) of correct answers, with their respective means displayed as *inset*. Tied scores are resolved by the McSherry–Najork algorithm [14], which effectively averages AP and RR over all possible permutations of tied entries. (b) Test results on a subset of 990 WikiQA questions (see Algorithm 3.12 in §3 for detailed screening criteria) that do not require logic inference beyond associative reasoning. (The Brin–Page damping factor 0.85 [15] is used in both panels a and b.) (a′)–(b′) Control experiments using the Bressan–Peserico damping factor 0.5 [16]. (a*)–(b*), Non-zero differences between Brin–Page and Bressan–Peserico values (upper panel minus lower panel).

Avrupalılaştırılamayacaklardansınız

You (all) are among those who will not be able to be caused to become Europeanized ≈ You (all) are among those who will not be made Europeanizable

Here, in both languages, the relevance of a morpheme roughly decays monotonically with respect to the distance from the central morpheme (Avrupa/Europe). The distance correspondence is generally stabler for larger separations, but is highly unstable when we count distances between nearest (or next-to-nearest) neighboring morphemes.

Table S7. **Length/time scales in natural language processing (NLP)**

Scale	Friederici's hierarchy [18]	Jakobson's metaphor	Rôle in NLP	Evolutionary susceptibility
short-range	phonological level	"elementary particle" [19]	acoustic encoding	climatic/environmental [20, 21, 22]
↓	lexical level	"atom"	concept tagging	cognitive/cultural/social [23, 24, 25, 26]
	sentence level	"molecule"	syntactic computing	cognitive/cultural/social [23, 24, 25, 26]
long-range	text/discourse level	"bulk material"	semantic processing	——

In our current work, we mainly concern ourselves with theories and applications for the recurrence statistics of word patterns that exhibit universal behavior, with a focus on semantic processing [27] rather than syntactic computation [28]. Practically, this means that we will suppress the fine structures in the three short scales in Friederici's hierarchy (Table S7), which vary by language typology. Similar exclusions of short-range word contacts are also implemented in the *n*-gram language model of Brown *et al.* [29].

2.2 Word disambiguation and stylistic adaptability

In theory, we expect that each Markov state corresponds to exactly one concept (a family of words that are related to each other by inflection and derivation). In practice, there are homographs (words with the identical spelling but unrelated meanings), whose denotations must be resolved through careful examination of context (either by experienced human readers, or by sophisticated algorithms in supervised learning). In numerical implementations of our model, we do not make any attempt to solve the polysemy puzzle, so we do run the risk of conflating homographs into one Markov state. In languages like Danish and Dutch, this conflation is serious enough to affect the performance of certain computational tasks related to the following

Table S8. **Parallel versions for a conversation in Chapter 1 of *Pride and Prejudice***

English	French	Russian	Finnish
"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?" Mr. Bennet replied that he **had not**.	— Savez-vous, mon cher ami, dit un jour Mrs. Bennet à son mari, que Netherfield Park est enfin loué ? Mr. Bennet répondit qu'il l'**ignorait**.	— Дорогой мистер Беннет, — сказала как-то раз миссис Беннет своему мужу, — слышали вы, что Незерфилд-парк наконец больше не будет пустовать? Мистер Беннет ответил, что он этого **не слышал**.	"Rakas Bennet", sanoi tämän arvon herran puoliso miehelleen eräänä päivänä, "oletko kuullut, että Netherfield Parkin kartano on vihdoinkin saanut vuokraajan?" Hra Bennet vastasi, ettei hän **ollut kuullut**.

Notes:
(1) See Table S1 for provenances.
(2) The underlined texts are roughly equivalent across four versions; the same can be said for the texts in **boldface**.

Table S9. **Parallel translations of the opening line in *Pride and Prejudice***

German (Werner Beyer)	(Christian Grawe & Ursula Grawe)	(Karin von Schwab)
In der ganzen Welt gilt es als ausgemachte Wahrheit, daß ein begüterter Junggeselle unbedingt nach **einer Frau** Ausschau halten muß.	Es ist eine allgemein anerkannte Tatsache, dass ein alleinstehender Mann im Besitz eines hübschen Vermögens angeblich nichts dringender braucht als **eine Frau**.	Es ist eine Wahrheit, über die sich alle Welt einig ist, daß ein unbeweibter Mann von einigem Vermögen unbedingt auf der Suche nach **einer Lebensgefährtin** sein muß.
Spanish (Ana M. Rodríguez)	(José de Urríes y Azara)	(Marciano Guerrero)
Es una verdad reconocida por todo el mundo que un soltero dueño de una gran fortuna siente un día u otro la necesidad de **una mujer**.	Es verdad universalmente admitida que un soltero poseedor de buena fortuna tiene que necesitar **una mujer**.	Es una verdad mundialmente reconocida que un hombre soltero y poseedor de una buena fortuna, debe estar en busca de **una esposa**.

Notes:
(1) See Table S1 for provenances.
(2) The underlined texts are roughly equivalent across four versions; the same can be said for the texts in **boldface**.

(common but ambiguous) words:[13]

Dutch	English gloss	(German cognate)	Danish	English gloss
meer	*more*	(*mehr*)	*så*	*saw* (past of *see*)
	lake	(*Meer*)		*so*
				(to) *sow*
vroeg	*asked*	(*frug*)	*ved*	*at*
	early	(*früh*)		*knows*
				wood

We refer our readers to §9.2 for a discussion of homographs in Korean.

Besides homographs, there are three more subtler noise sources that may negatively impact the translational invariance of the Markov spectrum, hence our automated word translation via cross-lingual matching of semantic fingerprints.

The first kind of noise is attributed to free variations of translators (Tables S8 and S9). As shown in Table S8, translators may take the liberty of using free (as opposed to literal) interpretations: the underlined texts in the three translations essentially read "Mrs. Bennet said to her husband one day", which paraphrases the interpersonal relationship expressed in the original English text; the bold texts in the three translations all explicitly supply the negated verb ("had not known"/"had not heard"), which is omitted in the English original. In Table S9, all the underlined texts correspond to "a truth universally acknowledged" in the English original and all the phrases in boldface mean "a wife", despite nuances in their wording and phrasing.

The second kind of noise is due to algorithmic imperfection in morphological classifications of words into word patterns. There are irregularities in word morphologies that cannot be fully covered by our automated word clustering algorithms described in Part III. Furthermore, the heterogeneous etymological origins[14] of the English language may bring us additional challenges. For example, our programs do not recognize *heart* and *cordial* as morphologically related in English, even though their German or Hungarian translations display such relationship manifestly:

English	German	Hungarian
heart	*Herz*	*szív*
hearty, *cordial*	*herzlich*	*szíves*
heartily, *cordially*	*herzlich*	*szívesen*

The third kind of noise is due to our insufficient treatment of space, time and causality in the lexicosemantics of verbs (see [30, Chap. 4] or [31, Chap. 5]). For the time being, we have swept all prepositions (which are vital to the space component of verbs) into the class of function words, and have ignored them in our text mining tasks. Thus, our numerical translation algorithm cannot handle polysemic verbs like *go*, *make*, *take*, *turn*, whose meanings are best clarified by subsequent prepositions

[13] In modern German, *Meer* means "sea". In archaic German, *Meer* means "lake". In modern German, the strong conjugation *frug* for first- or third-person singular preterite is rare, while the weak conjugation *fragte* is common. We have included these German cognates, to serve as visual aids to their Dutch counterparts.

[14] English lexicon possesses a native Germanic stock (exemplified by *hearty*, *heartily* and so forth), but is simultaneously under strong Latinate influence (shown in near-synonyms of the aforementioned words *cordial*, *cordially*), via Norman French. A more challenging scenario appears in Korean word families with disparate (Chinese/English/Korean) origins. See §9.2 for a discussion.

(if present). We have chosen to ignore prepositions (in either isolated form or as separable prefixes[15]) in our numerical text analysis, because they do not exhibit universal behavior, in at least two senses: (1) Short-range relations between verbs and prepositions (like *takes up*) in English may translate into long-range relations in certain Germanic languages (like *nimmt ... auf* in German and *neemt ... op* in Dutch, where the written space occupied by the ellipsis can be arbitrarily long). (2) The rôles of locative prepositions (like *on*, *in* and *from*) are largely subsumed by certain case endings in Uralic languages (such as Finnish and Hungarian). In our work, we not only ignore prepositions, but also strip away case endings in a stemming procedure (Part III). We believe that such simultaneous treatments of prepositions and case endings are consistent: they generate similar levels of coarse-graining in words across different language families. In addition to neglecting prepositions, we have designed our stemming procedure so as to merge verb forms in different tenses and aspects (time component), but it may also conflate causative and non-causative forms (causality component), in a fashion that is not necessarily consistent across all languages. For example, our stemming algorithm treats Finnish *syödä*/Korean 먹다 "(to) eat" (non-causative) and Finnish *syöttää*/Korean 먹이다 "(to) feed" (causative) as the same verb in their respective language, but not so for their English counterparts.

The language-specific usage of function words is another practical hindrance to lossless translation, but this hindrance does not contribute to our Markov spectrum (if we can ignore effects from pronominal coreferences, as discussed below) for content words. As a glance of the highly diverse typologies of function words found in different languages, we say a few more words about the gender agreement of pronouns. All the 8 modern Indo-European languages under our consideration in Part III distinguish "he" from "she" when referring to humans as subjects in a sentence, while the 5 non-Indo-European languages do not.[16] Gender agreement of possessive pronouns is a subtler issue for Indo-European languages. Expanding the example given by Willim and Chomsky [32, p. 100], we have

English:	*John told Mary about his father.*	*John told Mary about her father.*
French:	*Jean a parlé de son père à Marie.*	*Jean a parlé de son père à Marie.*
German:	*Johann hat Maria von seinem Vater erzählt.*	*Johann hat Maria von ihrem Vater erzählt.*
Polish:	*Jan opowiadał Marii o swoim ojcu.*	*Jan opowiadał Marii o jej ojcu.*
English:	*John told Mary about his mother.*	*John told Mary about her mother.*
French:	*Jean a parlé de sa mère à Marie.*	*Jean a parlé de sa mère à Marie.*
German:	*Johann hat Maria von seiner Mutter erzählt.*	*Johann hat Maria von ihrer Mutter erzählt.*
Polish:	*Jan opowiadał Marii o swojej matce.*	*Jan opowiadał Marii o jej matce.*

Here, the pronouns *his*, *her*, *seinem*, *seiner*, *ihrem*, *ihrer* and *jej* agree with the gender of the possessor; meanwhile the pronouns *son*, *sa*, *seinem*, *seiner*, *ihrem*, *ihrer*, *swoim* and *swojej* agree with the gender of the possessed. Automated resolution of such agreements is beyond the scope of our current research, as this language-specific problem occurs on a different neuro-linguistic scale from that of our concern. However, if pronominal coreferences become ambiguous enough after literal translation, then the translator might choose to supply personal names in place of pronouns, and such rephrasing may partly affect the performance of our Markov algorithms. (See Example 4.15.2 and Fig. S3e.)

2.3 Causal inference and deductive reasoning

Admittedly, our semantic cliques (see [1, §2.3.2] and §1.2 above) have some limitations. At relatively low computational cost, our criterion for semantic dependence [see (1.5)] does not distinguish between causal [33, 34] and non-causal relations, nor does it generate semantic webs with hierarchical topologies [35, 36]. For comprehension tasks requiring high-precision causal inference and relationship mining, Bayesian networks [33, 34] and persistent homologies [35, 36] offer more reliable guidance than our approach.

The lack of causal inference and deductive reasoning in our question-answering machine places a cap on its MAP and MRR scores ([1, Fig. 7b] and Fig. S2). No matter how we improve the stemming/clustering algorithms for English (§4), there are certain fractions of WikiQA questions that our method cannot handle properly (Table S12). Typical failures include, but are not limited to the following categories (see Table S10 for evaluations of our performance):

- Our semantic cliques do not always enable us to address causation questions (like *What causes ...?*) correctly, with only about a fifty-fifty chance of hitting the right answer. These cliques indicate association and connection between concepts, which may or may not be of causal nature.

- Our semantic cliques do not automatically sieve answers with numbers (spelt out in English or Arabic numerals), even if the question (such as *How big ...? How long ...? How many ...? How old ...? What year ...?*) elicits one. Perhaps

[15] For example, the prefixes (*ab*/*af*, *auf*/*op*, and *zu*/*toe*) in German *abnehmen*/Dutch *afnemen* "(to) decrease", German *aufnehmen*/Dutch *opnemen* "(to) take up", and German *zunehmen*/Dutch *toenemen* "(to) increase" must be separated from the verb stems in certain conjugated forms. Note however, that the prefixes in English words "decrease" and "increase" are not separable as prepositions.

[16] In some rare circumstances, the lack of gender-specific pronouns may also hamper topic analysis in our numerical experiments. (See Example 4.15.2 and Fig. S3e.)

Table S10. **Distributions of hits and misses in certain types of questions in our WikiQA experiments**

Question type	#Hits	#Misses
What caused ...?	1	1
What causes ...?	4	4
What is caused ...?	1	0
How big ...?	0	2
How deep ...?	0	1
How long ...?	3	5
How many ...?	19	51

Question type	#Hits	#Misses
How much ...?	5	9
How often ...?	0	2
How old ...?	4	5
What became of ...?	0	1
What happened to ...?	0	4
What year ...?	7	14
When ...?	46	67

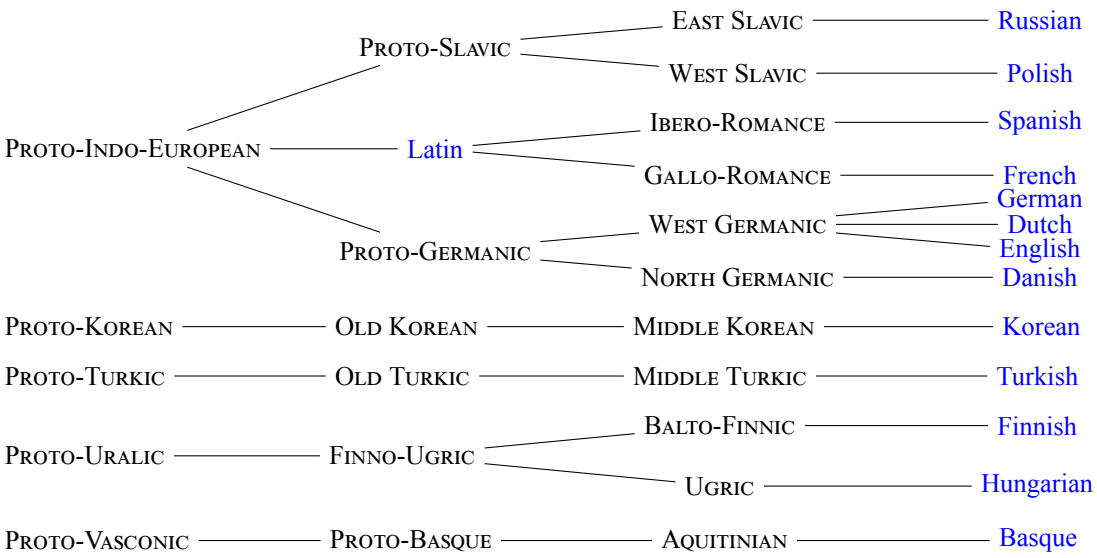
the statistical behavior of recurring numbers in a text document does not neatly fit into our Markov model for word patterns, so the semantic cliques do not easily detect numbers as synonyms to certain question words.

- Our semantic cliques stumble over questions that ask for superlatives, either explicitly or implicitly. Some of these questions (*What became of ...? What happened to ...?*) can be very subtle, as they require finding the most recent date out of a list, going beyond the reach of our Markov model.

Part II
Protocols for text cleansing

After a brief summary of notations (§3.1), we present in this Part cleansing protocols for certain articles in the the English Wikipedia (§3.2) and electronic books written in 14 languages spoken in Europe and Asia (§3.3), in preparation for word clustering and word translation experiments in Part III.

Hereafter in Parts II and III, aside from English (§4), we pick Danish, German and Dutch as representative Germanic languages (§5); Spanish, French and Latin[17] as representative Romance languages (§6); Polish and Russian as representative Slavic languages (§7); Finnish and Hungarian as representative Uralic languages (§8); Basque, Korean and Turkish as representatives from various other language families (§9). A putative evolutionary history of these languages is sketched below:



From now on, the languages other than English are first sorted alphabetically according to their genealogical affiliation (Germanic, Romance, Slavic, Uralic, "Various"), and then sorted within their respective affiliation according to their ISO codes [*e.g.* Spanish `(es)` precedes French `(fr)`][18].

All these Asian and European languages under investigation have clearly defined word boundaries (typically marked by spaces and punctuation marks, and additionally by apostrophes and hyphens in the case of French) and rich word morphologies

[17] Strictly speaking, Latin is not a Romance language, but an ancestor to all the modern Romance languages. Since a modern Romance language, Romanian, still shares a lot of conservative features with Latin, it may not hurt to label Latin as a Romance language.

[18] The individual sections from §5 to §9 can be read in any order, depending on need. To help our readers navigate through these protocols, we have reproduced certain explanatory texts from one section to another.

(*e.g.* adjective comparisons, noun declensions and verb conjugations). Inspired by the sequence alignment procedure [37, 38] in bioinformatics, we devise algorithms that employ string matching to find affinity between words, with a limited amount of input for grammatical rules.

The Germanic, Romance and Slavic languages all belong to the Indo-European family, which distinguishes from the other language families under consideration, both lexically and syntactically. The languages within the same affiliation share a lot of common features in vocabulary and syntax. We note however that English, though officially a Germanic language, has borrowed extensively from Latin and (Norman) French, so it has blended characteristics of both Germanic and Romance subgroups. In the light of this, we have singled out the algorithmic treatment of English texts in §4.

Our word clustering algorithms are based on Porter stemming [39, 40], along with extensive additions that accommodate to irregularities in word morphologies. Since regular and irregular word morphologies are handled by different neurobiological mechanisms in the human brain [41, 42], we are not going to unify their treatment under a single algorithmic framework. Regular morphologies will be identified by substitution rules (as in Porter stemming), while irregular morphologies in each language will be accommodated by a finite (though sometimes long) list of exceptions. It has been observed that grammatical irregularities tend to be associated with highly common words [43]: the frequent uses of these words make their peculiarities harder to be forgotten. It is our hope that, via many explicitly coded "exceptions to general rules", we can achieve higher precision in text mining tasks related to those highly frequent yet highly irregular words.

3 String manipulations and text normalizations

3.1 Notational conventions

In the source codes accompanying this work, we have implemented our text mining algorithms in *Mathematica* (including both the cleansing procedures in Part II and the clustering procedures in Part III). However, we hope that readers of this document could easily adapt our algorithms to other programming languages. Towards this end, we will present our methods of text processing using some standard notations in linguistic analysis (applicable to both Parts II and III), as declared below.

Definition 3.1 (String length)**.** The length of a text string $\hat{\sigma}$ is a non-negative integer that equals the number of Unicode characters used to write out the string, denoted by $\ell(\hat{\sigma})$.

The empty string $\varnothing$ (not to be confused the Danish letter Ø/ø) is the only text string with length zero: $\ell(\varnothing) = 0$.

For a positive integer n less than or equal to $\ell(\hat{\sigma})$, the nth position in string $\hat{\sigma}$ is denoted by $\hat{\sigma}^{\{n\}}$, and the first n characters in string $\hat{\sigma}$ is denoted by $\hat{\sigma}^{[n]}$. For a non-empty string $\hat{\sigma}$, its "first character" is $\hat{\sigma}^{\{1\}}$ and its "last character" is $\hat{\sigma}^{\{\ell(\hat{\sigma})\}}$. For the empty string, its "first character" and "last character" are both defined as the empty string. In general, the last character of a string $\hat{\sigma}$ is denoted by $\Omega(\hat{\sigma})$.

The notation $\hat{\sigma}^{-1}$ stands for the reverse of the string $\hat{\sigma}$. In particular, we have $\varnothing^{-1} = \varnothing$; when $\ell(\hat{\sigma}) > 0$, we have $(\hat{\sigma}^{-1})^{\{n\}} = \hat{\sigma}^{\{\ell(\hat{\sigma})+1-n\}}$ for all integers in the range $1 \le n \le \ell(\hat{\sigma})$. □

Definition 3.2 (String spelling and string alternatives)**.** For all the languages written in either the Latin or the Cyrillic script, letters (including apostrophes and hyphens) within a text string in an algorithm is always spelt out in slant typeface, such as *word* (English) and *слово* (Russian). Boxed characters set in the `typewriter` typeface are taken verbatim. A boxed expression in red, in the form of `U+xxxx`, refers to the Unicode character with hexadecimal code `xxxx`. The expression $\hat{\sigma}^{+}$ (resp. $\hat{\sigma}_{-}$) results from converting $\hat{\sigma}$ to upper (resp. lower) case.

The notation for "alternatives" $\hat{\sigma}_1|\cdots|\hat{\sigma}_n$ refers to an arbitrary text string in the set $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_n\}$. Parentheses in string spellings, usually in the form of $\hat{\sigma}(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$, are used to enclose alternative portions that are not shared by a group of words.

The rules above also apply to algorithms for the Korean language, except that slant typefaces will not be used, and the native Korean alphabet does not distinguish upper and lower cases. □

Example 3.2.1. The boxed pattern `, ` is interpreted literally, as a comma followed by a white space. The pattern `(a|b)` is a concatenation of five Unicode characters, while $(a|b)$ refers to either member of the set $\{a, b\}$.

Example 3.2.2. For English, the string pattern *word*$(\varnothing|s|$'s$|s')$ refers to any text string in the set {*word*, *words*, *word's*, *words'*}. For Russian, the string pattern *слов*$(о|а|\varnothing)$ refers to any text string in the set {*слово*, *слова*, *слов*}. The expression $(word(\varnothing|s|$'s$|s'))^{+}$ has the same effect as $WORD(\varnothing|S|$'S$|S')$, while $((СЛОВ(О|А|\varnothing))_{-}$ is equivalent to *слов*$(о|а|\varnothing)$.

Definition 3.3 (Multiple occurrences of string patterns)**.** For a string pattern $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$, the notation $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_m$ (resp. $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_{m_0}$) denotes one (resp. zero) or more successive occurrences of the same pattern in a text string, and the notation $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_{[n_1,n_2]}$ refers to between n_1 and n_2 (inclusive) consecutive appearances of the pattern $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$.[19] When $\nu = n_1 = n_2$, we also write $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_{[\nu,\nu]}$ as $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_{\times\nu}$, to denote exactly ν repeats of the pattern $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$. □

[19] In *Mathematica* codes, $(\hat{\sigma})_m$ is represented by `(`σ`)..` or `Repeated[`σ`]`, $(\hat{\sigma})_{m_0}$ by `(`σ`)...` or `RepeatedNull[`σ`]`, and $(\hat{\sigma})_{[n_1,n_2]}$ by `Repeated[`σ`, {n1, n2}]`.

Example 3.3.1. The pattern $(d|e)_m$ may be matched by strings *d*, *e*, *dd*, *de*, *ed*, *ee*, *ddd*, *dde*, *ded*, *dee*, *edd*, *ede*, *eed*, *eee* and so on. If we consider the pattern $(d|e)_{m_0}$, then the empty string $\varnothing$ also forms a match. The pattern $(d|e)_{[0,2]}$ is only matched by the strings $\varnothing$, *d*, *e*, *dd*, *de*, *ed*, *ee*.

Definition 3.4 (Wildcard symbols)**.** We write **X** for a generic text string of zero or positive length. At times, we may restrict **X** to a specific set, or may use subscripts (like $\mathbf{X}_1$ and $\mathbf{X}_2$) to put the generic string in context.

We write **V** and **C** for generic vowels and consonants, respectively. The classification schemes for vowels and consonants may depend on language and context. Like the case of **X**, the exact identity of the wildcard symbol **V** or **C** may vary from entry to entry, line to line.

The symbol $\hat{\chi}$ denotes a single letter character (a Unicode character other than digit or punctuation[20]). When we repeat it *n* times (as in $\underbrace{\hat{\chi}\cdots\hat{\chi}}_{n \text{ times}}$), we mean exactly *n* letter characters (not necessarily identical) appearing successively in a text string.

The symbol $\hat{\xi}$ stands for a single letter character (which serves the same function as $\hat{\chi}$ when used alone). However, whenever $\hat{\xi}$ is repeated, it always stands for the same character. □

Example 3.4.1. The string pattern **X***aint* (**X**∈{*rt*, *mpl*, *tr*}) is synonymous with the notation (*rt*|*mpl*|*tr*)*aint*, denoting any member among {*rtaint, mplaint, traint*}. For succinctness, we also abbreviate **X***aint* (**X**∈{*rt*, *mpl*, *tr*}) in a superscript notation, as $^{\mathbf{X}\in}(rt|mpl|tr)aint$. It is sometimes useful to specify the identity of a wildcard symbol by the set it does not belong to, such as $\hat{\chi}e$ ($\hat{\chi} \notin \{e, i\}$) denotes a string pattern where the letter *e* does not follow a letter *e* or *i*. In such cases, we will introduce a shorthand using subscript: $_{\hat{\chi}\notin}(e|i)e$, or use an overline[21] for the negation of a parenthesized string pattern: $^{\hat{\chi}\in}\overline{(e|i)}e$.

Example 3.4.2. The multiplicity notation (introduced in Definition 3.3) also applies to negated string patterns. For instance, in the notation $^{\mathbf{X}\in}\left(\overline{(e|i)}_{m_0}\right)e$, the wildcard **X** is restricted to zero or more repeated occurrences of any letter other than *e* or *i*. When *n* repetitions of the same letter character are intended, we write $\hat{\chi}_{\times n}$. Thus $\hat{\chi}_{\times 2}$ refers to double letters, which contrasts with $\hat{\chi}\hat{\chi}$ (two letter characters, either identical or distinct).

Example 3.4.3. In Finnish, there are word endings in the form of (*ahan*|*ehen*|*ihin*|*ohon*|*ähän*|*öhön*). Such patterns can be represented by $\hat{\xi}h\hat{\xi}n$ for $^{\hat{\xi}\in}(a|e|i|o|ä|ö)$.

Example 3.4.4. The string pattern **CVC***e* in English, when **V** and **C** are appropriately defined, usually implies that the **V** in question is pronounced as a long vowel.

Example 3.4.5. We write $\hat{\alpha} = \hat{\beta}$ if the two strings in question match exactly. With string alternatives and wildcard symbols, we can interpret many other string equations and string inequalities. For example, $\hat{\beta} = \hat{\alpha}\mathbf{X}$ means that string $\hat{\beta}$ can be constructed from string $\hat{\alpha}$, together with zero or more trailing characters (which is usually a linguistically meaningful scenario, where $\hat{\beta}$ is a morphological derivative of $\hat{\alpha}$); the equation $\hat{\beta} = \hat{\alpha}(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$ means that at least one of the *n* statements $\hat{\beta} = \hat{\alpha}\hat{\sigma}_1, \ldots, \hat{\beta} = \hat{\alpha}\hat{\sigma}_n$ holds (where $\hat{\sigma}_1, \ldots, \hat{\sigma}_n$ can be regarded as *n* candidate suffixes); the string inequality $\hat{\beta} \neq \hat{\alpha}(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)$ means that none of the *n* statements $\hat{\beta} = \hat{\alpha}\hat{\sigma}_1, \ldots, \hat{\beta} = \hat{\alpha}\hat{\sigma}_n$ is true.

Definition 3.5 (Partial and full matches)**.** A string $\hat{\sigma}$ written in the plain may occur anywhere in a word. The underlined notation $\underline{\hat{\sigma}}$ refers to a string that matches the entire word exactly.

The notation $\hat{\sigma}\sim$ (resp. $\sim\hat{\sigma}$) refers to any word that "starts" (resp. "ends") with the string $\hat{\sigma}$. Here, an isolated word may "start" or "end" at external and internal word boundaries. The external word boundaries are the two ends of the text string representing the isolated word, while the internal word boundaries occur at any non-letter character (apostrophe, digit, hyphen, etc.) in the string.[22] □

Definition 3.6 (Substitution rules)**.** When we say "do $\hat{\sigma} \to \hat{\sigma}'$", we mean to replace any occurrence(s) of the string $\hat{\sigma}$ in a word by a new string $\hat{\sigma}'$. Here in the substitution rule, the string $\hat{\sigma}$ may also be augmented with wildcard symbols and partial/full match notations. The letters behind the wildcard symbols are left intact in substitutions, if $\hat{\sigma}'$ contains the respective wildcard symbols.

When we say "do $\hat{\sigma}_1 \to \hat{\sigma}'_1, \ldots, \hat{\sigma}_n \to \hat{\sigma}'_n$" on a particular word, we mean to perform *n* types of substitution tasks in a single sweep — there might be multiple replacements in a single word, but the replacements do not overlap each other.

Instead of saying "do $\hat{\sigma}_1 \to \hat{\sigma}'_1, \ldots, \hat{\sigma}_n \to \hat{\sigma}'_n$", we may also tabulate the substitution rules as follows:

$\hat{\sigma}_1$	$\cdots$	$\hat{\sigma}_n$
$\hat{\sigma}'_1$	$\cdots$	$\hat{\sigma}'_n$

[20] Roughly speaking, this corresponds to `LetterCharacter` in *Mathematica* for versions 10.3 and higher. Earlier versions of *Mathematica* did not treat some symbols in non-Latin writing systems as `LetterCharacter`.

[21] In general, we avoid applying the overline directly to letters, such as $\bar{e}$, because this may be confused with the macron diacritic (as in the long vowel markings $\bar{a}$, $\bar{e}$, $\bar{\imath}$, $\bar{o}$, $\bar{u}$, $\bar{y}$ employed in Latin dictionaries). When $_{\hat{\chi}\notin}(e)$ is intended, we write $^{\hat{\chi}\in}\overline{(e)}$.

[22] Both external and internal word boundaries are represented by `WordBoundary` in *Mathematica*.

where the strings in the shaded entries are replaced by their counterparts that lie immediately below. The tabulation is particularly helpful when n is large and/or some structures in the substitution rules are better visualized in table form. □

Example 3.6.1. Application of the substitution rule "do $\sim s\rightarrow\varnothing$" will remove the plural markers in the word *daughters* as well as the compound *daughters-in-law*. If we "do $r\sim\rightarrow R$" on the word *rock-and-roll*, we obtain *Rock-and-Roll*. If we "do $ass\rightarrow\varnothing$, $sas\rightarrow\varnothing$" on the word *assassinate*, we obtain *inate*, as a result of $assassinate \xrightarrow{\text{remove } sas} assinate \xrightarrow{\text{remove } ass} inate$.[23]

Example 3.6.2. The following tabulated substitution rules

$\sim(go(es\|ing\|ne)\|went)$	*taught*	$\mathbf{X}^{\epsilon}(\varnothing\|in\|mis\|off\|out\|over\|p\|under)led$
go	*teach*	$\mathbf{X}lead$

regularize some verb forms in English.

Definition 3.7 (NW and SW)**.** For two text strings $\hat{\sigma}_1$ and $\hat{\sigma}_2$, the output of the function $\mathrm{NW}(\hat{\sigma}_1,\hat{\sigma}_2)$ [resp. $\mathrm{SW}(\hat{\sigma}_1,\hat{\sigma}_2)$] is a string pattern that results from the sequence alignment from $\hat{\sigma}_1$ to $\hat{\sigma}_2$ according to the Needleman–Wunsch (resp. Smith–Waterman) algorithm. □

Example 3.7.1. $\mathrm{NW}(keep, kept) = \mathrm{SW}(keep, kept) = k[e, \varnothing]ep[\varnothing, t]$, where the bracketed portions show mismatches of the two strings. Here, in the presentation of the output string pattern, we have used $[e, \varnothing]$ and $[\varnothing, t]$ to show the respective contributions from the string $\hat{\sigma}_1 = keep$ and $\hat{\sigma}_2 = kept$. Unlike the situation in notation for "alternatives" $(\hat{\sigma}'_1|\cdots|\hat{\sigma}'_n)$, the arguments within the square brackets no longer commute, and must reflect the order in the input arguments of the functions $\mathrm{NW}(\hat{\sigma}_1,\hat{\sigma}_2)$ and $\mathrm{SW}(\hat{\sigma}_1,\hat{\sigma}_2)$.

Example 3.7.2. The outputs of the functions $\mathrm{NW}(\hat{\sigma}_1,\hat{\sigma}_2)$ and $\mathrm{SW}(\hat{\sigma}_1,\hat{\sigma}_2)$ may not necessarily be identical. For example, $\mathrm{NW}(infer, inferior) = infe[\varnothing, rio]r$ while $\mathrm{SW}(infer, inferior) = infer[\varnothing, ior]$.

3.2 Normalizations of Wikipedia pages and WikiQA dataset

Each question in the WikiQA dataset [44] is associated with a unique Wikipedia page. We need to pick a sentence from a small subset (usually just the opening paragraph, as provided by the WikiQA dataset) of this Wikipedia page that best answers the question. In the main text of this research, we use the single Wikipedia page (in full, not just the opening paragraph) as the sole training source of our artificial neural network, before we score and screen individual sentences in the candidate answer pool.

To construct the knowledge base used in this work, we downloaded the static dump of English Wikipedia dated Sept. 1, 2015 (file name: `enwiki-20150901-pages-articles-xml.bz2`) from the URL `https://meta.wikimedia.org/wiki/Data_dump_torrents#enwiki` and saved it as an XML file. We chose this static dump because its time stamp is close to the publication date of the WikiQA dataset [44].

Once we have located the exact Wikipedia page suitable for a particular WikiQA task, we need to process the raw XML code of such a page, to generate (nearly human-readable) plain text strings. This normalization procedure is described below.

Algorithm 3.8 (Normalization of English Wikipedia Articles)**.** *Let* $\mathbf{D} = (0|1|2|3|4|5|6|7|8|9)$ *be any one of the digit characters, and* $\mathbf{W} = (\hat{\chi}|\mathbf{D})$ *be either a letter character or a digit character. The XML codes of a page from the filtered knowledge base are normalized by sequential applications of the following substitution rules, where all the boxed portions are spelt out in verbatim mode.*

(1) *Do* `&` → `&`.

(2) *Replace the shaded patterns by their respective counterparts that lie immediately below:*

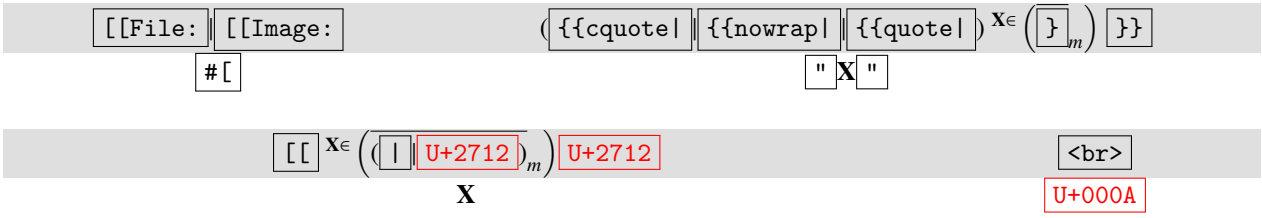
` `	`<`	`>`	`&quot;`	`]]`
∅	<	>	" = U+0022	U+2712

(3) *Replace*

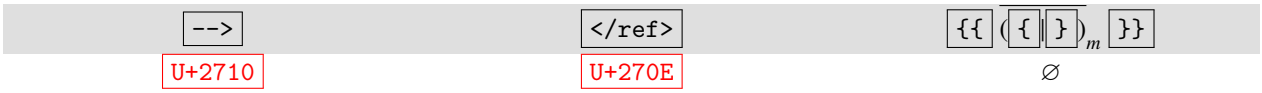
[23] Written in *Mathematica* codes, these statements amount to the following results:

```
StringReplace["daughters", "s"~~WordBoundary :> ""] yields "daughter";
StringReplace["daughters-in-law", "s"~~WordBoundary :> ""] yields "daughter-in-law";
StringReplace["rock-and-roll", WordBoundary~~"r" :> "R"] yields "Rock-and-Roll";
StringReplace["assassinate", {"ass" :> "", "sas" :> ""}] yields "inate".
```

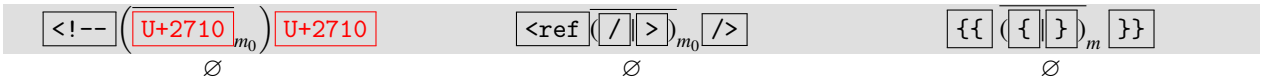
Note in particular that in the last example, the substitution rules `{"ass" :> "", "sas" :> ""}` and `{"sas" :> "", "ass" :> ""}` will generate the same output.



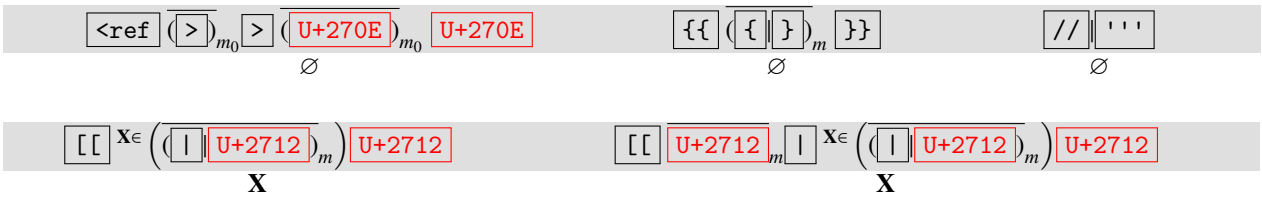
(4) *Replace*



(5) *Replace*

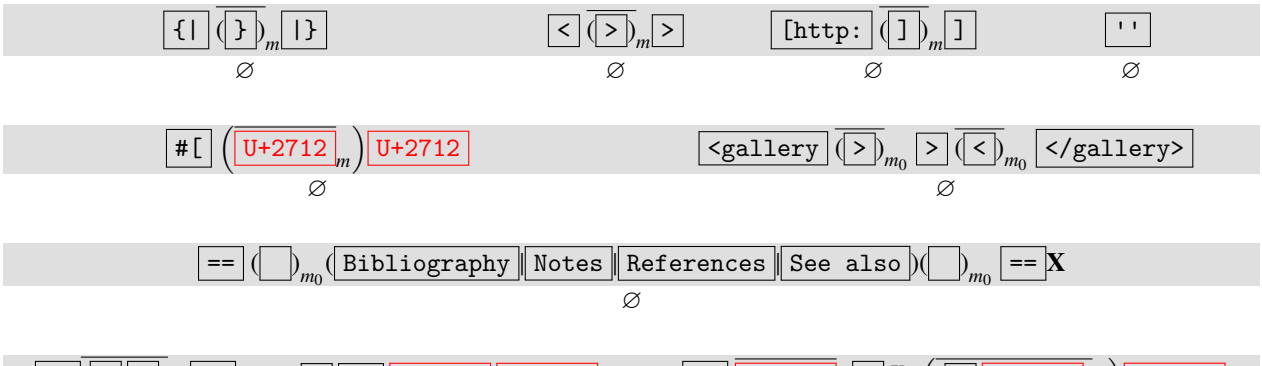


(6) *Replace*



where `//` *denotes two consecutive appearances of* SOLIDUS `U+002F` *and* `'''` *denotes three consecutive appearances of* APOSTROPHE `U+0027` .

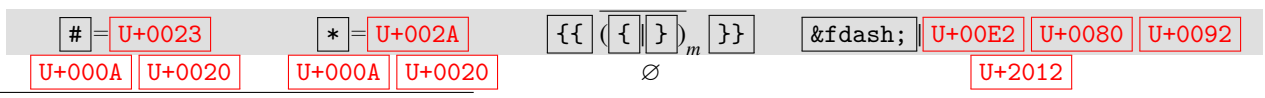
(7) *Replace*[24]



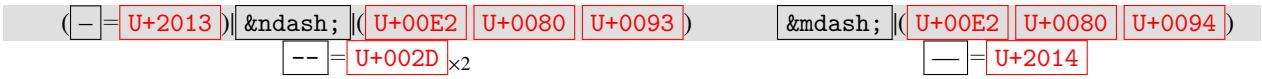
and also delete the string pattern $\hat{T}$**X** `text/x-wiki` *, where* $\hat{T}$ *is the text string matching the title of the Wikipedia page.*

(8) *Delete the string pattern* `{{` `{` `}` `}}` .

(9) *Replace*



[24] Here, the string pattern `==` `Bibliography` `Notes` `References` `See also` `==` must be deleted along with anything thereafter, because the wildcard symbol **X** does not reappear in the destination string.



(10) *Do* $\overline{\mathbf{W}}$ U+0027 → ␣.

(11) *If the text string so far ends with* **W**, *then append* . *to it; otherwise, leave it as is.*

(12) *Call the result so far as string* P_0. *Denote the input WikiQA question by* Q *and the Wikipedia page title by* T *(both of which must be "corrected" by Algorithm 3.11 below). If* $\mathsf{P}_0 = \varnothing$, *then send* P = ␣ *as output; otherwise send* P = Q U+000A×2 T U+000A×2 P_0 U+000A×2 Q␣ *as output.*

Here, in the algorithm above, we have tried our best to eliminate structured data (infoboxes, references, as well as hyperlinks to other pages and external files) from the XML file, and restore all clickable links to normal text. Some replacements are done in multiple sweeps to accommodate to nested structures in Wikipedia XML files. However, we note that due to the versatility of Wikipedia XML files, our cleansing algorithm may still have some blind spots. In practice, this heuristic algorithm handles most Wikipedia pages well, so long as there are no special characters in hyperlinks and there is no extensive coverage of TEXnical content.

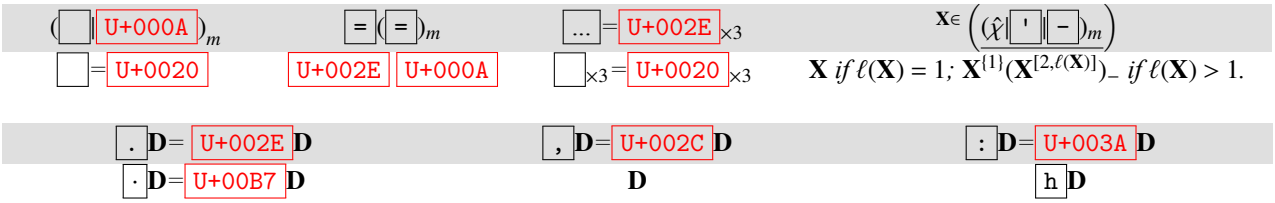
The WikiQA dataset contains some Wikipedia pages that are simply lists or tables. If we process such Wikipedia pages with Algorithm 3.8, then we will end up with (almost) an empty string. To compensate for this, we generate backup versions of Wikipedia pages, using the algorithm below.

Algorithm 3.9 (Backup version of Wikipedia pages)**.** *Define* B = Q U+000A×2 T U+000A×2 A U+000A×2 Q␣, *where* A *is the concatenation of all the candidate answers (≈ the opening section of the accompanying Wikipedia page, by design of the WikiQA dataset), riffled by* ␣. *If* $\ell(\mathsf{P}) < \ell(\mathsf{B})$, *then redefine* P *as the backup version* B. *If* T *matches* List of **X**, *regardless of letter case, then redefine* P *as* B␣B␣B.

After processing a Wikipedia page by Algorithms 3.8 and 3.9, we arrive at a text string P, which will be suitable for sentence extractions later. To facilitate the measurement of gaps between words, we need to regularize spacing, punctuation and capitalization, as indicated below.

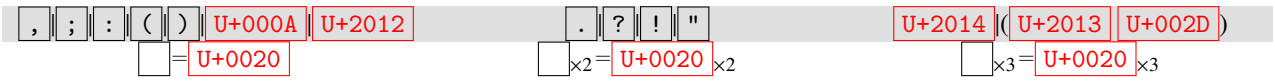
Algorithm 3.10 (Regularisation of Text String)**.** *Let* **D** = (0|1|2|3|4|5|6|7|8|9) *be any one of the digit characters, and* **W** = ($\hat{\chi}$|**D**) *be either a letter character or a digit character. A processed Wikipedia page must be regularized by the following substitution rules before temporal statistics are computed.*

(1) *Replace*



where $\mathbf{X}^{\{1\}}(\mathbf{X}^{[2,\ell(\mathbf{X})]})_{-}$ *reduces* **X** *to lower case except the initial character.*[25]

(2) *Replace*



(3) *Prepend a single white space character* U+0020 *to the entire text string.*

(4) *Record the positions of all the* U+0020$_m$ *patterns.*

(5) *Split the text string at* U+0020$_m$, *to obtain an ordered list of separate words.*

Upon conclusion of the aforementioned algorithm, we have effectively obtained the start and end positions of every word appearing in the normalized Wikipedia page P.

After scrutinizing the WikiQA dataset, we have detected a few typographical flaws:

[25] This means that the capitalization status of a word is determined by its initial character: *eBay* → *ebay*, *ELIZABETH* → *Elizabeth*, *McDonald* → *Mcdonald* etc.

- A few Wikipedia page names listed in WikiQA do not match our record.
- There are non-standard uses of capitalizations and diacritics.
- There are non-standard abbreviations that cannot be inferred from the related Wikipedia pages.
- There are other kinds of misspellings (including unnecessary use of foreign language) in the questions that do not match the information available from the related Wikipedia pages.

Since our purpose in the current research is to test a numerical question answering machine, rather than an automated spell checker, we decide to normalize the WikiQA dataset according to the following *ad hoc* procedures.

Algorithm 3.11 (Normalization of WikiQA page names and questions)**.** *We normalize the reference Wikipedia page names by the following replacements:*

Automatic Document Feeder	*High-Sticking*	*Judas (song)*	*Julius caesar*
Automatic document feeder	*High-sticking*	*Judas (Lady Gaga song)*	*Julius Caesar*
What It Takes (song)		*World Trade Center*	
What It Takes (Aerosmith song)		*World Trade Center (2001* U+00E2 U+0080 U+0093 *present)*	

We use the following heuristics to determine the correct capitalizations and diacritics in a question Q*, by checking it against all the candidate answers (≈ the opening section of the accompanying Wikipedia page, by design of the WikiQA dataset)* A*:*

(1) *Take the union of all the words in* A*, and call it* $\mathscr{A}$*. Remove diacritics from* $\mathscr{A}$*, and call it* $\mathscr{A}'$*. Use set-theoretic complements to define* $\mathscr{D} = \mathscr{A}' \setminus \mathscr{A}$ *and* $\mathscr{D}_0 = \mathscr{A} \setminus \mathscr{A}'$*.*[26] *Define* $\mathscr{A}_-$ *(resp.* $\mathscr{D}_-$*) as the lowercase form of* $\mathscr{A}$ *(resp.* $\mathscr{D}$*).*

(2) *For each word in* Q*, convert it to lower case, remove diacritics if any, call the result* $\hat{q}^*$*, and then do the following:*

(2.1) *If* $\hat{q}^*$ *matches* $\mathscr{A}_-(\varnothing|s|'s|s')$ *and does not match* $\mathscr{A}(\varnothing|s|'s|s')$*, then capitalize* $\hat{q}^*$ *and quit. Otherwise, go to next step.*

(2.2) *If* $\hat{q}^*$ *matches* $\mathscr{D}_-(\varnothing|s|'s|s')$*, landing on the n-th member of alphabetized* $\mathscr{D}$ *upon its first hit, then determine the correct spelling by the n-th member of alphabetized* $\mathscr{D}_0$*. Otherwise, leave* $\hat{q}^*$ *as is and quit.*

Furthermore, we correct misspelt questions by the following replacements:

a full job time	*cono sur*	*Điện Chí Minh*	□*fy*□	*general chu chicken*
a full-time job	*Southern Cone*	*Hồ Chí Minh*	□*fiscal year*□	*General Tso's Chicken*
Google in math	*jagger bomb*	*st patty*	*ti 82*	*trinity 5 7*
googol in math	*Jägerbomb*	*Saint Patrick*	*TI-82*	*Trin-i-tee 5:7*
tri tip	□*ua's*□	□*ww*□	□*wwii*□	
tri-tip	□*urinalysis*□	□*World War*□	□*World War II*□	

In Table 2 and Fig. S2, we showed that our numerical question answering machine had better performance on a subset of the WikiQA questions. Our criteria for defining such a subset (via its complement) are stated below.

Algorithm 3.12 (Heuristic classification of WikiQA questions)**.** *We classify a WikiQA question as "quantitative" (requiring rule-based reasoning), if it matches one of the following string patterns (regardless of letter case):*

how(□)$_m$(*big*|*deep*|*far*|*frequent*|*high*|*long*|*many*|*much*|*often*|*old*|*tall*)**X**,

*when***X**,

X(#|*number*|*percentage*|*proportion*)**X**,

what(□)(*bec*|*happen*|*year*)**X**.

Otherwise, we classify a WikiQA question as "qualitative" (requiring only associative reasoning).

[26] Note that if the reference page uses diacritics consistently and we alphabetize both $\mathscr{D}$ and $\mathscr{D}_0$, then the corresponding positions are occupied by the same words, without and with diacritics. For example, set A = *Estée Lauder and Eva Perón* and Q = *péron esTee lauder*, then we have $\mathscr{D}$ = {*Estee*, *Peron*} and $\mathscr{D}_0$ = {*Estée*, *Perón*}. After going through the remaining procedures in Step (2), we will end up with a corrected version of Q as *Perón Estée Lauder*. Our heuristic algorithm will malfunction, however, if P contains both *Estée* and *Estee* (*i.e.* the reference page uses diacritics inconsistently), or if alphabetic ordering is seriously affected by diacritics. In practice, the only glitch occurs in WikiQA-Q2158, where *Ho Chi Minh* becomes misspelt as *Điện Chí Minh*, due to unexpected alphabetic ordering of Vietnamese characters in *Mathematica* v11.3 (which brought us $\mathscr{D}$ = {*Ai*, *Bien*, *Chi*, *Cong*, *Dien*, *Ho*, *Nguyen*, *Phu*, *Quoc*, *Tat*, *Thanh*, *Viet*} and $\mathscr{D}_0$ = {*Ái*, *Biên*, *Chí*, *Cộng*, *Hồ*, *Điện*, *Nguyễn*, *Phủ*, *Quốc*, *Thành*, *Tất*, *Việt*}). This is subsequently corrected by a substitution rule.

3.3 Normalizations of ebooks for text mining

For the topic extraction and word translation experiments (Figs. S3–S16), we use the following algorithms for cleansing electronic books written in 12 European languages and 2 Asian languages, sorted in the same order as the remaining sections of this document. Hereafter, in accordance with *Mathematica* terminology, we use the word character $\mathbf{W} = (\hat{\chi}|\mathbf{D})$ to refer to either a letter or a digit.

Algorithm 3.13 (Text cleansing for English ebooks). *We process our text (ebooks from Gutenberg Project*[27] *and elsewhere) according to the following procedures:*

(1) *Do* $\overline{\mathbf{W}}$ U+0027 → $\square_{\times 2}$.

(2) *Replace*

$$(\square_m|\text{U+000A}_m|\texttt{--}|\texttt{*}|\texttt{\_}|(|)|\square) \qquad \mathbf{X}\in\left((\hat{\chi}|\texttt{'}|\texttt{-})_m\right)$$
$$\square \qquad\qquad \mathbf{X} \text{ if } \ell(\mathbf{X})=1;\ \mathbf{X}^{\{1\}}(\mathbf{X}^{[2,\ell(\mathbf{X})]})_{\_} \text{ if } \ell(\mathbf{X})>1.$$

where $\mathbf{X}^{\{1\}}(\mathbf{X}^{[2,\ell(\mathbf{X})]})_{\_}$ *reduces* $\mathbf{X}$ *to lower case except the initial character.*

Algorithm 3.14 (Text cleansing for Danish ebooks). *We normalize our text by doing*

$$(\square_m|\text{U+000A}_m|\texttt{--}|\texttt{*}|\texttt{\_}|\text{U+002D}|\text{U+2013}|\text{U+2014}|\text{U+201E}|(|)|\text{U+00AB}|\text{U+00BB}|\text{U+2026})\to\square.$$

Algorithm 3.15 (Text cleansing for German ebooks). *We normalize our text by doing* (U+0027|U+2019)→□, (U+2039|U+203A)→□, (□$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→□, u.s.w. → $\square_{\times 6}$.

Algorithm 3.16 (Text cleansing for Dutch ebooks). *We normalize our text by doing* (U+0027|U+2019)→□, (U+2039|U+203A)→□, (□$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→□.

Algorithm 3.17 (Text cleansing for Spanish ebooks). *We normalize our text by doing* (¿|¡)→ ∅, (□$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→□.

Algorithm 3.18 (Text cleansing for French ebooks). *We normalize our text (ebooks from Wikisource and Gutenberg Project) according to the following procedures:*

(1) *Do* M. → Mr.

(2) *Do* (□$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→□, (U+0027|U+2019)→□.

Algorithm 3.19 (Text cleansing for Latin ebooks). *We normalize our text by doing*

$$(\square_m|\text{U+000A}_m|\texttt{--}|\texttt{*}|\texttt{\_}|\text{U+002D}|\text{U+2013}|\text{U+2014}|\text{U+201E}|(|)|\text{U+00AB}|\text{U+00BB}|\text{U+2026})\to\square.$$

Algorithm 3.20 (Text cleansing for Polish ebooks). *We normalize our text according to the following procedures:*

(1) *Do* $(\varnothing|\text{U+002D}|\text{U+2014})(\text{U+0020}|\text{U+000A})_{m_0}[\mathbf{D}_{m_0}](\text{U+0020}|\text{U+000A})_{m_0} \to \varnothing$, $\hat{\chi}_1(\text{U+0027}|\text{U+2019})\hat{\chi}_2 \to \hat{\chi}_1\sharp\hat{\chi}_2$.

(2) *Do* (□$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→□, (U+0027|U+2019)→□.

Algorithm 3.21 (Text cleansing for Russian ebooks). *We normalize our text according to the following procedures:*

(1) *Do* $(\varnothing|\text{U+002D}|\text{U+2014})(\text{U+0020}|\text{U+000A})_{m_0}[\overline{(]}_{m_0}](\text{U+0020}|\text{U+000A})_{m_0} \to \varnothing$, **BookTitle**$(\square|\texttt{-})_{m_0}\mathbf{D}_{m_0}[\texttt{/}]\mathbf{D}_{m_0}(\square|\texttt{-})_{m_0}\to\square$.

(2) *Do* $\hat{\chi}_1[\texttt{?}]\hat{\chi}_2 \to \hat{\chi}_1 y \hat{\chi}_2$.

[27] It is always understood that the front and back matters (publisher's boilerplate, table of contents, introductory chapter, glossary, index, endnotes, etc.) in ebooks are manually removed. Footnotes that scatter in a few pages are left intact.

(3) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+0027|U+2019)→␣.

Algorithm 3.22 (Text cleansing for Finnish ebooks)**.** *We normalize our text (ebooks from the Gutenberg Project) according to the following procedures:*

(1) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+201C|U+201D)→␣, [**D**$_{m_0}$]→∅, **W**:$\hat{x}$→**W**♯$\hat{x}$.

(2) *Do* U+0027 $\overline{\mathbf{W}}$→␣, $\overline{\mathbf{W}}$ U+0027→␣.

Algorithm 3.23 (Text cleansing for Hungarian ebooks)**.** *We normalize our text (ebooks from* `http://mek.oszk.hu`*) according to the following procedures:*

(1) *Do* (∅|U+002D|U+2014)(U+0020|U+000A)$_{m_0}$[**D**$_{m_0}$](U+0020|U+000A)$_{m_0}$→∅, **BookTitle**(␣|-)$_{m_0}$**D**$_{m_0}$/**D**$_{m_0}$(␣|-)$_{m_0}$→␣.

(2) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+0027|U+2019|U+201C|U+201D)→␣.

Algorithm 3.24 (Text cleansing for Basque ebooks)**.** *We normalize our text according to the following procedures:*

(1) *Do* (∅|U+002D|U+2014)(U+0020|U+000A)$_{m_0}$[**D**$_{m_0}$](U+0020|U+000A)$_{m_0}$→∅.

(2) *Do eliz*→*ëliz.*[28]

(3) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+0027|U+2018|U+2019|U+201C|U+201D)→␣.

Algorithm 3.25 (Text cleansing for Korean ebooks)**.** *We normalize our text according to the following procedures:*

(1) *Do* (∅|U+002D|U+2014)(U+0020|U+000A)$_{m_0}$[**D**$_{m_0}$](U+0020|U+000A)$_{m_0}$→∅.

(2) *Do* 제**D**$_m$(권|장)→␣.

(3) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+0027|U+2018|U+2019|U+201C|U+201D|U+2500|U+25CB|U+3008|U+3009|U+300C|U+300D)→␣.

(4) *Adjust spacing in the text as follows:*[29]

(4.1) *Do* **D**백→**D**␣백, **D**천→**D**␣천, **D**만→**D**0␣천, **D**일→**D**␣, 지난**W**→지난␣**W**, 다음**W**→다음␣**W**, ␣다시␣$\hat{x}_{\notin}$(씨|양)→␣땄씆␣$\hat{x}$;

(4.2) *Do* $\hat{x}_{\in}$(가|나)␣싶→␣$\hat{x}$싶, 고␣나(~$^{\mathbf{X}_{\in}}$(∅|고|기|는|니|라|면|서|셔|신|실|심|야|요|자))→␣꼮나**X**, 고␣(~$\hat{x}_{\in}$(난|난다|날|남))→␣꼮$\hat{x}$, 고␣$\hat{x}_{\in}$(나겠|나느|나는|나니|나더|나려|나세|나셔|나셨|나시|나신|나십|나자|나지|납니|납시|놨|보(이)|본|볼|봄|봅|봐|봤(자))→␣꼮$\hat{x}$, 고␣$\hat{x}_{\in}$(말|싶)→␣고$\hat{x}$, $\hat{x}_{\in}$(과|앤|와)␣땄씆␣→␣다시␣, 까␣$^{\mathbf{X}_{\in}}$(말|보(이)|본|볼|봄|봅|봐|봤(자)|싶)→␣까**X**, $\hat{x}^{\hat{x}_1\in}$(다|아|어)␣$^{\hat{x}_2\in}$(보(이)|본|볼|봄|봅|봐|봤(자))→$\hat{x}$␣$\hat{x}_1\hat{x}_2$, $^{\hat{x}_1\in}$(냔|단|란|잔)␣말$^{\hat{x}_2\in}$(씀|이)→␣$\hat{x}_1$말$\hat{x}_2$, 지␣$^{\mathbf{X}_{\in}}$(마느|마세|마셨|마시|마십|말겠|말더|말려|말았|말지|맙니|맙시)→␣지**X**, 지␣(~$^{\mathbf{X}_{\in}}$(마는|마니|마셔|마신|마실|마심|만|말|말고|말기|말면|말아|말자|맒))→␣지**X**,

Algorithm 3.26 (Text cleansing for Turkish ebooks)**.** *We normalize our text according to the following procedures:*

(1) *Do* (∅|U+002D|U+2014)(U+0020|U+000A)$_{m_0}$[**D**$_{m_0}$](U+0020|U+000A)$_{m_0}$→∅, $\hat{x}_1$(U+2019)$\hat{x}_2$→$\hat{x}_1$♯$\hat{x}_2$, *I*→*ı*, *İ*→*i*, *BÖLÜM*(␣)$_m$**D**$_m$→*BÖLÜM*(␣).

(2) *Do* (␣$_m$|U+000A$_m$|--|*|_|U+002D|U+2013|U+2014|U+201E|(|)|U+00AB|U+00BB|U+2026)→␣, (U+0027|U+2018|U+2019|U+201C|U+201D)→␣.

[28] The Basque word *eliza* "church" should not be confused with the proper name *Eliza*.

[29] Some Korean words are best disambiguated by their immediate neighbours, such as ~고␣싶~ "want to do" [45, p. 251] and ~(가|나)␣싶~ "think it might" [45, p. 256]. This is the main reason why we reassign spaces to our text.

To obtain individual words from a normalized text, we split the text string at (‖ , ‖ . ‖ ? ‖ ! ‖ : ‖ ; ‖ U+0022) (for French, also apostrophe ' and hyphen -). To measure gaps between words in a normalized text, we further substitute spaces and punctuation marks according to the rules: (, ‖ : ‖ ;)→ , (. ‖ ? ‖ ! ‖ U+0022)→ $_{\times 2}$.

Part III
Protocols for word clustering

All the word clustering algorithms in this Part are available as *Mathematica* source codes (`English.nb` for word clustering and analysis of English texts, etc.), posted to Github

`https://github.com/yajun-zhou/linguae-naturalis-principia-mathematica`

To reproduce all the figures in this Part from these source codes, the readers need to download and/or purchase text corpora (not posted to our Github repository), following the instructions in Table S1.

Additionally, a separate source code `WikiQA.nb` is available, which implements our word clustering algorithm for English to the WikiQA dataset [44]. Based on the list of page titles in the WikiQA dataset, we extract `WikiSelecta.XML` (deposited as a zipped file `WikiSelecta.zip` in our Github folder) from the static dump of English Wikipedia dated Sept. 1, 2015 (file name: `enwiki-20150901-pages-articles-xml.bz2`). This subset of 1242 Wikipedia pages forms the only training source and knowledge base for our question-answering machine.

4 Approximate word clustering in English

In this section, we describe an automatic algorithm to cluster English words according to their morphologies. Due to the complicated linguistic history of English that derives from both Germanic heritage and Latin/Romance influence, the English language is rich in both inflectional complexity (such as vowel alternations in verb forms, a feature common to Germanic languages) and derivational complexity (such as many Latinate suffixes that distinguish words within the same word family). Our English word clustering algorithm thus involves methods that will be later used in the studies of Germanic (§5) and Romance (§6) languages.

4.1 Capitalization and stop words

Before presenting our algorithms in detail, we need to briefly discuss the capitalization problem in English and the list of English stop words.

English orthography requires capitalization of proper nouns and their derivatives.[30] It turns out that proper nouns and common nouns in English tend to have different etymological sources, and consequently, different morphological structures. Thus, it is helpful to single out "intrinsically capitalized words" used in an English text before performing word clustering algorithms. The precise identification of the capitalization status of a particular word is a non-trivial task: words that start any sentence are always capitalized; some authors may occasionally choose to use ALL CAPS for emphasis; some "intrinsically capitalized words" may share the same spelling with other ordinary words, up to capitalization (such as *Miss* and *miss*). For our text mining purposes, we adhere to the following heuristic determination of the capitalization status.

Definition 4.1 (Heuristic capitalization)**.** We say that a word is "heuristically capitalized" in a given text, if its all-lowercase form never appears in the same document. Otherwise, we consider the word's spelling "heuristically uncapitalized".[31] □

A small collection of English words perform important grammatical/logical functions, but transmit very little information on their own. These words need to be excluded in most natural language processing tasks, and are thus referred to as "stop words". There is not a universal standard for the list of English stop words. We state our empirical list in the definition below.

[30] Among the major European languages, this capitalization rule perhaps only applies to English. In all the other European languages studied in this work, adjectives derived from proper nouns are spelt in lower case forms. In modern German and pre-1948 Danish, all nouns are capitalized. In Modern Turkish (which we regard as an Asian language), derivatives of proper nouns are also capitalized, as in English. However, we do not find it particularly helpful to treat Turkish proper nouns differently from other nouns in word clustering. Therefore, we will restrict Definition 4.1 to English only.

[31] According to this definition, a word's spelling must be either "heuristically capitalized" or "heuristically uncapitalized". If *Miss* and *miss* both appear in the same text, then they will be regarded as the same word, in the "heuristically uncapitalized" status of our classification scheme. On the other hand, if a text contains *Elizabeth* and *ELIZABETH*, but does not contain *elizabeth*, then we merge *Elizabeth* and *ELIZABETH* to the same word, in the "heuristically capitalized" status.

Definition 4.2 (English stop words)**.** If a word belongs to the following list[32]:

> –, *&c*, *a*, *about*, *above*, *across*, *after*, *afterward*, *afterwards*, *again*, *against*, *ago*, *ahead*, *albeit*, *all*, *almost*, *alone*, *along*, *already*, *also*, *although*, *always*, *am*, *amid*, *amidst*, *among*, *amongst*, *an*, *and*, *another*, *any*, *anybody*, *anybody's*, *anyhow*, *anyone*, *anything*, *anyway*, *anywhere*, *apart*, *are*, *aren't*, *around*, *as*, *aside*, *at*, *away*, *back*, *be*, *became*, *because*, *become*, *becomes*, *becoming*, *been*, *before*, *behind*, *being*, *below*, *beneath*, *beside*, *besides*, *between*, *beyond*, *both*, *but*, *by*, *can*, *can't*, *cannot*, *could*, *couldn't*, *did*, *didn't*, *do*, *does*, *doesn't*, *doing*, *don't*, *done*, *down*, *during*, *each*, *either*, *else*, *elsewhere*, *enough*, *etc*, *even*, *ever*, *every*, *everybody*, *everybody's*, *everyone*, *everyone's*, *everything*, *everywhere*, *except*, *few*, *for*, *from*, *front*, *full*, *further*, *furthermore*, *get*, *gets*, *getting*, *got*, *gotten*, *had*, *hadn't*, *hardly*, *has*, *hasn't*, *have*, *haven't*, *having*, *he*, *he'd*, *he'll*, *he's*, *hence*, *her*, *here*, *here's*, *hereabout*, *hereabouts*, *hereafter*, *hereby*, *herein*, *hereinafter*, *hereof*, *hereon*, *hereto*, *heretofore*, *hereunder*, *hereunto*, *hereupon*, *herewith*, *hers*, *herself*, *him*, *himself*, *his*, *hither*, *how*, *how's*, *however*, *howsoever*, *i*, *I*, *i'd*, *I'd*, *i'll*, *I'll*, *i'm*, *I'm*, *i've*, *I've*, *if*, *in*, *inside*, *instead*, *into*, *is*, *isn't*, *it*, *it's*, *its*, *itself*, *just*, *last*, *least*, *less*, *let*, *let's*, *letting*, *likewise*, *many*, *may*, *me*, *mere*, *merely*, *might*, *more*, *moreover*, *most*, *mostly*, *much*, *must*, *mustn't*, *my*, *myself*, *neither*, *never*, *next*, *no*, *nobody*, *nobody's*, *none*, *nor*, *not*, *nothing*, *now*, *nowhere*, *of*, *off*, *often*, *on*, *once*, *one*, *one's*, *ones*, *ones'*, *oneself*, *only*, *or*, *other*, *others*, *otherwise*, *ought*, *our*, *ours*, *ourselves*, *out*, *outside*, *over*, *own*, *per*, *perhaps*, *quite*, *rather*, *same*, *several*, *shall*, *shan't*, *she*, *she'd*, *she'll*, *she's*, *should*, *shouldn't*, *since*, *so*, *some*, *somebody*, *somebody's*, *somehow*, *someone*, *someone's*, *something*, *sometimes*, *somewhat*, *somewhere*, *soon*, *sooner*, *still*, *such*, *than*, *that*, *that's*, *the*, *thee*, *their*, *theirs*, *them*, *themselves*, *then*, *there*, *there's*, *thereabout*, *thereabouts*, *thereafter*, *thereat*, *thereby*, *therefor*, *therefore*, *therefrom*, *therein*, *thereof*, *thereon*, *thereto*, *theretofore*, *thereunder*, *thereunto*, *thereupon*, *therewith*, *these*, *they*, *they'd*, *they'll*, *they're*, *they've*, *thine*, *this*, *those*, *thou*, *though*, *through*, *thus*, *thy*, *thyself*, *till*, *to*, *together*, *too*, *toward*, *towards*, *under*, *unless*, *until*, *up*, *upon*, *us*, *versus*, *very*, *via*, *was*, *wasn't*, *we*, *we'd*, *we'll*, *we're*, *we've*, *well*, *were*, *weren't*, *what*, *what's*, *whatever*, *whatsoever*, *when*, *whenever*, *when's*, *where*, *where's*, *whereabouts*, *whereas*, *whereat*, *whereby*, *wherefore*, *wherein*, *whereof*, *whereon*, *wheresoever*, *whereto*, *whereupon*, *wherever*, *wherewith*, *wherewithal*, *whether*, *which*, *whichever*, *whichsoever*, *while*, *who*, *who's*, *whoever*, *whole*, *whom*, *whomever*, *whomsoever*, *whose*, *whosoever*, *why*, *why's*, *whyever*, *whysoever*, *will*, *with*, *within*, *without*, *won't*, *would*, *wouldn't*, *yes*, *yet*, *you*, *you'd*, *you'll*, *you're*, *you've*, *your*, *yours*, *yourself*, *yourselves*,

then we consider it an English stop word.[33] All the English stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

4.2 Modified Porter stemming algorithm for English

The Porter stemming algorithm is a very useful word clustering tool for information retrieval in English [39] and many other European languages [40]. Through a series of string manipulations, the Porter stemming algorithm generates a token (known as "Porter stem") for an arbitrary word in a given language, without consulting a dictionary. Words sharing the same token are identified with each other in the Porter stemming procedure. While this procedure correctly lumps together morphologically related words in many scenarios, it ignores morphological irregularities that prevail in the vocabularies of many European languages, including English.

In this work, we modify Porter's stemming algorithm so that it works more accurately on realistic vocabulary lists generated from text mining tasks. Our modification has two aspects. First, instead of associating each word with a unique Porter stem, we construct the token in two stages: a longer stem called *effective spelling*, and a shorter stem called *essential root* (§4.2.1). Second, instead of using exact matching of tokens as a criterion for word clustering, we add some *admissible mutation* rules to accommodate to morphological irregularities, and build an *approximate clustering* algorithm thereupon (§4.2.2).

In English, there are about two hundred irregular verbs in common use. A limited subset of these verbs are highly irregular, and they need to handled by explicitly written exceptions to rules during the stemming procedure (§4.2.1). The remaining

[32] Our list of English stop words differ from the built-in list in *Mathematica* v11.0 in the following aspects:
(1) Our list includes – (two hyphens joined together), which accommodates to vintage encoding in certain ebooks;
(2) Our list includes *shall* and *towards*, to accommodate to British English;
(3) Our list includes some common function words: *afterward*, *afterwards*, *ago*, *ahead*, *albeit*, *amid*, *amidst*, *amongst*, *anyhow*, *anyway*, *apart*, *aside*, *away*, *beneath*, *beside*, *besides*, *beyond*, *else*, *elsewhere*, *everybody*, *everybody's*, *everyone's*, *except*, *front*, *hardly*, *hence*, *hither*, *inside*, *instead*, *just*, *likewise*, *neither*, *oneself*, *otherwise*, *outside*, *somebody*, *somebody's*, *someone's*, *sometimes*, *somewhat*, *soon*, *sooner*, *thereafter*, *therein*, *thereof*, *thereupon*, *unless*, *versus*, *via* and *yes*;
(4) Our list includes *thee*, *thine*, *thou*, *thy* and *thyself*, to accommodate to Early Modern English (such as in the works of Shakespeare);
(5) Our list excludes *find*, *first*, *give*, *go*, *interest*, *keep*, *made*, *noone*, *part*, *put*, *see*, *seem*, *seemed*, *seeming*, *seems*, *show*, *side* and *take*—while some of these words might appear disposable in certain rudimentary information retrieval tasks, we find it harmful to regard them as stop words for machine comprehension of English texts.

[33] In the list above (and also the substitution rules below), the apostrophe refers to either Apostrophe U+0027 or Right Single Quotation Mark U+2019, as is appropriate for the particular text mining task.

majority of these irregular verbs, however, do exhibit systematic vowel/consonant alternation patterns, such as *keep/kept*, *sleep/slept*, *sweep/swept*, *weep/wept*, *break/broke/broken*, *speak/spoke/spoken*, *steal/stole/stolen*. These systematic irregularities are reflexes of historical Germanic vowel mutations, which are also found in modern Danish, German and Dutch (§5). We will treat these vowel alternations with an *admissible mutation* mechanism (§4.2.2).

4.2.1 Effective spelling and essential root

Before deducing analogs of the Porter stem as the *effective spelling* and the *essential root*, we need to regularize some English verbs (Algorithm 4.3) and normalize the vowel lengths in English spelling (Algorithm 4.4). It should be noted that our purpose is to deduce a token for an English word that is suitable for morphological studies, so the token itself does not have to resemble the word's etymological stem (as was in the Porter stemming algorithm): in fact, we are going to exploit a wide assortment of special Latin characters (like the German *ä*, *ö*, *ü*, *ß*, the Spanish *ñ*, and various accented vowels appearing in Romance languages) as well as some Greek letters[34] to construct the effective spelling and essential root of an English word.

Algorithm 4.3 (Regularization of English verbs)**.** *For an English word $\hat{\sigma}$, its "regularized verb form"* RegVb($\hat{\sigma}$) *is deduced by the following four steps:*[35]

(1) *Do* $\mathbf{X}^{\in}$*(a|i)bilit***X** $\to$ **X***ble, $\hat{\chi}$ful(ler|lest|ness)***X** $\to \hat{\chi}$*, feet* $\to$ *foot, heaven* $\to$ *ηεανεν̃, ificat***X** $\to$ *ify, ili(s|z)(e|ing)***X** $\to$ *ility, inabil***X**~ $\to$ *unable, inalien*~ $\to$ *unalien, istic***X** $\to$ *ist, positiv* $\to$ *ποσιτιν, reciprocal* $\to$ *reciprocity, toric***X** $\to$ *tory,* ~$^{\mathbf{X}^{\in}}$*(barb|hil)arit(ies|y)* $\to$ **X***arious,* ~$^{\mathbf{X}^{\in}}$*(c|d|g|n)uit(ies|y)* $\to$ **X***uous, ~alit(ies|y)* $\to$ *al, ~arit(ies|y)* $\to$ *ar, ~charit(ies|y)* $\to$ *charitable, ~circuit(ies|y)* $\to$ *circuitous, ~enmit(ies|y)* $\to$ *enemy, ~femini(∅|ni)t(ies|y)* $\to$ *feminine, ~idit(ies|y)* $\to$ *id, ~lar(ly|s)* $\to$ *lar, ~necessit(ies|y)* $\to$ *necessary, ~ntit(ies|y)* $\to$ *ntify, ~simplicit(ies|y)* $\to$ *simple, ~tricit(ies|y)* $\to$ *tric, ~vanit(ies|y)* $\to$ *vain, ~virgin(∅|s)* $\to$ *virginity.*

(2) *In a single sweep, perform replacements according to the following substitution rules:*

$\hat{\chi}\notin$*(i)aris(m|t)***X**	$\hat{\chi}\notin$*(s|t)tress*	*(Us|Usa)*	$\mathbf{X}^{\in}$*(∅|un)dece(it|pt)***X**~							
$\hat{\chi}$*ial*	$\hat{\chi}$*tor*	*American*	**X***deceive*							
$\mathbf{X}^{\in}$*(alchem|narch)is(m|t)***X**	$\mathbf{X}^{\in}$*(ap|im|multi)pl(y)(ie***X***)*		$\mathbf{X}^{\in}$*(at|g|ic|it|iv|nt)ious*							
X*y*	**X***plication*		**X***ion*							
$\mathbf{X}^{\in}$*(b|h)ang(ed|ing)*~	$\mathbf{X}^{\in}$*(cent|cloist|integ|neut|sepulch|spect)er*									
X*ang*	**X***re*									
$\mathbf{X}^{\in}$*(con|miscon|per|precon)ceptX*~	*a(c|n)eous***X**	*ancestress*	*avuncular***X**~	*Belgium*	*breadth*	*buri*	*bury*	*capric*		
X*ceive*	*al*	*ancestor*	*uncle*	*Belgian*	*broad*	*βυρι*	*βυργ*	*καπric*		
children	*choice*	*clar*~	*comic*~	*condition*	*contempor***X**	*crash*	*dempt***X**	*Denmark*	*earn*	*ength*
child	*choose*	*clear*	*comedic*	*conδτιον*	*contempor*	*κrash*	*deem*	*Danish*	*εαarn*	*ong*
*flexi(on|ve)***X**	*flower*	*focus***X**	*France*	*fratern***X**~	*grove*	*hall*	*homage*	*hour*	*iaris(m|t)***X**	
flect	*φλower*	*focus*	*French*	*brother*	*γρove*	*ζαλl*	*homαγe*	*ορα*	*iary*	
idiot	*inguish*	*instab*~	*Ireland*	*ironi(c|es)***X**	*ironing*	*irony*	*Italy*	*Korean*		
idiοτ	*inct*	*unstab*	*Irish*	*iρony*	*iron*	*iρony*	*Italian*	*Korea*		
*langu(id|or)***X**~	*laughter*~	*leg*~	*lesson*	*ma('|')d)am*	*marit***X**~	*matern***X**~	*matrimon***X**~	*membran*		
languish	*laugh*	*λεγ*	*λεκcjon*	*φru*	*marry*	*mother*	*marry*	*meμβρaν̃*		
menace	*mentalit***X**	*Mexico*	*mistress*	*mound*	*mourn*~	*pacif***X**~	*patern***X**~	*polite*	*promis*	*prophet***X**
μeν̃ace	*mental*	*Mexican*	*master*	*mωund*	*μουρn*	*peace*	*father*	*πλίτε*	*πrωmis*	*prophetic*
publish	*rash*~	*rebellion***X**	*rece(ipt|pt)***X**~	*roof*~	*salar*	*satin*	*satir*			
publication	*ρash*	*rebel*	*receive*	*ρωoof*	*σαλαρ*	*σατin*	*σβατir*			
*scienti(fic|st)***X**	*Scotland*	*severed*	*soror***X**~	*sow*~	*Spain*	*species*	*sticit***X**	*tenab*	*tenac***X**~	*tenant*
science	*Scottish*	*sever*	*sister*	*σow*	*Spanish*	*σπiecεσ*	*stic*	*τteν̃ab*	*τenac*	*τeν̃ant*
tender	*tenet*	*tenta*	*theori***X**~	*tician***X**	*town*	*traged*~	*Ukraine*	*usur*	*Wales*	
teν̃der	*τenετ*	*τentα*	*theory*	*tical*	*towν̃*	*tragic*	*Ukrainian*	*usuρ*	*Welsh*	

[34] The Greek letter *β* should not be confused with German *ß*.

[35] Not all these substitutions operate on verbs, yet most of them are conducive to correct clustering of certain verbs. For example, *missile* and *mission* should not be merged with *miss(∅|ed|es|ing)*, and *passion* should be separated from *pass(∅|ed|es|ing)*. The word family *attent(ion|ive|ively)* was indeed etymologically related to the verb *attend*, but their modern meanings are well separated. To avoid such a conflation, while guaranteeing the correct grouping of *inten(d|t)*, *exten(d|t)* and so on, we need to alter the spelling *attend*~. A diligent reader may find the rationale behind the other entries in our substitution rules.

weari$_{\hat{\chi}\notin}$*(n)***X**	*weary***X**	*woke*	*worr*	*chagrin(∅|s)*	*conic(∅|al)(∅|ly)*	*good*	*hung*	*made*	*radii*	*ran*
ωεαρι	*ωεαρι*	*wake*	*woρr*	*chagrine*	*cone*	*γοωδ*	*hang*	*make*	*radius*	*run*

sis	*ten(∅|s)(th***X***)*	*~*$_{\hat{\chi}\notin}$*(a|e|i|o|u)ar(∅|ies|ily|ize|y)*	*~*$_{\hat{\chi}\notin}$*(p)lus*
sister	*10ten*	*χ̂ial*	*χ̂li*

~(is|iz)(ab|able|ation|e|ed|ement|er|ing)(∅|ly|s)	*~*$^{\mathbf{X}\in}$*(atr|fer|prec)ocit(ies|y)*
ize	**X***ocious*

~acit(ies|y)	*~antal*	*~antas(ies|y)*	*~ics*	*~istry*	*~monopol(ies|y)*	*~nness*	*~radical(∅|s)*	*~uum*
acious	*ance*	*antasize*	*ical*	*ister*	*monopolize*	*n*	*radicalize*	*ua*

(3) *Do εαarnest→ εαaρneστ,* $^{\mathbf{X}\in}$*(a|i|o|u)si(on|ve)***X***→***X***de, abdomen→abdomin, beautical→beauty, began→begin, caus*$^{\mathbf{X}\in}$*(a|e|i) → καωζ***X***, comedo~→ κομεδο, criti(c|que)(∅|al)(∅|ly)~→criticize, deer→dεer, hung~→ φaiμ, pretent~→ pretend, repeat***X***→repetition, savage→saβaγe, serious→sεrious, sever(e|it) → σεωεre, suffer→sωuffer, suspic→suspect, tent~→tenτ, tranquil(∅|l)iz***X***→tranquil, weariness***X***→ ωεαρι, ~(risen|rose) →rize, ~χ̂tic(∅|al)(∅|ly)()(iz***X***)→ χ̂sis, ~s's→∅.*

(4) *Replace*

'~	$^{\mathbf{X}\in}$*(ab|pre)sence*	$^{\mathbf{X}\in}$*(ad|de)vi(c|z)*	$^{\mathbf{X}\in}$*(col|e)lid*	$^{\mathbf{X}\in}$*(de|of)fen(c|s)*	$^{\mathbf{X}\in}$*(di|pro)vid*
∅	**X***sent*	**X***vis*	**X***lis*	**X***fend*	**X***vis*

$^{\mathbf{X}\in}$*(dissatis|lique|petri|putre|rare|satis|stupe|vitri)f(ie|ied|ies|y|ying)*
X*fact*

$^{\mathbf{X}\in}$*(g|i|r|s)onal(∅|i|ly)***X**	$^{\mathbf{X}\in}$*(k|K)ing~*	$^{\mathbf{X}\in}$*(mpl|rt|tr)aint*	*attend~*	*aur*	*automation*
X*on*	**X***oenig*	**X***ain*	*attenzd*	*aωr*	*automasis*

begg(ar|ed)~	*counten*	*de*$^{\mathbf{X}\in}$*(c|r)id*	*duct(∅|ion)*	*dye*	*expence*	*explain~*	*flatter*
beg	*kouten*	*de***X***is*	*duce*	*dzye*	*expense*	*explan*	*flaττer*

*forg(e|o)t(∅|ten)***X**	*former~*	*four*	*inal(∅|i|ly)***X**	*licen(c|s)~*	*listen*	*mad*	*mention*	*migrant*	*missile*
vγeβ	*φormer*	*φour*	*in*	*licent*	*lisztzn*	*μφaδ*	*menzzion*	*migrate*	*missle*

mission	*passion*	*pool*	*possess*	*preceden(c|t)*	*prett*	*prevail~*	*prevent~*	*rouse*	*solde~*	*soldi~*
mit	*patzion*	*poωl*	*pozzezz*	*precezβ*	*prεeτ*	*preval*	*prezvent*	*rzouse*	*soldq*	*soldw*

sump	*temper*	*thought*	*typ~*	*warn~*	*(flew|flown)*	*(mete|meted|metes|meting|metings)*
sume	*tempεr*	*think*	*zyp*	*wxarn*	*fly*	*mzet*

(ridden|rode)	*~*$^{\mathbf{X}\in}$*(∅|in|mis|off|out|over|p|under)led*	*~*$^{\mathbf{X}\in}$*(∅|s)lain*	*~*$^{\mathbf{X}\in}$*(b|br|c|l|r)ook*
ride	**X***lead*	**X***lay*	**X***ooked*

brought	*caught*	*eat(∅|en|ing|s)*	*lad(∅|die|dish|hood)(∅|s)*	*meant*	*sought*	*taught*	*won*
bring	*catch*	*ate*	*λaδ*	*mean*	*seek*	*teach*	*win*

(~sis|siz)	*~(did|does|doing|done)*	*~(goes|going|gone|went)*	*~bought*	*~came*
ses	*do*	*go*	*buy*	*come*

~ced(e|ed|es|ing)	*~ered*	*~gave*	*~***X***-in-law*	*~heard*	*~itten*	*~rnt*	*~see(∅|ing|n|s)*	*~stood*	*~visor*
cezβ	*er*	*give*	*λαω***X**	*hear*	*iten*	*rn*	*saw*	*stand*	*vise*

The silent *e* in English spelling may affect the length of the immediately preceding vowel. Pairs of words like *hat*/*hate* and *rat*/*rate* are etymologically unrelated. Motivated by modern Dutch orthography (§5.3.1), we will introduce a heuristic procedure to mark the vowel lengths in English words explicitly. This procedure not only detects long vowels in **CVC***e* patterns, but also words derived therefrom. (For example, the vowel *o* in *devote*, *devoting* and *devotion* should all be consistently marked as long.)

Algorithm 4.4 (Normalization of English vowel lengths)**.** *For a text string* $\hat{\sigma}$*, its "normalized English vowel length form"* NormVL($\hat{\sigma}$) *is determined through the following procedure:*

(1) *If* $\hat{\sigma}$ *comes from a word that is "heuristically capitalized" (Definition 4.1), then* NormVL($\hat{\sigma}$) = *zx*$\hat{\sigma}'$ *where* $\hat{\sigma}'$ *results from doing ~ean(∅|s) → ∅ on* $\hat{\sigma}$*.*[36]

[36]Here, the prefix "zx" push the word down to the end of an alphabetized list later. In English, the ending *~ean* is found mainly in adjectives (and nouns) deriving from proper nouns. The procedure here ensures that words like *Argentine*(∅|*an*)(∅|*s*), *Boole*(∅|*an*) and *Euclid*(∅|*ean*) *Europe*(∅|*an*)(∅|*s*) are clustered properly.

(2) *Otherwise, perform the following substitutions, in five sequential sweeps:*

(2.1) *Do* $\sim osit(ies|y) \to ous$, $\sim\hat{\chi}\hat{\chi}oe(\varnothing|d|s) \to \hat{\chi}\hat{\chi}o$.

(2.2) *Replace*[37]

$(lose\|loss)$	$\hat{\chi}tr$	$\mathbf{X}^{\in}(dis\|ex\|pre)tens$	$\mathbf{X}^{\in}(ex\|sus)pens$
lost	$\hat{\chi}ter$	**X***tend*	**X***pend*

$\mathbf{X}^{\in}(\hat{\chi}(a\|o))tion$	*au*	*capab*~	*deep*$(\varnothing\|e\mathbf{X})$	*earl*$(i\|y)$~	*eed*	*fer*	*fine*~	*freedom*	*gas*~	*generat*~	*genero*~	*ll*
X*tiv*	*aau*	*kapab*	*depth*	*earzli*	*ezß*	*fér*	*fîzne*	*free*	*gazß*	*jenerat*	*gjenero*	*λ*

manner	*mere*~	*mode*~	*neer*	*oe*	*oo*	*pre*~	*puls*	*quiet*	*rough*	*side*~	*some*~	*soul*	*ss*	*static*	*stel*~
maznner	*mzere*	*modz*	*ne*	*oue*	*ú*	*pré*	*pel*	*cquiet*	*rouffgh*	*szide*	*somz*	*szoul*	*zß*	*sztatic*	*sztel*

sudden~	*summer*	*ton*~	*water*~	*women*	*idl*$(est\|y)$	*genera*	*gently*	*singly*	*ski*$(\varnothing\|ed\|ing\|s)$
swudden	*suzmmer*	*twon*	*wazter*	*woman*	*idle*	*genus*	*gentle*	*single*	*szki*

(2.3) *Replace every occurrence of double letter by a corresponding single capital letter (i.e.* $ee \to E$ *etc.).*

(2.4) *Set* $\mathbf{C}_1$ *as any lowercase English letter*[38] *other than* $\{a, e, i, o, u, y\}$, $\mathbf{V}$ *as any one among* $\{a, e, i, o, u\}$, *and* $\mathbf{C}_2$ *as any lowercase English letter other than* $\{a, e, i, o, u, w, y\}$. *Define* $\mathbf{V}^+$ *as the capital form of the vowel* $\mathbf{V}$. *Replace*[39]

$\mathbf{C}_1\mathbf{V}(^{\mathbf{X}\in}((\mathbf{C}_2(ag\|al\|a\lambda\|at\|e\|ing\|ion\|ish\|it\|iv\|ous))\|ck\|gh\|nd))$	$ang(e\|ing)$
$\mathbf{C}_1\mathbf{V}^+\mathbf{X}$	*Ange*

(2.5) *Replace*[40]

A	*E*	*é*	*I*	*lezß*	*O*	*T*	*ú*	*U*	*λ*	*gOv*~	*inter*~	*be*$(Ter\|st)$
aa	*ee*	*e*	*í*	*zlezß*	*óó*	*tt*	*oo*	*u*	*ll*	*góv*	*jntr*	*γοωδ*

before converting the string to lowercase.

With the preparations above, we can derive the effective spelling of an English word. The effective spelling (to be derived in Algorithm 4.5) is a conservative estimate for the (etymologically correct) word stem, which usually has longer string length than the essential root (to be derived in Algorithm 4.8).

Algorithm 4.5 (English effective spelling)**.** *For an English word* $\hat{\sigma}$*, its effective spelling* EffSpell$(\hat{\sigma})$ *is constructed in seven sequential steps:*

(1) *Take* NormVL(RegVb$(\hat{\sigma})$)*, and perform the following replacements in a single sweep:*

$\hat{\chi}aid$	*ck*	*clamat*	*emn*	*enten*	*family*	*gral*	*ióón*	*iv*	*la*	*ld*	*lóó*	*ng*	*oi*	*ou*	*ply*
$\hat{\chi}ayed$	*c*	*claimat*	*μn*	*entzen*	*families*	*graation*	*ion*	*ív*	*lôa*	*llz*	*lo*	*zñ*	*zö*	*í*	*plie*

rd	*sh*	*tim*	*ui*	*ye*	~'$(\varnothing\|s)(\varnothing\|')$	*~ant*	*~ble*$(\varnothing\|d\|s)$	*~hie*$(r\|st)$	*~ook*	*~s*	- (hyphen)
rzd	*š*	*ztim*	*zü*	*yme*	∅	*ance*	*bly*	*hy*	*aaken*	∅	∅

(2) *Replace*

$\hat{\chi}ied$~	$\mathbf{C}_1^{\in}(c\|w)a$~	*ba*~	$\mathbf{C}_2^{\in}(s\|t\|w)^{\mathbf{X}\in}(e\|o\|óó)$~	*dóó*~	$f^{\mathbf{V}_1\in}(a\|e)$~	$f^{\mathbf{V}_2\in}(i\|í)$~	*móó*~	*dad*$(\varnothing\|y)$	*m*$(o\|u)m$
$\hat{\chi}ieed$	$\mathbf{C}_1\beta a$	*βa*	$\mathbf{C}_2 z\mathbf{X}$	*δo*	$\varphi\mathbf{V}_1$	$ff\mathbf{V}_2$	*μo*	*φaθer*	*mother*

(3) *Do* $\sim f(\varnothing|\ e) \to ve$, $\sim xe \to x$, $\sim\hat{\chi}\hat{\chi}\hat{\chi}ly \to \hat{\chi}\hat{\chi}\hat{\chi}$ *(i.e. remove final "ly" if it is preceded by at least three letters).*

(4) *Do* $h(i|í) \to hî$, $lu \to lvu$, $y \to ie$.

(5) *Do* $\sim\hat{\chi}\hat{\chi}\hat{\chi}(tic|tion) \to \hat{\chi}\hat{\chi}\hat{\chi}t$, $\sim\hat{\chi}\hat{\chi}\hat{\chi}(age|ed|ful|ical|izñ|îzñ|ment|nezß) \to \hat{\chi}\hat{\chi}\hat{\chi}$.

[37] As we do *deep*$(\varnothing|e\mathbf{X}) \to$ *depth* to the words *deep*, *deeper*, *deepest*, *deeply*, we obtain *depth*, *depth*, *depth*, *depthly*. (In *Mathematica*, the command `StringReplace[{"deep", "deeper", "deepest", "deeply"}, "deep"~~""|("e"~~___) :> "depth"]` results in `{"depth", "depth", "depth", "depthly"}`.) Here, the wildcard symbol $\mathbf{X} \in \{r, st\}$ is deleted after replacement, because $\mathbf{X}$ does not reappear in the destination string $\hat{\sigma}' =$ *depth* (cf. Definition 3.6).

[38] A letter is an English letter if and only if it is one of the 26 members in the ISO basic Latin alphabet.

[39] For readers' benefit, we point out that the corresponding *Mathematica* substitution rules are `{(C1: Except["e"|"i"|"o"|"u"|"y", CharacterRange["b", "z"]])~~(V: "a"|"e"|"i"|"o"|"u"~~(X: (Except["e"|"i"|"o"|"u"|"w"|"x"|"y", CharacterRange["b", "z"]]~~"ag"|"al"|"aλ"|"at"|"e"|"ing"|"ish"|"it"|"ous")|"ck"|"gh"|"nd") :> C1 <> ToUpperCase[V] <> X, "ang"~~"e"|"ing" :> "Ange"}`. As our substitutions disallow overlaps, the letter *e* in patterns like *ndering* (as in *plundering*, *pondering*) will not be marked as a long vowel.

[40] Note that during a single sweep, the longer matches take priority over shorter matches. Therefore, the result NormVL(*government*) = *góvernment* does not contain *óó*.

(6) *Replace*

$(eft\|iev\|\hat{\imath}ev)$	$(o\|\acute{o}\acute{o})^{\mathbf{X}\in}(p\|r)$	ch	ea	ir	ll	$mi\sim$	ph	th	$\sim_{\hat{\chi}\notin}(d)ent(\varnothing\|ive)$	$\sim^{\mathbf{X}\in}(pel\|s)ion$	$\sim\hat{\chi}\hat{\chi}\hat{\chi}i\check{s}$	$\sim lt$	$\sim own$
$\ddot{a}\ddot{a}ve$	$\hat{o}\mathbf{X}$	$\check{c}$	$\ddot{a}\ddot{a}$	$\ddot{\imath}r$	λ	μi	$\varphi\varphi$	θ	$\hat{\chi}eend$	$\mathbf{X}\acute{\imath}v$	$\hat{\chi}\hat{\chi}\hat{\chi}$	l	ow

(7) *Replace*[41]

$\underline{beg}$	$\underline{d\ddot{a}\ddot{a}(d\|\theta)}$	$\underline{\hat{\chi}_1\mathbf{X}_1{}_{\hat{\chi}_2\notin}(e\|i)e}$	$\underline{zx\mathbf{X}_3}$
$bbeg$	die	$\hat{\chi}_1\mathbf{X}_1\hat{\chi}_2 ae$, *if* $\hat{\chi}_1\mathbf{X}_1 = zx\mathbf{X}_2$; $\hat{\chi}_1\mathbf{X}_1\hat{\chi}_2$ *otherwise.*	$zx\widetilde{\mathbf{X}}_3$, *where* $\widetilde{\mathbf{X}}_3$ *results from doing* $\sim ie \to y$ *on* $\mathbf{X}_3$.

Here, "$\hat{\chi}_1\mathbf{X}_1 = zx\mathbf{X}_2$*" means that the string* $\hat{\chi}_1\mathbf{X}_1$ *begins with two letters "zx".*

Up to the current stage, we have introduced many non-English letters into the tokens assigned to English words. Some of these non-English letters will be regarded as extensions to the set of English vowels.

Definition 4.6 (English Vowel Extensions)**.** Hereafter in §4, the symbol $\mathbf{V}^*$ stands for any member from the list {*a*, *ä*, *e*, *i*, *í*, *î*, *ï*, *o*, *ó*, *ô*, *ö*, *u*, *ü*}, the so-called English vowel extensions. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}^*_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of English vowel extensions.

Dual to the notations above, the symbol $\mathbf{C}^*$ stands for any character that does not belong to the list {*a*, *ä*, *e*, *i*, *í*, *î*, *ï*, *o*, *ó*, *ô*, *ö*, *u*, *ü*}, and $\mathbf{C}^*_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list. □

As mentioned before, the conservative effective spelling algorithm assigns a long token to an English word, while the aggressive essential root algorithm will produce a short token, removing as many final letters as "appropriate". To ensure "appropriateness" during the short token generation, we need to explicitly label some part of an English word as unremovable. The string length of this unremovable part is called the English protected range, as defined below.

Definition 4.7 (English protected range)**.** Let $\hat{\sigma}$ be the effective spelling of an English word, its protected range ProtRg($\hat{\sigma}$) is an integer determined as follows:[42]

- Look for the string pattern $(\varnothing|be|cínter|co|cô|dee|di|fô|ha|ma|pre|pro|próó|r\mathbf{V}^*_m|ster|su|te|\mathbf{V}^*_m)\mathbf{C}^*_{m_0}\mathbf{V}^*_m\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines ProtRg($\hat{\sigma}$); otherwise, set ProtRg($\hat{\sigma}$) = 0. □

Example 4.7.1. Let $\hat{\sigma} = TNT$, then EffSpell($\hat{\sigma}$) = *zxTNT*, and ProtRg(EffSpell($\hat{\sigma}$)) = 0; let $\hat{\sigma} = trinitrotoluene$ (the IUPAC name for TNT), then EffSpell($\hat{\sigma}$) = *tríniterotolvuen*, and ProtRg(EffSpell($\hat{\sigma}$)) = 3.

Example 4.7.2. Let $\hat{\sigma} = DDT$, then EffSpell($\hat{\sigma}$) = *zxDDT*, and ProtRg(EffSpell($\hat{\sigma}$)) = 0; let $\hat{\sigma}$ = *1,1'-(2,2,2-trichloroethane-1,1-diyl)bis(4-chlorobenzene)* (the IUPAC name for DDT), then

$$\text{EffSpell}(\hat{\sigma}) = 1,1(2,2,2tričlôríeθaane1,1diiel)bi(4člôróóbenzeen),$$

and ProtRg(EffSpell($\hat{\sigma}$)) = 39. Here is how the effective spelling looks like with its first 39 characters underlined:

$$\underline{1,1(2,2,2tričlôríeθaane1,1diiel)bi(4člô}róóbenzeen).$$

We note that the string pattern of interest here is counted from an internal word boundary (the numerical digit 4).

Algorithm 4.8 (English essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of an English word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\text{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \text{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$.

(2) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*[43]

[41] Here, the status of heuristic capitalization affects our final treatments. If a word is heuristically capitalized, its final *e* (not preceded by *e* or *i*) will be altered to *ae*; if a word is heuristically uncapitalized, its final *e* (not preceded by *e* or *i*) will be removed. Furthermore, If a word is heuristically capitalized, its final *ie* will be changed to *y*.

[42] Roughly speaking, we want to protect an English word up to its first vowel (or vowel cluster) counting from an internal or external word boundary, bypassing some common prefixes.

[43] In other words, the core algorithm for essential root extraction runs as follows: keep the last "strong" vowel *a*, *ï*, *o* or *u* in non-final position, plus one subsequent letter; delete final *a*; erase the final appearance of *e* and all the letters thereafter.

(2.1) *Do* $\sim\hat{\chi}^{\in}(a|i|o|u)^{\mathbf{X}\in\overline{(a|i|o|u)}_m} \to \hat{\chi}\mathbf{X}_1$, *where* $\mathbf{X}_1$ *is the first character in* $\mathbf{X}$.

(2.2) *Do* $\sim a \to \varnothing$.

(2.3) *Do* $\sim e\overline{(e)}_{m_0} \to \varnothing$.

The result after these three steps of operations is called $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(4) *Do* $\sim pe\lambda \to pel$, $\underline{reebel} \to reebe\lambda$.

Example 4.8.1. Let $\hat{\sigma}$ be the IUPAC name for TNT, then EssRoot(EffSpell($\hat{\sigma}$)) = *tríniterotolvu*. Let be the IUPAC name for DDT, then EssRoot(EffSpell($\hat{\sigma}$)) = *1,1(2,2,2tričlóríeθaane1,1diiel)bi(4člóróóbenze)*. Note, in particular, that in the latter example, the last internal word boundary (closing parenthesis) is preserved in the essential root.

Example 4.8.2. Here are more examples showing how the functions RegVb, NormVL, EffSpell, EssRoot act on some etymologically related words.

$\hat{\sigma}$	RegVb($\hat{\sigma}$)	NormVL(RegVb($\hat{\sigma}$))	EffSpell($\hat{\sigma}$)	EssRoot(EffSpell($\hat{\sigma}$))
environment	*environment*	*environment*	*envïron*	*envïron*
environmental	*environmental*	*environmental*	*envïronmental*	*envïronm*
environmentalist	*environmentalist*	*environmentalist*	*envïronmentalist*	*envïronm*
protect	*protect*	*próótect*	*próótect*	*próótect*
protection	*protection*	*próótection*	*próótect*	*próótect*
protector	*protector*	*próótector*	*próótectôr*	*próótectôr*
protectorate	*protectorate*	*próótectóórate*	*próótectôrat*	*próótectôrat*

We note that the vowel *e* in *protect* had not been deleted in the construction of essential root. In fact, it remains in the unremovable $\hat{\sigma}_1$ in Algorithm 4.8(1).

4.2.2 Admissible mutation and approximate clustering

Unlike the Porter stemming algorithm, we shall not use the exact match of tokens as the sole criterion for word clustering. Instead, we shall sort words by their tokens, and compare consecutive neighbors in such an alphabetized list. Roughly speaking, our clustering algorithm (see Algorithm 4.15) runs as follows: if the tokens of two neighbors in such a list fail certain "approximate matching criteria", we add a demarcation line between them; after comparisons between all the consecutive neighbors, we split the alphabetized list into word clusters.

As noted before, some English irregular verbs exhibit systematic patterns of vowel alternations, such as *drink/drank/drunk* and *sing/sang/sung*. However, the tokens for the separate forms of an irregular verb may not sit next to each other in an alphabetized list. To resolve this issue in token sorting order, we need to blot certain vowels (more precisely, the extended vowels prescribed by Definition 4.6) occurring in our token, by the following algorithm.

Algorithm 4.9 (English vowel blotting)**.** *For a token string* $\hat{\sigma}$*, its blotted form* $\mathrm{BlotV}_1(\hat{\sigma})$ *is constructed as follows:*

- *If the first appearance of pattern* $\mathbf{V}_m^*$ *occurs neither word-initially nor word-finally (as measured by external word boundaries of* $\hat{\sigma}$*), then replace the aforementioned pattern with a single letter "a".*
- *Otherwise, leave the string* $\hat{\sigma}$ *intact.*

Example 4.9.1. As a continuation of Example 4.8.2, one can append one more column BlotV_1(EssRoot(EffSpell($\hat{\sigma}$))) with entries *envïron*, *envïronm*, *envïronm*, *pratect*, *pratect*, *pratectôr*, *pratectôrat*.

Example 4.9.2. A few more examples:

$\hat{\sigma}$	*automatic*	*bee*	*find*	*join*	*solo*
BlotV_1(EssRoot(EffSpell($\hat{\sigma}$)))	*aautóómat*	*bee*	*ffand*	*jzan*	*szalo*

Our "approximate matching criteria" are encoded in a bivariate Boolean-valued function HrdTest($\hat{\alpha}$,$\hat{\beta}$), which tests whether string $\hat{\beta}$ is a "legitimate heir" to string $\hat{\alpha}$. Before stating this "heredity test function" in its entirety (Algorithm 4.14), we shall describe a "simple heredity test function" in Algorithm 4.10, which judges similarities between two strings, without worrying about the problems associated with irregular verbs.

Algorithm 4.10 (Simple heredity test)**.** *Let $\hat{\alpha}'$ be the result of doing $\sim e \to \varnothing$, $\hat{\chi} lie(er|st) \to \hat{\chi}$ on $\hat{\alpha}$. The Boolean-valued function* SimpHrdTest($\hat{\alpha},\hat{\beta}$) *returns* **TRUE** *if the lowercase form of $\hat{\alpha}$ contains at least one instance of* $\mathbf{V}^*$ *(Definition 4.6)* **AND** *at least one of the following 12 conditions holds:*[44]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\alpha} = \hat{\beta}^{\circ}$, *where* $\hat{\beta}^{\circ}$ *results from doing* $\sim_{\hat{\chi}\notin}(a|e|i|o|u)ster \to \hat{\chi}$ *on* $\hat{\beta}$.

(iii) $\hat{\alpha} = \hat{\beta}^{*}$, *where* $\hat{\beta}^{*}$ *results from doing* $\hat{\chi} lie(er|st) \to \hat{\chi}$ *on* $\hat{\beta}$;

(iv) $\hat{\alpha} = \hat{\beta}^{\dagger}$, *where* $\hat{\beta}^{\dagger}$ *results from doing* $\sim ie \to i$ *on* $\hat{\beta}$;

(v) $\hat{\alpha} = \hat{\beta}^{\ddagger}$, *where* $\hat{\beta}^{\ddagger}$ *results from doing* $\sim us \to u$ *on* $\hat{\beta}$;

(vi) *Appending the last letter of string* $\hat{\alpha}$ *to itself, one obtains* $\hat{\beta}$;

(vii) $\hat{\beta} = \hat{\alpha}' i$;

(viii) $\hat{\beta} = \hat{\alpha} iv$;

(ix) $\hat{\beta} = \hat{\alpha} n$;

(x) $\hat{\beta} = (\hat{\alpha}|\hat{\alpha}')šip$;

(xi) $\hat{\beta} = \hat{\alpha}\theta$;

(xii) $\ell(\hat{\beta}) - 1 > \ell(\hat{\alpha}') \geq \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha} = \hat{\beta}^{[\ell(\hat{\alpha}')]}(\varnothing|e)$ **AND** $\hat{\alpha}'\mathbf{X}_1(gu|iô|ow|zβ)\mathbf{X}_2 \neq \hat{\beta}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha}')+1\}} = (a|e|h|i|o|ó|ô|r|u)$. *(See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

When SimpHrdTest($\hat{\alpha},\hat{\beta}$) = **FALSE**, it is still possible that the two underlying words are etymologically related. To detect affinity between words in such scenarios, we need to invoke the Needleman–Wunsch (NW) and Smith–Waterman (SW) algorithms for more detailed string comparisons (cf. Definition 3.7).

Algorithm 4.11 (Roots and suffixes by NW and SW)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the function values* RootNW($\hat{\alpha},\hat{\beta}$), SuffixNW($\hat{\alpha},\hat{\beta}$) *and* NW*($\hat{\alpha},\hat{\beta}$) *are determined through the following procedure:*

- *Use the Needleman–Wunsch algorithm to align the sequences as* NW($\hat{\alpha},\hat{\beta}$).
- *If* NW($\hat{\alpha},\hat{\beta}$) *ends with a mismatch (shown in brackets in Examples 3.7.1 and 3.7.2), use this mismatch to define* SuffixNW($\hat{\alpha},\hat{\beta}$)*; otherwise, define* SuffixNW($\hat{\alpha},\hat{\beta}$) = $[\varnothing,\varnothing]$.
- *If* NW($\hat{\alpha},\hat{\beta}$) *starts with a matching string, define* RootNW($\hat{\alpha},\hat{\beta}$) *as the concatenation of all the (not necessarily contiguous) matching string portions in* NW($\hat{\alpha},\hat{\beta}$)*, and define* NW*($\hat{\alpha},\hat{\beta}$) *by dropping the first matching string from* NW($\hat{\alpha},\hat{\beta}$) *as well as its contribution (if non-void) to* SuffixNW($\hat{\alpha},\hat{\beta}$)*; otherwise, define* RootNW($\hat{\alpha},\hat{\beta}$) = #% = U+0023 U+0025 *(a nonsensical string containing two characters) and define* NW*($\hat{\alpha},\hat{\beta}$) *by deleting from* NW($\hat{\alpha},\hat{\beta}$) *the contribution (if non-void) to* SuffixNW($\hat{\alpha},\hat{\beta}$)*;*

The function values RootSW($\hat{\alpha},\hat{\beta}$), SuffixSW($\hat{\alpha},\hat{\beta}$) *and* SW*($\hat{\alpha},\hat{\beta}$) *are determined similarly, through the* SW *function.*

Algorithm 4.12 (Admissible suffix mismatch)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmSM}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if the lowercase form of* RootNW($\hat{\alpha},\hat{\beta}$) *contains at least one instance of* $\mathbf{V}^*$ *(Definition 4.6)* **AND** NW*($\hat{\alpha},\hat{\beta}$) *is void* **AND** SuffixNW($\hat{\alpha},\hat{\beta}$) $\neq [(nd|ow|zβ), (nd|ow)]$[45] **AND** *at least one of the following four conditions holds:*

(i) SuffixNW($\hat{\alpha},\hat{\beta}$) = $[i, n]$;

(ii) SuffixNW($\hat{\alpha},\hat{\beta}$) = $[i, u]$;

(iii) SuffixNW($\hat{\alpha},\hat{\beta}$) = $[\varnothing, s]$ **AND** Ω(RootNW($\hat{\alpha},\hat{\beta}$)) = $í$;

(iv) SuffixNW($\hat{\alpha},\hat{\beta}$) = $[(\varnothing|i|((a|e|h)\mathbf{X})), (en|í|íz|îz\mathbf{X}|(i|ia|io|ió|iv|m|r|\hat{\chi}')\hat{\chi}\mathbf{X})]$ *where* $\hat{\chi}'$ = RootNW($\hat{\alpha},\hat{\beta}$)$^{\{\ell(\text{RootNW}(\hat{\alpha},\hat{\beta})\}}$ *is the last character in* RootNW($\hat{\alpha},\hat{\beta}$).

[44] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

[45] Hereafter, $[\hat{\sigma}_1, \hat{\sigma}_2] \neq [\hat{\tau}_1, \hat{\tau}_2]$ means ($\hat{\sigma}_1 \neq \hat{\tau}_1$ **AND** $\hat{\sigma}_2 \neq \hat{\tau}_2$), while $[\hat{\sigma}_1, \hat{\sigma}_2] = [\hat{\tau}_1, \hat{\tau}_2]$ means ($\hat{\sigma}_1 = \hat{\tau}_1$ **AND** $\hat{\sigma}_2 = \hat{\tau}_2$).

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmSM}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 4.13 (Admissible vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmVA}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^{*}(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if* $\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [(\varnothing|(e|i|n)\mathbf{X}), (\varnothing|(e|i|n|t|z)\mathbf{X})]$ **AND** $\text{NW}^{*}(\hat{\alpha},\hat{\beta})$ *consists of one mutation bracket and one non-void matching string* **AND** $\text{RootNW}(\hat{\alpha},\hat{\beta}) \neq \mathbf{X}_1(i|u)\mathbf{X}_2$ **AND** *at least one of the following three conditions holds:*

(i) $\text{NW}^{*}(\hat{\alpha},\hat{\beta}) = ([a,\ e]|[a,\ i]|[a,\ u]|[\ddot{a}a,\ \hat{o}]|[\ddot{a}a,\ \acute{o}\acute{o}]|[e,\ \varnothing]|[e,\ o]|[ee,\ oo]|[ee,\ \acute{o}\acute{o}]|[i,\ u]|[\acute{i},\ \hat{o}]|[\acute{i},\ \acute{o}\acute{o}]|[oo,\ \acute{o}\acute{o}])\mathbf{X}$

(ii) $\text{NW}^{*}(\hat{\alpha},\hat{\beta}) = [\varnothing,\ \hat{\chi}']\mathbf{X}$ *where* $\hat{\chi}' = \text{RootNW}(\hat{\alpha},\hat{\beta})^{\{\ell(\text{RootNW}(\hat{\alpha},\hat{\beta})\}}$ *is the last character in* $\text{RootNW}(\hat{\alpha},\hat{\beta})$*;*

(iii) $\text{NW}^{*}(\hat{\alpha},\hat{\beta}) = [\varnothing,\ e]e$.

Similarly, one can evaluate $\text{AdmVA}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta}))$.

Algorithm 4.14 (Heredity test function)**.** *Let $\hat{\alpha}'$ be the result of performing a substitution $\sim e \rightarrow \varnothing$ on $\hat{\alpha}$. For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function* $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\text{SimpHrdTest}(\hat{\alpha},\hat{\beta}) =$ **TRUE***;*

(ii) $\ell(\text{RootNW}(\hat{\alpha},\hat{\beta})) \geq \frac{1}{2}\max\{\ell(\hat{\alpha}'),\ell(\hat{\beta})\}$ **AND** $(\text{AdmSM}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^{*}(\hat{\alpha},\hat{\beta})) =$ **TRUE OR** $\text{AdmVA}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^{*}(\hat{\alpha},\hat{\beta})) =$ **TRUE**)*;*

(iii) $\ell(\text{RootSW}(\hat{\alpha},\hat{\beta})) \geq \frac{1}{2}\max\{\ell(\hat{\alpha}'),\ell(\hat{\beta})\}$ **AND** $(\text{AdmSM}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta})) =$ **TRUE OR** $\text{AdmVA}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta})) =$ **TRUE**).

Algorithm 4.15 (Approximate clustering of English words)**.** *The approximate clustering of a list of English words $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1,\text{EffSpell}(\hat{\alpha}_1)), \ldots, (\hat{\alpha}_N,\text{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the second component (effective spelling) of each entry. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)},\text{EffSpell}(\hat{\alpha}_{(1)})), \ldots, (\hat{\alpha}_{(N)},\text{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfy* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n)}),\text{EffSpell}(\hat{\alpha}_{(n+1)})) =$ **FALSE**, *where $n \in \mathbb{Z} \cap [1, N)$, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)},\text{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)},\text{EffSpell}(\hat{\alpha}_{(1,n_1)}))\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, \text{EffSpell}(\hat{\alpha}_{(M,1)})), \ldots, (\hat{\alpha}_{(M,n_M)},\text{EffSpell}(\hat{\alpha}_{(M,n_M)}))\}\}$*, where each sublist also preserves alphabetic order. In particular, we have $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ and $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.*

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)},\text{EffSpell}(\hat{\alpha}_{(m,1)})), \ldots, (\hat{\alpha}_{(m,n_m)},\text{EffSpell}(\hat{\alpha}_{(m,n_m)}))\}$ *where $m \in \mathbb{Z} \cap [1, M]$, we augment it into a tagged entry* $G_m = (g_m,\text{EffSpell}(\hat{\alpha}_{(m,1)}),\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})),\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})))$ *(with highest priority),* $\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with medium priority), and* $\text{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)},\hat{\gamma}'_{(1)},\hat{\gamma}''_{(1)},\hat{\gamma}'''_{(1)}), \ldots, G_{(M)} = (g_{(M)},\hat{\gamma}'_{(M)},\hat{\gamma}''_{(M)},\hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\text{HrdTest}(\hat{\gamma}''_{(m)},\hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)},\hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, then add a demarcation line between these two entries. In this manner, the alphabetized list $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots, vowel blotted forms) from each Γ_k, where $k \in \mathbb{Z} \cap [1, K]$.*

Example 4.15.1. To partially demonstrate the capabilities of our clustering algorithm, we pick the following list of English words:

abolish, *abolished*, *abolishes*, *abolishing*, *abolishment*, *abolition*, *abolitionism*, *abolitionist*, *abolitionists*, *admonish*, *admonishable*, *admonisher*, *admonishes*, *admonishing*, *admonishingly*, *admonishment*, *admonition*, *admonitory*, *aid*, *aided*, *aids*, *America*, *American*, *analyses*, *analysis*, *analytic*, *analytically*, *analyze*, *analyzed*, *ant*, *antenna*, *antennae*, *anterior*, *anxieties*, *anxiety*, *anxious*, *anxiously*, *arise*, *arisen*, *arose*, *automatic*, *automatically*, *bad*, *bat*, *beauties*, *beautiful*, *beautifully*, *beauty*, *bed*, *bedding*, *beds*, *beg*, *began*, *beggar*, *begged*, *begin*, *beginning*, *begun*, *belie*, *belied*, *belief*, *beliefs*, *belies*, *believe*, *believed*, *believes*, *believing*, *bereave*, *bereft*, *best*, *bestow*, *bestowed*, *bet*, *better*, *bid*, *bit*, *bitten*, *break*, *breaking*, *breaks*, *Britain*, *British*, *broke*, *broken*, *brow*, *brows*, *browse*, *browsed*, *browses*, *browsing*, *brutal*, *brutality*, *brutalize*, *brutally*, *brute*, *brutish*, *child*, *childhood*, *children*, *children's*, *child's*, *conquer*, *conquered*, *conquest*, *cycle*, *cylinder*, *dance*, *danced*, *dances*, *dancing*, *dead*, *deal*, *dealt*, *death*, *destroy*, *destroyed*, *destruction*, *die*, *died*, *dies*, *differ*, *difference*, *different*, *difficult*, *difficulty*, *dosage*, *dose*, *drain*, *drainage*, *dried*, *dries*, *dry*, *dryly*, *dye*, *dyed*, *dyeing*, *dyes*, *dying*, *elect*, *elected*, *election*, *elections*, *electoral*, *emphases*, *emphasis*, *emphasize*, *emphasized*, *emphasizes*, *emphatic*, *employ*, *employed*, *employee*, *employer*, *employment*, *England*, *English*, *enjoy*, *enjoyed*, *enjoying*, *enjoyment*, *enjoys*, *environment*, *environmental*, *environmentalist*, *establish*, *established*, *establishes*, *establishing*, *establishment*, *estimate*, *estimated*, *estimates*, *estimation*, *fall*, *fallen*, *falling*, *falls*, *fame*, *family*, *famous*, *fat*, *feel*, *feeling*, *feels*, *fell*, *fellow*, *fellows*, *fellowship*, *felt*, *fight*, *fighting*, *fill*, *find*, *fit*, *fold*, *follow*, *fond*, *fought*, *found*, *fret*, *fretting*, *full*, *fund*, *funding*, *gentleman*, *gentlemen*, *German*, *Germany*, *good*, *happen*, *happened*, *happening*, *happens*, *happier*, *happiest*, *happily*, *happiness*, *happy*, *hat*, *hate*, *hated*, *hating*, *hatred*, *hats*, *heavy*, *heft*, *held*, *hell*, *hill*, *hillside*, *hit*, *hits*, *hitting*, *hold*, *increase*, *increased*, *increases*, *increasing*, *incredible*, *incredulity*, *incredulous*, *infer*, *inference*, *inferior*, *inferiority*, *inferred*, *infers*, *insist*, *insisted*, *insistence*, *integral*, *integrally*, *integrate*, *integration*, *interfere*, *interference*, *interfering*, *introduce*, *introduced*, *introduces*, *introduction*, *introductory*, *lack*, *lacking*, *ladies*, *ladies'*, *lady*, *lady's*, *law*, *lawyer*, *leave*, *leaves*, *leaving*, *left*, *lick*, *lie*, *lied*, *lies*, *life*, *life's*, *lift*, *lived*, *lives*, *living*, *lock*, *love*, *loved*, *lovely*, *loves*, *loving*, *low*, *lower*, *lowest*, *lowing*, *lowly*, *luck*, *luckily*, *lucky*, *lying*, *maid*, *maiden*, *maids*, *male*, *man*, *manage*, *manipulate*, *mankind*, *manner*, *manners*, *man's*, *mansion*, *mansions*, *marriage*, *married*, *marries*, *marrow*, *marry*, *marrying*, *men*, *mend*, *men's*, *merry*, *mile*, *mind*, *mole*, *mule*, *natural*, *naturally*, *nature*, *nature's*, *paid*, *pain*, *painful*, *painfully*, *pay*, *paying*, *payment*, *payments*, *pays*, *plan*, *planned*, *planning*, *plans*, *plant*, *plantation*, *play*, *played*, *player*, *players*, *playing*, *plays*, *presidency*, *president*, *presidential*, *presidents*, *prince*, *princes*, *princess*, *princesses*, *protect*, *protection*, *protector*, *protectoral*, *protectorate*, *ran*, *rang*, *range*, *ranged*, *ranges*, *ranging*, *rank*, *rant*, *rate*, *rated*, *rates*, *rating*, *real*, *realization*, *realize*, *refer*, *reference*, *referral*, *rid*, *ridden*, *ride*, *ring*, *ringing*, *rings*, *road*, *rode*, *rude*, *ruin*, *ruined*, *ruins*, *run*, *rung*, *running*, *runs*, *Russell*, *Russia*, *Russian*, *sad*, *sadden*, *saddened*, *sadly*, *said*, *sang*, *sat*, *say*, *saying*, *says*, *secede*, *seceded*, *secedes*, *seceding*, *secession*, *secessionist*, *secessionists*, *secessions*, *sell*, *selling*, *send*, *sending*, *sent*, *sentence*, *sentences*, *sentiment*, *sentimental*, *sentiments*, *set*, *sets*, *setting*, *sing*, *singing*, *sit*, *sits*, *sitting*, *slain*, *slave*, *slave-holder*, *slave-holding*, *slavery*, *slavery-restricting*, *slaves*, *slay*, *slow*, *slowly*, *sold*, *solution*, *solutions*, *solve*, *solved*, *solves*, *son*, *song*, *sons*, *sore*, *sorely*, *sorrow*, *sorry*, *sort*, *speak*, *speaker*, *speaking*, *speaks*, *speech*, *speeches*, *spoke*, *spoken*, *spokesman*, *strata*, *stratum*, *strife*, *strifes*, *strive*, *strived*, *striven*, *strove*, *sun*, *sung*, *swam*, *swear*, *swell*, *swelled*, *swelling*, *swells*, *swim*, *swimming*, *swims*, *swollen*, *swoon*, *swooned*, *swoons*, *sword*, *sworn*, *swum*, *tall*, *tame*, *tamed*, *tax*, *taxes*, *teeth*, *tell*, *telling*, *thank*, *thanking*, *thanks*, *theft*, *thief*, *thieve*, *thieved*, *thieves*, *thin*, *thing*, *things*, *think*, *thinking*, *thrift*, *thrifty*, *thrive*, *thrived*, *thriven*, *thrives*, *throve*, *time*, *times*, *timing*, *told*, *tooth*, *traffic*, *trafficking*, *transfer*, *transference*, *transferral*, *want*, *wear*, *weave*, *weep*, *weeping*, *weeps*, *weft*, *wept*, *wet*, *wife*, *wife's*, *winter*, *wipe*, *wiped*, *wipes*, *wit*, *wives*, *woman*, *woman's*, *women*, *women's*, *word*, *words*, *wore*, *world*, *worn*, *woven*.

Applying Algorithm 4.15 to the list above, we arrive at the following result

{*abolish*, *abolished*, *abolishes*, *abolishing*, *abolishment*, *abolition*, *abolitionism*, *abolitionist*, *abolitionists*}, {*admonish*, *admonishable*, *admonisher*, *admonishes*, *admonishing*, *admonishingly*, *admonishment*}, {*admonition*, *admonitory*}, {*aid*, *aided*, *aids*}, {*America*, *American*}, {*analyses*, *analysis*, *analytic*, *analytically*, *analyze*, *analyzed*}, {*ant*}, {*antenna*, *antennae*}, {*anterior*}, {*anxieties*, *anxiety*, *anxious*, *anxiously*}, {*arise*, *arisen*, *arose*}, {*automatic*, *automatically*}, {*bad*}, {*bat*}, {*beauties*, *beautiful*, *beautifully*, *beauty*}, {*bed*, *bedding*, *beds*}, {*beg*, *begged*}, {*began*, *begin*, *beginning*, *begun*}, {*beggar*}, {*belie*, *belied*, *belies*}, {*belief*, *beliefs*, *believe*, *believed*, *believes*, *believing*}, {*bereave*, *bereft*}, {*best*, *better*, *good*}, {*bestow*, *bestowed*}, {*bet*}, {*bid*}, {*bit*}, {*bitten*}, {*break*, *breaking*, *breaks*, *broke*, *broken*}, {*Britain*, *British*}, {*brow*, *brows*}, {*browse*, *browsed*, *browses*, *browsing*}, {*brutal*, *brutality*, *brutalize*, *brutally*, *brute*, *brutish*}, {*child*, *child's*, *childhood*, *children*, *children's*}, {*conquer*, *conquered*}, {*conquest*}, {*cycle*}, {*cylinder*}, {*dance*, *danced*, *dances*, *dancing*}, {*dead*, *death*, *die*, *died*, *dies*, *dying*}, {*deal*, *dealt*}, {*destroy*, *destroyed*, *destruction*}, {*differ*, *difference*, *different*}, {*difficult*, *difficulty*}, {*dosage*, *dose*}, {*drain*, *drainage*}, {*dried*, *dries*, *dry*, *dryly*}, {*dye*, *dyed*, *dyeing*, *dyes*}, {*elect*, *elected*, *election*, *elections*, *electoral*}, {*emphases*, *emphasis*, *emphasize*, *emphasized*, *emphasizes*, *emphatic*}, {*employ*, *employed*, *employee*, *employer*, *employment*}, {*England*, *English*}, {*enjoy*, *enjoyed*, *enjoying*, *enjoyment*, *enjoys*}, {*environ-*

ment, *environmental*, *environmentalist*}, {*establish*, *established*, *establishes*, *establishing*, *establishment*}, {*estimate*, *estimated*, *estimates*, *estimation*}, {*fall*, *fallen*, *falling*, *falls*, *fell*}, {*fame*, *famous*}, {*family*}, {*fat*}, {*feel*, *feeling*, *feels*, *felt*}, {*fellow*, *fellows*, *fellowship*}, {*fight*, *fighting*, *fought*}, {*fill*}, {*find*, *found*}, {*fit*}, {*fold*}, {*follow*}, {*fond*}, {*fret*, *fretting*}, {*full*}, {*fund*, *funding*}, {*gentleman*, *gentlemen*}, {*German*, *Germany*}, {*happen*, *happened*, *happening*, *happens*}, {*happier*, *happiest*, *happily*, *happiness*, *happy*}, {*hat*, *hats*}, {*hate*, *hated*, *hating*, *hatred*}, {*heavy*, *heft*}, {*held*, *hold*}, {*hell*}, {*hill*}, {*hillside*}, {*hit*, *hits*, *hitting*}, {*increase*, *increased*, *increases*, *increasing*}, {*incredible*}, {*incredulity*, *incredulous*}, {*infer*, *inference*, *inferred*, *infers*}, {*inferior*, *inferiority*}, {*insist*, *insisted*, *insistence*}, {*integral*, *integrally*, *integrate*, *integration*}, {*interfere*, *interference*, *interfering*}, {*introduce*, *introduced*, *introduces*, *introduction*, *introductory*}, {*lack*, *lacking*}, {*ladies*, *ladies'*, *lady*, *lady's*}, {*law*}, {*lawyer*}, {*leave*, *leaves*, *leaving*, *left*}, {*lick*}, {*lie*, *lied*, *lies*, *lying*}, {*life*, *life's*, *lived*, *lives*, *living*}, {*lift*}, {*lock*}, {*love*, *loved*, *lovely*, *loves*, *loving*}, {*low*, *lower*, *lowest*, *lowing*, *lowly*}, {*luck*, *luckily*, *lucky*}, {*maid*, *maiden*, *maids*}, {*male*}, {*man*, *man's*, *men*, *men's*}, {*manage*}, {*manipulate*}, {*mankind*}, {*manner*, *manners*}, {*mansion*, *mansions*}, {*marriage*, *married*, *marries*, *marry*, *marrying*}, {*marrow*}, {*mend*}, {*merry*}, {*mile*}, {*mind*}, {*mole*}, {*mule*}, {*natural*, *naturally*, *nature*, *nature's*}, {*paid*, *pay*, *paying*, *payment*, *payments*, *pays*}, {*pain*, *painful*, *painfully*}, {*plan*, *planned*, *planning*, *plans*}, {*plant*, *plantation*}, {*play*, *played*, *player*, *players*, *playing*, *plays*}, {*presidency*, *president*, *presidential*, *presidents*}, {*prince*, *princes*, *princess*, *princesses*}, {*protect*, *protection*, *protector*, *protectoral*, *protectorate*}, {*ran*, *run*, *running*, *runs*}, {*rang*, *ring*, *ringing*, *rings*, *rung*}, {*range*, *ranged*, *ranges*, *ranging*}, {*rank*}, {*rant*}, {*rate*, *rated*, *rates*, *rating*}, {*real*, *realization*, *realize*}, {*refer*, *reference*, *referral*}, {*rid*}, {*ridden*, *ride*, *rode*}, {*road*}, {*rude*}, {*ruin*, *ruined*, *ruins*}, {*Russell*, *Russia*, *Russian*}, {*sad*, *sadden*, *saddened*, *sadly*}, {*said*, *say*, *saying*, *says*}, {*sang*, *sing*, *singing*, *sung*}, {*sat*, *sit*, *sits*, *sitting*}, {*secede*, *seceded*, *secedes*, *seceding*, *secession*, *secessionist*, *secessionists*, *secessions*}, {*sell*, *selling*, *sold*}, {*send*, *sending*, *sent*}, {*sentence*, *sentences*}, {*sentiment*, *sentimental*, *sentiments*}, {*set*, *sets*, *setting*}, {*slain*, *slay*}, {*slave*, *slave-holder*, *slave-holding*, *slavery*, *slavery-restricting*, *slaves*}, {*slow*, *slowly*}, {*solution*, *solutions*, *solve*, *solved*, *solves*}, {*son*, *sons*}, {*song*}, {*sore*, *sorely*, *sorry*}, {*sorrow*}, {*sort*}, {*speak*, *speaker*, *speaking*, *speaks*, *spoke*, *spoken*, *spokesman*}, {*speech*, *speeches*}, {*strata*, *stratum*}, {*strife*, *strifes*, *strive*, *strived*, *striven*, *strove*}, {*sun*}, {*swam*, *swim*, *swimming*, *swims*, *swum*}, {*swear*, *sworn*}, {*swell*, *swelled*, *swelling*, *swells*, *swollen*}, {*swoon*, *swooned*, *swoons*}, {*sword*}, {*tall*}, {*tame*, *tamed*}, {*tax*, *taxes*}, {*teeth*, *tooth*}, {*tell*, *telling*, *told*}, {*thank*, *thanking*, *thanks*}, {*theft*, *thief*, *thieve*, *thieved*, *thieves*}, {*thin*}, {*thing*, *things*}, {*think*, *thinking*}, {*thrift*, *thrifty*}, {*thrive*, *thrived*, *thriven*, *thrives*, *throve*}, {*time*, *times*, *timing*}, {*traffic*, *trafficking*}, {*transfer*, *transference*, *transferral*}, {*want*}, {*wear*, *wore*, *worn*}, {*weave*, *weft*, *woven*}, {*weep*, *weeping*, *weeps*, *wept*}, {*wet*}, {*wife*, *wife's*, *wives*}, {*winter*}, {*wipe*, *wiped*, *wipes*}, {*wit*}, {*woman*, *woman's*, *women*, *women's*}, {*word*, *words*}, {*world*},

which contains only very few errors. For such a list of test words, the time cost of our algorithm is only twice as long as conventional Porter stemming[46], the latter of which yields the following clustering result:

{*abolish*, *abolished*, *abolishes*, *abolishing*, *abolishment*}, {*abolition*}, {*abolitionism*}, {*abolitionist*, *abolitionists*}, {*admonish*, *admonishable*, *admonisher*, *admonishes*, *admonishing*, *admonishment*}, {*admonishingly*}, {*admonition*}, {*admonitory*}, {*aid*, *aided*, *aids*}, {*America*}, {*American*}, {*analyses*}, {*analysis*}, {*analytic*, *analytically*}, {*analyze*, *analyzed*}, {*ant*}, {*antenna*, *antennae*}, {*anterior*}, {*anxieties*, *anxiety*}, {*anxious*}, {*anxiously*}, {*arise*}, {*arisen*}, {*arose*}, {*automatic*, *automatically*}, {*bad*}, {*bat*}, {*beauties*, *beautiful*, *beauty*}, {*beautifully*}, {*bed*, *bedding*, *beds*}, {*beg*, *begged*}, {*began*}, {*beggar*}, {*begin*, *beginning*}, {*begun*}, {*belie*, *belied*, *belies*}, {*belief*, *beliefs*}, {*believe*, *believed*, *believes*, *believing*}, {*bereave*}, {*bereft*}, {*best*}, {*bestow*, *bestowed*}, {*bet*}, {*better*}, {*bid*}, {*bit*}, {*bitten*}, {*break*, *breaking*, *breaks*}, {*Britain*}, {*British*}, {*broke*}, {*broken*}, {*brow*, *brows*}, {*browse*, *browsed*, *browses*, *browsing*}, {*brutal*, *brutality*, *brutalize*, *brutally*}, {*brute*}, {*brutish*}, {*child*}, {*childhood*}, {*children*}, {*children's*}, {*child's*}, {*conquer*, *conquered*}, {*conquest*}, {*cycle*}, {*cylinder*}, {*dance*, *danced*, *dances*, *dancing*}, {*dead*}, {*deal*}, {*dealt*}, {*death*}, {*destroy*, *destroyed*}, {*destruction*}, {*die*}, {*died*, *dies*}, {*differ*, *difference*, *different*}, {*difficult*}, {*difficulty*}, {*dosage*}, {*dose*}, {*drain*}, {*drainage*}, {*dried*, *dries*}, {*dry*}, {*dryly*}, {*dye*, *dyeing*, *dyes*}, {*dyed*, *dying*}, {*elect*, *elected*, *election*, *elections*}, {*electoral*}, {*emphases*, *emphasize*, *emphasized*, *emphasizes*}, {*emphasis*}, {*emphatic*}, {*employ*, *employed*}, {*employee*}, {*employer*, *employment*}, {*England*}, {*English*}, {*enjoy*, *enjoyed*, *enjoying*, *enjoys*}, {*enjoyment*}, {*environment*}, {*environmental*}, {*environmentalist*}, {*establish*, *established*, *establishes*, *establishing*, *establishment*}, {*estimate*, *estimated*, *estimates*, *estimation*}, {*fall*, *falling*, *falls*}, {*fallen*}, {*fame*}, {*family*}, {*famous*}, {*fat*}, {*feel*, *feeling*, *feels*}, {*fell*}, {*fellow*, *fellows*}, {*fellowship*}, {*felt*}, {*fight*, *fighting*}, {*fill*}, {*find*}, {*fit*}, {*fold*}, {*follow*}, {*fond*}, {*fought*}, {*found*}, {*fret*, *fretting*}, {*full*}, {*fund*, *funding*}, {*gentleman*}, {*gentlemen*}, {*German*}, {*Germany*}, {*good*}, {*happen*, *happened*, *happening*, *happens*}, {*happier*}, {*happiest*}, {*happily*}, {*happiness*, *happy*}, {*hat*, *hats*}, {*hate*, *hated*, *hating*}, {*hatred*}, {*heavy*}, {*heft*}, {*held*}, {*hell*},

[46] We use the Porter stemming functionality built in *Mathematica* v11.0 for this numerical experiment. Certain variations on the Porter stemming algorithm may have better performances in word clustering.

{*hill*}, {*hillside*}, {*hit*, *hits*, *hitting*}, {*hold*}, {*increase*, *increased*, *increases*, *increasing*}, {*incredible*}, {*incredulity*, *incredulous*}, {*infer*, *inference*, *inferred*, *infers*}, {*inferior*, *inferiority*}, {*insist*, *insisted*, *insistence*}, {*integral*, *integrally*, *integrate*, *integration*}, {*interfere*, *interference*}, {*interfering*}, {*introduce*, *introduced*, *introduces*}, {*introduction*}, {*introductory*}, {*lack*, *lacking*}, {*ladies*, *lady*}, {*ladies'*}, {*lady's*}, {*law*}, {*lawyer*}, {*leave*, *leaves*, *leaving*}, {*left*}, {*lick*}, {*lie*}, {*lied*, *lies*}, {*life*}, {*life's*}, {*lift*}, {*lived*, *lives*, *living*}, {*lock*}, {*love*, *loved*, *lovely*, *loves*, *loving*}, {*low*, *lowing*}, {*lower*}, {*lowest*}, {*lowly*}, {*luck*}, {*luckily*}, {*lucky*}, {*lying*}, {*maid*, *maids*}, {*maiden*}, {*male*}, {*man*}, {*manage*}, {*manipulate*}, {*mankind*}, {*manner*, *manners*}, {*man's*}, {*mansion*, *mansions*}, {*marriage*}, {*married*, *marries*, *marry*, *marrying*}, {*marrow*}, {*men*}, {*mend*}, {*men's*}, {*merry*}, {*mile*}, {*mind*}, {*mole*}, {*mule*}, {*natural*, *naturally*, *nature*}, {*nature's*}, {*paid*}, {*pain*, *painful*}, {*painfully*}, {*pay*, *paying*, *pays*}, {*payment*, *payments*}, {*plan*, *planned*, *planning*, *plans*}, {*plant*}, {*plantation*}, {*play*, *played*, *playing*, *plays*}, {*player*, *players*}, {*presidency*, *president*, *presidents*}, {*presidential*}, {*prince*, *princes*}, {*princess*, *princesses*}, {*protect*, *protection*}, {*protector*, *protectoral*, *protectorate*}, {*ran*}, {*rang*, *range*, *ranged*, *ranges*, *ranging*}, {*rank*}, {*rant*}, {*rate*, *rated*, *rates*, *rating*}, {*real*}, {*realization*, *realize*}, {*refer*, *reference*}, {*referral*}, {*rid*}, {*ridden*}, {*ride*}, {*ring*, *ringing*, *rings*}, {*road*}, {*rode*}, {*rude*}, {*ruin*, *ruined*, *ruins*}, {*run*, *running*, *runs*}, {*rung*}, {*Russell*}, {*Russia*}, {*Russian*}, {*sad*}, {*sadden*, *saddened*}, {*sadly*}, {*said*}, {*sang*}, {*sat*}, {*say*, *saying*, *says*}, {*secede*, *seceded*, *secedes*, *seceding*}, {*secession*, *secessions*}, {*secessionist*, *secessionists*}, {*sell*, *selling*}, {*send*, *sending*}, {*sent*}, {*sentence*, *sentences*}, {*sentiment*, *sentimental*, *sentiments*}, {*set*, *sets*, *setting*}, {*sing*, *singing*}, {*sit*, *sits*, *sitting*}, {*slain*}, {*slave*, *slaves*}, {*slave-holder*, *slave-holding*}, {*slavery*}, {*slavery-restricting*}, {*slay*}, {*slow*}, {*slowly*}, {*sold*}, {*solution*, *solutions*}, {*solve*, *solved*, *solves*}, {*son*, *sons*}, {*song*}, {*sore*, *sorely*}, {*sorrow*}, {*sorry*}, {*sort*}, {*speak*, *speaking*, *speaks*}, {*speaker*}, {*speech*, *speeches*}, {*spoke*}, {*spoken*}, {*spokesman*}, {*strata*}, {*stratum*}, {*strife*, *strifes*}, {*strive*, *strived*}, {*striven*}, {*strove*}, {*sun*}, {*sung*}, {*swam*}, {*swear*}, {*swell*, *swelled*, *swelling*, *swells*}, {*swim*, *swimming*, *swims*}, {*swollen*}, {*swoon*, *swooned*, *swoons*}, {*sword*}, {*sworn*}, {*swum*}, {*tall*}, {*tame*, *tamed*}, {*tax*, *taxes*}, {*teeth*}, {*tell*, *telling*}, {*thank*, *thanking*, *thanks*}, {*theft*}, {*thief*}, {*thieve*, *thieved*, *thieves*}, {*thin*}, {*thing*, *things*}, {*think*, *thinking*}, {*thrift*}, {*thrifty*}, {*thrive*, *thrived*, *thrives*}, {*thriven*}, {*throve*}, {*time*, *times*, *timing*}, {*told*}, {*tooth*}, {*traffic*}, {*trafficking*}, {*transfer*, *transference*}, {*transferral*}, {*want*}, {*wear*}, {*weave*}, {*weep*, *weeping*, *weeps*}, {*weft*}, {*wept*}, {*wet*}, {*wife*}, {*wife's*}, {*winter*}, {*wipe*, *wiped*, *wipes*}, {*wit*}, {*wives*}, {*woman*}, {*woman's*}, {*women*}, {*women's*}, {*word*, *words*}, {*wore*}, {*world*}, {*worn*}, {*woven*},

containing more errors than our output.

Example 4.15.2. In Fig. S3, we apply our modified Porter stemming algorithm to topic mining from three English masterpieces (see Table S1 for provenances), and compare the results with ground truths.

According to our stochastic model for recurrence time statistics, word patterns should stay either around the critical line of Poissonian banality (colored gray in Fig. S3) or below it (colored red in Fig. S3). The pattern *Eliza*(∅|*beth*)(∅|*'s*) in Fig. S3a is a prominent counterexample to this general rule. Our explanation for this is the aliasing of "Elizabeth" by pronouns and nicknames.

Furthermore, if a language lacks gender distinction in third-person pronouns, then its speaker might sometimes spell out a specific person's name to avoid confusion. This practice also complicates the statistical behavior of *Eliza*(∅|*beth*)(∅|*'s*) in translations of *Pride and Prejudice*. In Fig. S3e, we see that in four versions (Hungarian, Basque, Korean and Turkish), our statistical algorithm fails to identify the equivalent of *Eliza*(∅|*beth*)(∅|*'s*) as a topic, and the corresponding word counts also deviate significantly from the English original. These four languages, together with Finnish, do not have separate words for "he" and "she".[47] The use of pronouns (or possessive suffixes, or personal verb markers, or explicit names) in these five languages can be demonstrated by the following excerpt (see Table S1 for provenances) from Chapter 43 of *Pride and Prejudice*:

English

Mrs. Gardiner was standing a little behind; and on her pausing, he asked her if she would do him the honour of introducing him to her friends.

Finnish

Gardinerit olivat jääneet vähän taemmaksi; ja kun Elizabethilta puhe juuttui, pyysi Darcy tulla esitetyksi hänen ystävilleen.

Hungarian

Mrs. Gardiner kissé hátrább maradt, és amikor Elizabeth elhallgatott, Darcy megkérte, legyen szíves bemutatni őt ismerőseinek.

[47] This statement is true for colloquial Korean, where 그 means both "he" and "she". In written Korean texts that are translations of certain European languages, the word 그녀 is sometimes used for "she". This explains why the Korean word count statistics conform to the English original for *Eliza*(∅|*beth*)(∅|*'s*) better than Hungarian, Basque and Turkish.

Basque

Gardiner anderea atzeratxoago zegoen; eta Elizabeth isildu zelarik, Darcy jaunak galde egin zion ea ez al zion egingo lagun horien aurrean bera aurkezteko ohorea.

Korean

가디너 부인은 조금 뒤에 서 있었다. 엘리자베스가 말을 멈추자 다시 씨는 그녀에게 친구 분들을 소개하는 영광을 베풀어 주지 않겠냐고 물었다.

Turkish

Mrs. Gardiner biraz arkada duruyordu; Elizabeth'in duraklamasıyla Mr. Darcy, arkadaşlarıyla tanışma şerefine erişmeyi rica etti.

Here, in the parallel texts given above, the magenta parts point to Elizabeth herself, either in name or by coreference; the blue parts refer to Mr. Darcy. Agglutinative case prefixes/suffixes (other than those personal markers) are not shown in color. To avoid potential misinterpretations of pronominal coreferences, all the non-Indo-European translations above explicitly mentioned the names Elizabeth and Darcy as antecedents, and the subsequent pronouns (or personal markers) referred to these antecedents, according to their order of appearance in the translated texts.

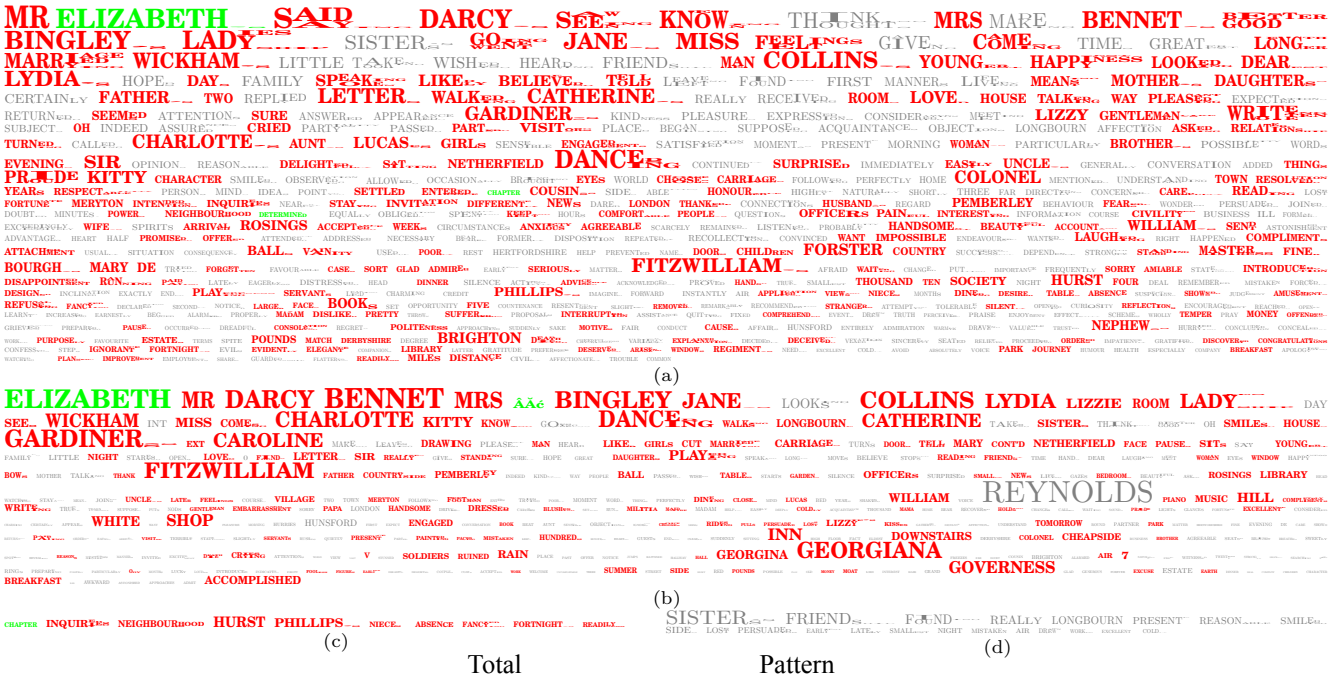


Fig. S3. Text mining in English. (a) Word patterns W_i in Jane Austen's *Pride and Prejudice*, sorted by descending $n_{ii} \geq 20$. Font sizes are set to visualize word statistics (cf. [1, Fig. 3]). Topical (*i.e.* significantly non-Poissonian) patterns painted in *red* (resp. *green*) reside below (resp. above) the critical line of Poissonian banality (*blue line* in [1, Fig. 2e]), where the deviations exceed the error margin prescribed in (1.5). (b) Word patterns W_i in Deborah Moggach's screenplay adaptation, sorted by descending $n_{ii} \geq 5$. (c) False positives: automatically extracted topical patterns (colored *green* or *red* in panel a) that are not mentioned anywhere in the entire screenplay. (d) False negatives: automatically identified non-topic patterns (*gray* in panel a) that are significantly non-Poissonian in the screenplay (colored *green* or *red* in panel b). (e) Counts of the pattern *Eliza*(∅|*beth*)(∅|*'s*) in 61 chapters from various versions of *Pride and Prejudice*.

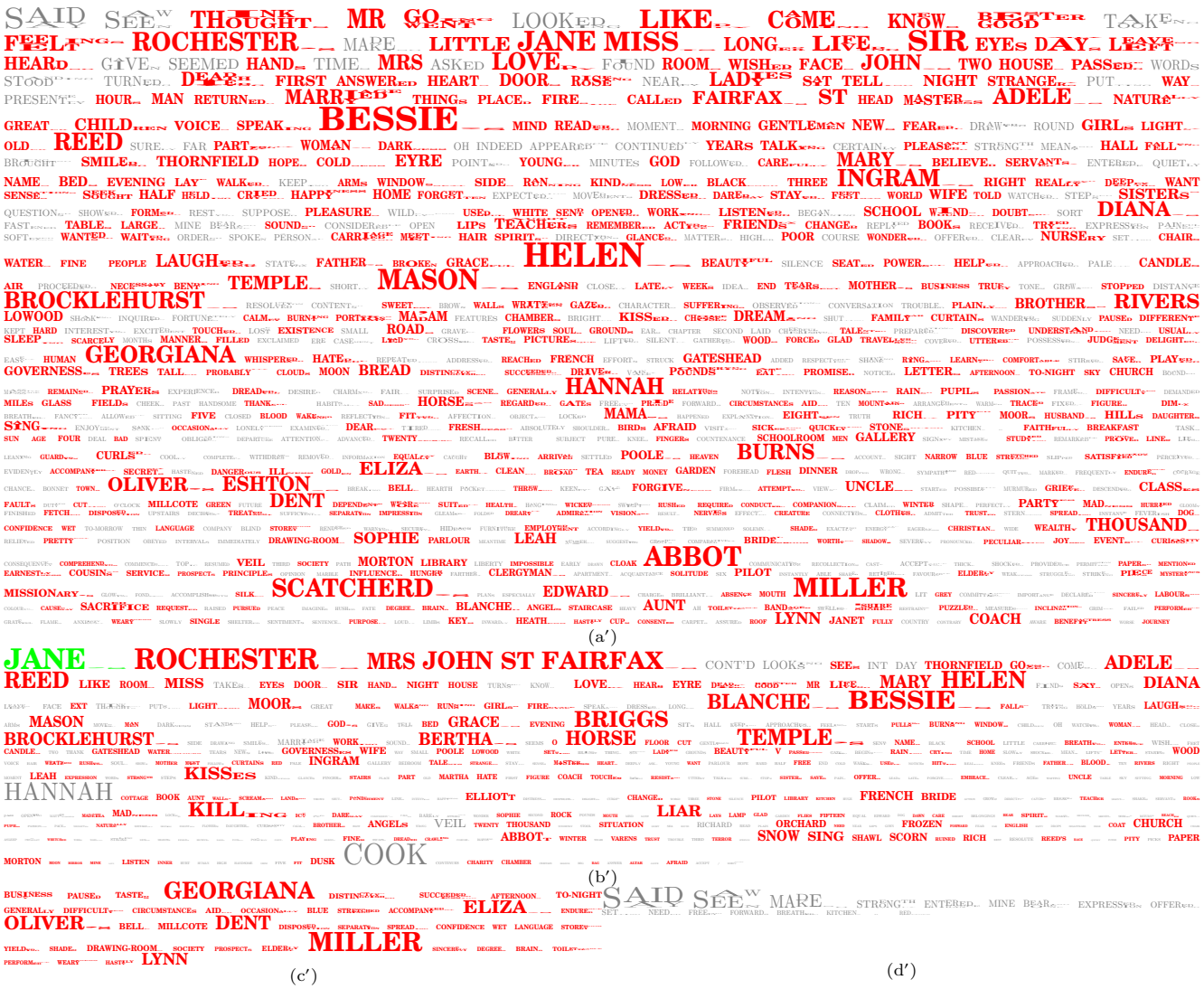
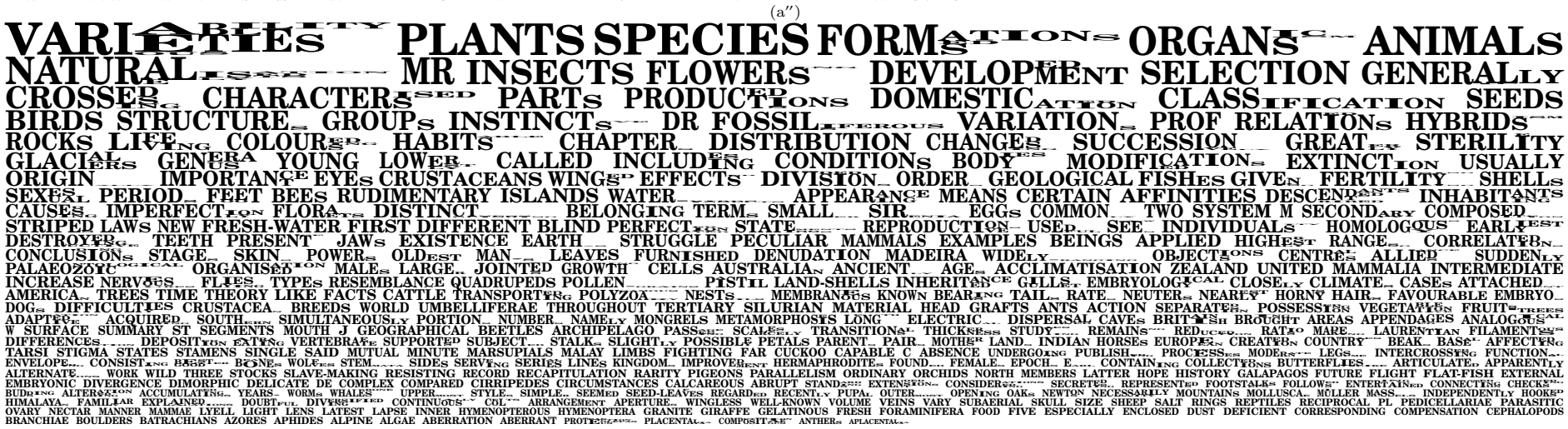
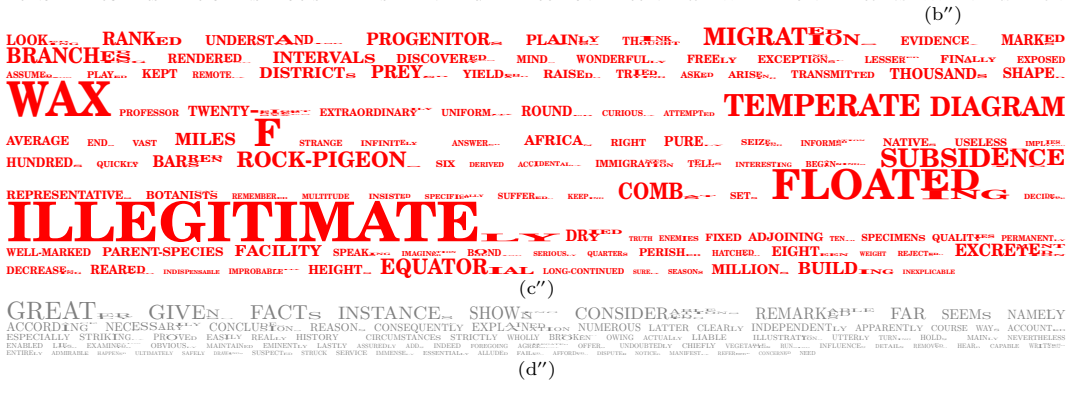


Fig. S3. Text mining in English. (*Continued*) A similar service on Charlotte Brontë's *Jane Eyre*. Moira Buffini's screenplay adaptation is chosen as the ground truth for topic extraction.

(a″)



(b″)



(c″)

(d″)

Fig. S3. Text mining in English. (*Continued*) (a″) Word patterns W_i in Charles Darwin's *Origin of Species* (including 15 chapters in the main text, but excluding the synopsis—the table of contents, the introduction, the glossary, and the index), sorted by descending $n_{ii} \geq 20$. (b″) Content word clusters W_i found in the synopsis of the aforementioned book, sorted by descending $n_{ii} \geq 3$. Font sizes are proportional to $\sqrt{n_{ii}}$. Temporal structures of the word patterns are ignored. (c″) False positives: automatically extracted topical patterns (colored *green* or *red* in panel a″) that are not mentioned anywhere in the synopsis. (d″) False negatives: automatically identified non-topic patterns (*gray* in panel a″) that are found in panel b″.

Example 4.15.3. We use exactly the same clustering algorithm to automatically generate words related to a particular question, in our WikiQA experiments. In Table S11 below, we list all the questions whose top-scoring answers (according to our algorithm) contain at least one sentence that has been officially labeled as correct by the WikiQA team. All the tabulated entries take the form of

[WikiQA-Q#]: [WikiQA question] [*Reference Wikipedia page*]
[Sentence(s) bearing the highest score in our algorithm.]

Here, while tabulating our results, we make no attempt to correct the spelling or capitalization in the question (see however, Algorithm 3.11), and the punctuation marks in the candidate answer(s) are kept in the same form as the original WikiQA dataset, such as

26: how did anne frank die *Anne Frank*
Anne Frank and her sister, Margot , were eventually transferred to the **Bergen-Belsen concentration camp** , where they **died** of **typhus** in March **1945**.

In the example above, the words related to the question (including the place of death "Bergen-Belsen concentration camp", cause of death "typhus" and year of death "1945", all discovered by our automated text mining) are highlighted in red.

Some (85 out of 580 total) question numbers in Table S11 are prefixed with ♯, because the corresponding questions are classified as "quantitative", by Algorithm 3.12. Intuitively speaking, one might answer such "quantitative" questions better by focusing on candidate sentences that mention numbers, dates and so on. However, in our current work, we have made no effort to treat "quantitative" questions with special screening procedures for the answers.

Table S11. List of WikiQA hits

0: HOW AFRICAN AMERICANS WERE IMMIGRATED TO THE US *African immigration to the United States*
As such, **African immigrants** are to be distinguished from **African American** people, the latter of whom are descendants of mostly West and **Central Africans** who were involuntarily brought to the United States by means of the historic Atlantic slave trade .

1: how are glacier caves formed? *Glacier cave*
A **glacier cave** is a **cave formed** within the ice of a **glacier** .

♯16: how much is 1 tablespoon of water *Tablespoon*
In the USA one **tablespoon** (measurement unit) is approximately 15 mL; the capacity of an actual **tablespoon** (dining utensil) ranges from 7 mL to 14 mL.
In countries where a **tablespoon** is a serving spoon, the nearest equivalent to the US **tablespoon** is either the dessert spoon or the soup spoon .
A **tablespoonful**, nominally the capacity of one **tablespoon**, is commonly used as a measure of volume in cooking .

18: how a rocket engine works *Rocket engine*
A **rocket engine**, or simply "**rocket**", is a **jet engine** that uses only stored propellant mass for **forming** its high speed propulsive **jet** .

26: how did anne frank die *Anne Frank*
Anne Frank and her sister, Margot , were eventually transferred to the **Bergen-Belsen concentration camp** , where they **died** of **typhus** in March **1945**.

♯31: how old was monica lewinsky during the affair *Monica Lewinsky*
Monica Samille **Lewinsky** (born July 23, 1973) is an American woman with whom United States President Bill Clinton admitted to having had an "improper relationship" while she worked at the White House in 1995 and 1996.

33: how are antibodies used in *antibody*
An **antibody** (Ab), also known as an immunoglobulin (Ig), is a large Y-shaped **protein produced** by B-cells that is **used** by the immune system to **identify** and neutralize foreign objects such as bacteria and viruses .

♯45: how old is kirk douglas, the actor? *Kirk Douglas*
Kirk Douglas (born Issur Danielovitch, ; December 9, 1916) is an American **stage** and film **actor**, film producer and author.

♯50: how long was richard nixon a president *Richard Nixon*
Richard Milhous **Nixon** (January 9, 1913 – **April** 22, **1994**) was the 37th **President** of the United States , **serving** from 1969 to 1974, when he became the only **president** to resign the office.

♯64: How long was Mickie James with WWE? *Mickie James*
James appeared in **World Wrestling** Entertainment (**WWE**) in October 2005 and was placed in a storyline with **Trish Stratus** , in which **James'** gimmick was that of **Stratus'** biggest fan turned obsessed stalker, an angle which lasted almost a year.

77: how are the # of electrons in each shell determined *Electron shell*
Each **shell** can **contain** only a fixed number of **electrons**: The 1st **shell** can hold up to two **electrons**, the 2nd **shell** can hold up to eight **electrons**, the 3rd **shell** can hold up to 18, and 4th **shell** can hold up to 32 and so on.

♯94: how old is beatrice author *Bea Arthur*
Beatrice "Bea" Arthur (May 13, 1922 – April 25, 2009) was an American actress, comedienne, and singer whose career spanned seven decades.

98: how are public schools funded *state school*
State schools (also **known** as **public schools** or **government schools**) generally **refer** to primary or secondary **schools** mandated for or offered to all children without charge paid for, in whole or in part, by taxation .

102: how does interlibrary loan work *Interlibrary loan*
Interlibrary loan (abbreviated ILL, and sometimes called interloan, document delivery, or document supply) is a service whereby a user of one library can borrow books or receive photocopies of documents that are owned by another library.
In many cases, nominal fees accompany **interlibrary loan** services.

104: what did mia hamm do his work *Mia Hamm*
Mariel Margaret "**Mia**" **Hamm** (born March 17, 1972) is a retired American professional soccer player.

105: what bacteria grow on macconkey agar *MacConkey agar*
MacConkey agar is a culture medium designed to **grow** Gram-negative **bacteria** and differentiate them for lactose fermentation .

110: how do forensic auditors examine financial reporting *Financial audit*
The purpose of an audit is provide and objective independent **examination** of the **financial statements**, which **increases** the value and credibility of the **financial statements** produced by management, thus **increase** user confidence in the **financial statement**, reduce investor **risk** and consequently reduce the cost of capital of the preparer of the **financial statements**.

127: What committees are joint committees *Joint committee*
A **Joint Committee** is a term in politics that is used to refer to a **committee** made up of members of both chambers of a bicameral legislature.

♯130: How many states and territories are within India? *States and territories of India*
India is a federal union of **states** comprising twenty-eight **states** and seven union **territories** .

140: how is single malt scotch made *Single malt Scotch*
Single Malt Scotch is **single malt** whisky **made** in Scotland using a pot still distillation process at a **single** distillery , with **malted** barley as the only grain ingredient.

164: what county in texas is conroe located in *Conroe, Texas*
Conroe is the seat of Montgomery **County** and falls within the metropolitan area.

178: how does a dredge work? *Dredging*
A **dredger** (or "dredge" as is the general usage in the Americas) is any device, **machine**, or vessel that is used to **excavate** and remove material from the bottom of a body of water.

♯181: how many world series did curt schilling have *Curt Schilling*
He helped lead the Philadelphia **Phillies** to the **World Series** in and **won World Series** championships in with the **Arizona Diamondbacks** and in and with the Boston Red Sox .

188: what area code is 479 *Area code 479*
Area code 479 serves Benton , Carroll (split with **area code** 870), Crawford , Franklin , Johnson , Logan , Madison , Polk , Pope , Scott , Sebastian , Washington and Yell Counties in Arkansas.
Area code 479 is the telephone **area code** serving thirteen counties in northwest Arkansas , including the cities of Fort Smith , Fayetteville , Springdale , Rogers and Bentonville : the second, third, fourth, eighth and tenth-largest cities in the state, respectively.

189: what does arraignment mean *Arraignment*
During the second **arraignment**, a post-indictment **arraignment** or PIA, the defendant is allowed to enter a plea.
The first is called the initial **arraignment** and must take place within 48 hours of an individual's arrest, 72 hours if the individual was arrested on the weekend and not able to go before a judge until Monday.
In response to **arraignment**, the accused is expected to enter a plea .
In federal courts of the United States, **arraignment** takes place in two stages.
In England, Wales, Northern Ireland and Australia, **arraignment** is the first of eleven stages in a criminal trial, and involves the clerk of the court reading out the indictment .
During this **arraignment** the defendant is informed of the pending legal charges and is informed of his or her right to retain counsel.
Arraignment is a formal reading of a criminal charging document in the presence of the defendant to inform the defendant of the charges against him or her.

190: what did chaucer do *Geoffrey Chaucer*
While he achieved fame during his lifetime as an author, philosopher, alchemist and astronomer , composing a scientific treatise on the astrolabe for his ten year-old son Lewis, **Chaucer** also maintained an active career in the civil service as a bureaucrat, courtier and diplomat.
Geoffrey **Chaucer** (; c. 1343 – 25 October 1400), known as the Father of English literature , is widely considered the greatest English poet of the Middle Ages and was the first poet to have been buried in Poet's Corner of Westminster Abbey .
Chaucer is a crucial figure in developing the legitimacy of the vernacular , Middle English , at a time when the dominant literary languages in England were French and Latin.

193: what does am and pm stand for *12-hour clock*
The 12-hour **clock** was developed over time from the mid-second millennium BC to the 16th century AD .
The 12-hour **clock** is a time convention in which the 24 hours of the day are divided into two periods: a.m. (from the Latin ante meridiem, meaning "before midday") and p.m. (post meridiem, "after midday").

203: how do pineapples grow *Pineapple*
The **Pineapple** (Ananas comosus) is a tropical plant with edible multiple fruit consisting of coalesced berries, and the most economically significant plant in the Bromeliaceae family.
Pineapples may be cultivated from a crown cutting of the fruit, possibly flowering in 20–24 months and fruiting in the following six months.
Pineapples are consumed fresh, cooked, juiced, and preserved, and are found in a wide array of cuisines .
Pineapple does not ripen significantly post-harvest.

221: what does righteousness mean *Righteousness*
Righteousness (also called rectitude) is an important theological concept in Christianity , Judaism , Islam , Hinduism (dharma), and Zoroastrianism .

241: what does the universal law of gravitation state *Newton's law of universal gravitation*
Newton's law of **universal gravitation states** that every point **mass** in the **universe** attracts every other point **mass** with a force that is directly proportional to the **product** of their **masses** and inversely proportional to the square of the distance between them.

242: how does a dim sum restaurant work *Dim sum*
Dim sum is also well known for the unique way it is served in some **restaurants**, wherein fully cooked and ready-to-serve **dim sum dishes** are carted around the **restaurant** for customers to choose their **orders** while seated at their tables.

♯253: how many planets is jupiter away from the sun? *Jupiter*
Jupiter is the fifth **planet** from the **Sun** and the largest **planet** in the Solar **System** .

254: what city was the convention when gerald ford was nominated *1976 Republican National Convention*
Held in Kemper Arena in Kansas **City** , Missouri , the **convention nominated** President **Gerald Ford** for a full term, but only after narrowly defeating a strong challenge from former California Governor Ronald Reagan .

261: what county is northville mi *Northville, Michigan*
Most of the **city** is in Oakland **County**, and is surrounded by the **city** of **Novi** .

289: what do porcupines eat *Porcupine*
It **eats** leaves, herbs, twigs and green plants like skunk cabbage and clover and in the winter it may **eat** bark.

290: what did elijah mccoy invent *Elijah McCoy*
Elijah J. **McCoy** (May 2, 1844 – October 10, 1929) was a black Canadian-American inventor and engineer, who was notable for his 57 U.S. patents , most to do with lubrication of steam engines.

299: What does the idea of effective nuclear charge tell us *Effective nuclear charge*
The term "**effective**" is used because the shielding **effect** of negatively **charged** electrons prevents higher orbital electrons from experiencing the full **nuclear charge** by the repelling **effect** of inner-layer electrons.

303: what domain is eu *.eu*
.eu is the country code top-level **domain** (ccTLD) for the European Union (EU).

304: What does the name "Meister" mean in German *Meister*
Meister means master in **German** (as in master craftsman , or as an honorific title such as **Meister** Eckhart).

309: what did tycho brahe do *Tycho Brahe*
Tycho Brahe (14 December 154624 October 1601), born Tyge Ottesen **Brahe**, was a **Danish** nobleman known for his accurate and comprehensive astronomical and planetary observations.

317: what holiday is first weekend of september *Labor day*
Labor Day is an American federal **holiday** observed on the **first** Monday in **September**, that celebrates the economic and social contributions of workers.

320: what does life insurance cover? *Life insurance*
Life insurance is a contract between an **insured** (**insurance** policy holder) and an **insurer** , where the **insurer** promises to pay a designated **beneficiary** a sum of money (the "**benefits**") upon the **death** of the **insured person**.

327: how is a backpack made *Backpack*
The bag itself was **made** up of different **animal** hide and skin (depending on what sorts of animals were in the area) and sewn together by the intestines of said animals, which were woven together tightly to **make** a sturdy thread-like material.

331: what day is st. patricks day *Saint Patrick's Day*
Saint **Patrick's Day** or the Feast of Saint **Patrick** (, "the **Day** of the Festival of **Patrick**") is a cultural and religious holiday celebrated on 17 March.

334: what cheese is made from goat's milk *Goat cheese*
Goat cheese, or chèvre (from the French word for goat), is **cheese made** out of the **milk** of **goats** .

338: what country is belize in *Belize*
Belize , is a **country** located on the northeastern **coast** of Central **America**.

342: how does a cat purr *Purr*
However, using a strict definition of **purring** that continuous sound production must alternate between pulmonic egressive and ingressive airstream (and usually go on for minutes), Peters (2002), in an exhaustive review of the scientific literature, reached the conclusion that until then only '**purring cats**' (Felidae) and two species of genets , Genetta tigrina, and most likely also Genetta genetta, had been documented to **purr**.

346: what county is St. Elizabeth MO in *St. Elizabeth, Missouri*
St. Elizabeth is a village in Miller **County** , Missouri , United States .

348: what county is wilton ca in *Wilton, California*
Wilton is a census-designated place (CDP) in Sacramento **County** , California , United States .

#354: how many gold gloves does barry larkin have *Barry Larkin*
Larkin is considered one of the top players of his era, winning nine Silver Slugger **awards** and three **Gold Glove awards** .

361: how did women's role change during the war *Women's roles in the World Wars*
Whether it was on the home front or the front-lines, for civilian or enlisted **women**, the World Wars started a new era for **women's** opportunities to contribute in **war** and be recognized for efforts outside of the home.

366: how does lsd impact the human body *Lysergic acid diethylamide*
Lysergic acid diethylamide, abbreviated **LSD** or LSD-25, also known as lysergide (INN) and colloquially as acid, is a semisynthetic **psychedelic** drug of the ergoline family, well known for its psychological effects which can include altered thinking processes, closed and open eye visuals, synesthesia , an altered sense of time and spiritual experiences , as well as for its key role in 1960s counterculture .
Currently, a number of organizations—including the Beckley Foundation , MAPS , Heffter Research Institute and the Albert Hofmann Foundation—exist to fund, encourage and coordinate research into the medicinal and spiritual uses of **LSD** and related **psychedelics**.

374: what county is orono maine in *Orono, Maine*
Orono is a town in Penobscot **County** , **Maine** , United States .

377: what it is a pilot study *Pilot experiment*
A **pilot** experiment, also called a **pilot study**, is a small scale preliminary **study** conducted in order to evaluate feasibility, time, cost, adverse events, and effect size (statistical variability) in an attempt to predict an appropriate sample size and improve upon the **study** design prior to performance of a full-scale research project.

#383: how many humps on a camel *Camel*
The two surviving species of **camel** are the dromedary , or one-humped **camel**, which is native to the Middle East and the Horn of Africa ; and the Bactrian , or two-humped **camel**, which **inhabits** Central Asia .

384: what can be powered by wind *Wind power*
Wind power is the conversion of **wind** energy into a useful form of energy, such as using **wind** turbines to make electrical **power** , windmills for mechanical **power**, **wind** pumps for water pumping or drainage , or sails to propel ships.

389: what chili wants wiki *What Chilli Wants*
What Chilli **Wants** is an American reality series on VH1 starring Chilli , one-third of the Grammy Award -winning R&B trio TLC .

398: what division is boise state football *Boise State Broncos football*
The **Boise State** Broncos **football** program represents **Boise State University** in college **football** and compete in the **Football** Bowl Subdivision (FBS) of **Division** I as a member of the Mountain West Conference .

#409: how much caffeine is in a shot of espresso *Espresso*
Espresso has more **caffeine** per unit **volume** than most beverages, but the usual serving **size** is smaller—a typical **60 mL** (2 US fluid ounce) of **espresso** has 80 to 150 mg of **caffeine**, **little** less than the 95 to 200 mg of a **standard** 240 **mL** (8 US fluid ounces) **cup** of drip-brewed coffee.

417: what does add my two cents mean *My two cents*
"My **two cents**" (2¢) and its longer version "put my **two cents** in" is an United States (US) idiomatic expression, taken from the original English idiom expression: to put in "my **two** pennies worth" or "my tuppence worth."

447: what does bruce jenner do *Bruce Jenner*
William **Bruce Jenner** (born October 28, 1949) is a former U.S. track and field athlete , motivational speaker , socialite , television personality and businessman .

448: what does alkali do to liquids? *Alkali*
In chemistry , an **alkali** (; from Arabic : al-qaly ,القالي القَالي) is a basic , ionic salt of an **alkali** metal or alkaline earth metal element .
This broad use of the term is likely to have come about because **alkalis** were the first bases known to obey the Arrhenius definition of a base and are still among the more common bases.
Some authors also define an **alkali** as a base that dissolves in water .

450: what area code is 217 *Area code 217*
Area code 217 is the North American telephone **area code** for much of western and central Illinois .

462: what does gloria in excelsis deo mean *Gloria in Excelsis Deo*
"**Gloria** in **excelsis Deo**" (Latin for "**Glory** to God in the highest") is a hymn known also as the Greater Doxology (as distinguished from the "Minor Doxology" or **Gloria** Patri) and the Angelic Hymn.

#463: how many grams in a troy ounce of gold *Troy weight*
The **troy ounce** is 480 **grains**, compared with the avoirdupois **ounce**, which is 437½ **grains**.

#473: how much does united states spend on health care *Health care in the United States*
According to the **World Health** Organization (WHO), the **United States spent** more on **health care** per **capita** ($7,146), and more on **health care** as percentage of its GDP (15.2%), than any other **nation** in 2008.

#488: how many died in hiroshima and nagasaki *Atomic bombings of Hiroshima and Nagasaki*
Within the first two to four months of the **bombings**, the acute **effects** killed 90,000–166,000 people in **Hiroshima** and 60,000–80,000 in **Nagasaki**, with roughly half of the **deaths** in each **city** occurring on the first day.

490: what county is Holly Ridge nc in? *Holly Ridge, North Carolina*
Holly Ridge is a town in Onslow **County** , North Carolina , United States .

503: what kind of books does debbie macomber writes *Debbie Macomber*
Debbie Macomber (born October 22, 1948 in Yakima, Washington) is a best-selling American author of over 150 romance novels and contemporary women's fiction.

504: what county is catonsville md in *Catonsville, Maryland*
Catonsville is the home of the University of Maryland, Baltimore **County** (UMBC), a public research university with over 12,000 students.
Catonsville is a census-designated place (CDP) in Baltimore **County** , Maryland , United States.

509: what kind of literature did john steinbeck writing *John Steinbeck*
As the author of twenty-seven **books**, including sixteen novels, six non-fiction **books**, and five collections of short stories, **Steinbeck** received the **Nobel Prize** for **Literature** in 1962.

510: whatever happened clint walker *Clint Walker*
Norman Eugene **Walker**, known as **Clint Walker** (born May 30, 1927), is a retired American actor .

512: what does the family leave act *Family and Medical Leave Act of 1993*
The **Family** and **Medical Leave Act** of 1993 (FMLA) is a United States federal law requiring covered employers to provide employees job-protected and unpaid **leave** for qualified **medical** and **family reasons**.

515: How is the pothole formed *Pothole*
A **pothole** (sometimes called a kettle and known in parts of the Western United States as a chuckhole) is a type of disruption in the surface of a roadway where a portion of the road material has broken away, leaving a hole.
A deep **pothole** on New York City's Second Avenue

517: what does estee lauder do *Estée Lauder Companies*
Estée Lauder Companies, Inc. is a manufacturer and marketer of prestige skincare, makeup, fragrance and hair care products.

534: what creates sonic boom *Sonic boom*
A **sonic boom** is the **sound** associated with the shock waves **created** by an object traveling through the air faster than the **speed** of **sound**.

551: what city is oregon state university in *Oregon State University*
Oregon State University (OSU) is a coeducational , public **research university** located in Corvallis , **Oregon** , United States .

#561: how many bones are in the skeletal system is composed of 306 bones *Human skeleton*
Humans are born with over 270 **bones**, some of which fuse together into a longitudinal axis, the axial **skeleton** , to which the appendicular **skeleton** is attached.

564: What county is NOrth Myrtle Beach in SC *North Myrtle Beach, South Carolina*
North Myrtle Beach is a coastal resort city in Horry **County** , **South** Carolina , United States .

#568: how much of our universe does plasma make up *Plasma (physics)*
In the **universe** , **plasma** is the most common state of **matter** for ordinary **matter** , most of which is in the rarefied intergalactic **plasma** (particularly intracluster medium) and in stars.

571: what does barefoot and pregnant mean *Barefoot and pregnant*
The phrase "**barefoot** and **pregnant**" was probably first used sometime before 1950.
The only way to keep a woman happy,' he said, 'is to keep her **barefoot** and pregnant.'"
"**Barefoot** and **pregnant**" is a phrase most commonly associated with the controversial idea that women should not work outside the home and should have many children during their reproductive years.

572: what county is oakhurst, nj in *Oakhurst, New Jersey*
Oakhurst is a census-designated place and unincorporated community within Ocean Township , in Monmouth **County** , New **Jersey** , United States.

574: What did Lawrence Joshua Chamberlain do? *Joshua Chamberlain*
Joshua Lawrence Chamberlain (September 8, 1828 – February 24, 1914), born as **Lawrence Joshua Chamberlain**, was an American **college** professor from the State of Maine , who volunteered during the American Civil War to join the Union Army .

575: what circuit court is maryland *Maryland Circuit Courts*
The **Circuit Courts** of **Maryland** are the state trial **courts** of general jurisdiction in **Maryland**.

577: what causes a deficiency in adenosine deaminase *Adenosine deaminase deficiency*
Adenosine deaminase deficiency, also called ADA **deficiency** or ADA-SCID, is an autosomal recessive metabolic disorder that **causes** immunodeficiency .

582: how is human height measured *Human height*
Human height is the distance from the bottom of the feet to the top of the head in a **human** body , standing erect.

587: what does a timing belt do *Timing belt*
Timing belt (camshaft) , a toothed **belt** used to drive the camshaft(s) within an internal combustion engine
A **timing belt** is a non-slipping mechanical drive **belt** and the term may refer to either:

604: what channel is shopnbc on *ShopNBC*
ShopNBC is an American broadcast and cable home shopping network, owned and operated by ValueVision Media , which is in turn 30% owned by GE Equity and NBC Universal .
ShopNBC headquarters is in Eden Prairie, Minnesota .

606: how post and lintels are used *Post and lintel*
Post and **lintel**, "prop and **lintel**" or "trabeated" is a simple construction method using a **lintel** , header, or architrave as the horizontal member over a building void supported at its ends by two vertical columns , , or .

616: what county is Cambria WI in *Cambria, Wisconsin*
Cambria is a village in Columbia **County** , Wisconsin , United States .

#622: HOW MANY STRIPES ARE ON THE AMERICAN FLAG *Flag of the United States*
The national **flag** of the United States of **America** , often simply referred to as the **American flag**, consists of **thirteen** equal horizontal **stripes** of **red** (**top** and **bottom**) **alternating** with **white**, with a **blue** rectangle in the canton (referred to specifically as the "union") bearing fifty small, **white**, **five-pointed stars** arranged in nine offset horizontal rows of **six stars** (**top** and **bottom**) **alternating** with rows of **five stars**.

625: what do biologists do *Biologist*
Biologists involved in basic research attempt to discover underlying mechanisms that govern how organisms work.
Biologists involved in applied research attempt to develop or improve medical, industrial or agricultural processes.
A **biologist** is a scientist who studies living organisms and their relationship to their environment.

626: what does base jumping stand for *BASE jumping*
BASE jumping, also sometimes written as B.A.S.E. **jumping**, is an activity where participants **jump** from **fixed** objects and use a **parachute** to break their fall.

631: What country is the largest stalagmite *Stalagmite*
The **largest stalagmite** in the world is high and is located in the cave of Cueva Martin Infierno, Cuba.

633: what does Gringo mean *Gringo*
Roger Axtell, a travel etiquette expert, notes that "[t]he word **gringo** is not necessarily a bad word.
Gringo (, ,) is a slang Spanish and Portuguese word used in Ibero-America , to denote foreigners, often from the United States .

638: what does hair testing show *Drug test*
A drug **test** is a technical analysis of a biological specimen – for example urine, **hair**, blood, sweat, or oral fluid / saliva – to determine the presence or absence of specified parent drugs or their metabolites .

644: what hormones produce thyroid *Thyroid hormone*
The **thyroid hormones**, triiodothyronine (T3) and thyroxine (T4), are tyrosine -based **hormones produced** by the **thyroid gland** that are primarily responsible for regulation of metabolism.

645: what does automatic paper feeder on printers mean *Automatic Document Feeder*
In multifunction or all-in-one **printers** , fax machines , photocopiers and scanners , an **automatic** document **feeder** or ADF is a feature which takes several pages and **feeds** the **paper** one page at a time into a scanner or copier, allowing the user to scan , and thereby copy , **print** , or fax , multiple-page documents without having to manually replace each page.

#660: how old a child pedophile crime *Pedophilia*
As a medical diagnosis, **pedophilia** or paedophilia is a psychiatric **disorder** in persons 16 years of **age** or **older** typically characterized by a primary or exclusive sexual interest toward **prepubescent children** (generally **age** 11 years or **younger**, though specific diagnosis criteria for the **disorder** extends the cut-off point for prepubescence to **age** 13).

#662: how many people were killed in the holocaust *The Holocaust*
Over one **million** Jewish children were **killed** in the **Holocaust**, as were **approximately** two **million** Jewish women and three **million** Jewish **men**.

679: what channel is letterman on *Late Show with David Letterman*
Late Show with David **Letterman** is an American late-night talk show **hosted** by David **Letterman** on CBS .

683: what does karma mean in buddhism *Karma in Buddhism*
Karma (Sanskrit , also **karman**, Pāli : Kamma) **means** "action" or "doing"; whatever one does, says, or thinks is a **karma**.

687: what kind a is tilapia *Tilapia*
Tilapia ( ) is the common name for nearly a hundred species of cichlid fish from the **tilapiine** cichlid tribe .

#695: how many grape farms in united states *Agriculture in the United States*
As of the last census of agriculture in 2007, there were 2.2 **million farms**, covering an area of , an average of per **farm**.

696: what freezes faster? hot or cold water? *Mpemba effect*
The Mpemba effect, named after Tanzanian student Erasto Mpemba , is the assertion that warmer **water** can **freeze faster** than **colder water**.

698: what causes a small bowel obstruction *Bowel obstruction*
Bowel obstruction (or intestinal **obstruction**) is a mechanical or functional **obstruction** of the intestines, preventing the normal transit of the products of digestion.

#699: how many qfc stores are there *QFC*
Quality Food Centers (**QFC**) is a supermarket chain based in Bellevue, Washington , with 64 **stores** in the Puget Sound region of Washington state and in the Portland, Oregon metropolitan area.

705: What Causes Brain Freeze *Ice-cream headache*
It is **caused** by having something cold touch the roof of the mouth (palate), and is believed to result from a nerve response **causing** rapid constriction and swelling of blood vessels or a " referring " of pain from the roof of the mouth to the head.

708: what does leeroy jenkins mean *Leeroy Jenkins*
Leeroy Jenkins, sometimes misspelled **Leroy Jenkins** and often elongated with numerous additional letters, is an Internet meme named for a player character created by Ben Schulz in Blizzard Entertainment 's MMORPG , World of Warcraft .

#722: how many episodes of Lost were there *List of Lost episodes*
A total of 121 **episodes** of **Lost** were produced, the last of which aired on May 23, 2010.
ABC announced that **Lost** would end after six seasons, having produced a total of 121 **episodes**.

#742: how much does a gold bar weigh *Gold bar*
The standard **gold bar** held as **gold** reserves by central banks and traded among bullion dealers is the 400-troy-ounce (12.4 kg or 438.9 ounces) Good Delivery **gold bar**.

743: what county is bolingbrook il in? *Bolingbrook, Illinois*
Bolingbrook is a large village in Will and DuPage **Counties** in the U.S. state of Illinois .

#745: how many kids does archie manning have *Archie Manning*
He is the father of current Denver Broncos quarterback Peyton **Manning** , current New York Giants **starting** quarterback Eli **Manning** , and former Ole Miss receiver **Cooper Manning** .

746: what does an advocacy website promote? *Advocacy*
Research is beginning to explore how **advocacy** groups in the U.S. and Canada are using social media to facilitate civic engagement and collective action.
Lobbying (often by lobby groups) is a form of **advocacy** where a direct approach is made to legislators on an issue which plays a significant role in modern politics.
Advocacy is a political process by an individual or group which aims to influence public-policy and resource allocation decisions within political, economic, and social systems and institutions.
Advocacy can include many activities that a person or organization undertakes including media campaigns, public speaking, commissioning and publishing research or polls or the filing of an amicus brief .

#750: how many amendments in us *United States Constitution*
The Constitution has been **amended** seventeen **additional** times (for a total of **twenty-seven** amendments).

751: what city is george washington university *George Washington University*
The **George Washington University** (GW, **GWU**, or **George Washington**) is a comprehensive private , coeducational research **university** located in **Washington**, D.C. .

#755: How many Muslims live in the United Kingdom? *Islam in the United Kingdom*
The vast majority of **Muslims** in the **United Kingdom live** in England and Wales : of 1,591,000 **Muslims** recorded at the 2001 Census, 1,536,015 were **living** in England and Wales , where they formed 3% of the **population** in 2001; 42,557 were **living** in Scotland , forming 0.84% of the **population**; and 1,943 were **living** in Northern Ireland .

#763: how long to take iwo jima *Battle of Iwo Jima*
The Battle of **Iwo Jima** (19 February – 26 March 1945), or **Operation** Detachment, was a major battle in which the **United States** Armed Forces fought for and captured the island of **Iwo Jima** from the Japanese Empire .

764: What a Margarita contains *Margarita*
The **margarita** is a Mexican cocktail consisting of tequila mixed with Cointreau or similar orange -flavoured liqueur and lime or lemon juice , often served with salt on the glass rim.
The drink is served shaken with ice (on the rocks), blended with ice (frozen margarita), or without ice (straight up).

766: what color is burgundy *Burgundy (color)*
Burgundy is a dark red **color** associated with the **Burgundy** wine of the same name, which in turn is named after the **Burgundy** region of France .

773: what day is the feast of st joseph's? *Saint Joseph's Day*
Saint **Joseph's Day**, March 19, the **Feast** of **St**. **Joseph** is in Western Christianity the principal **feast day** of Saint **Joseph** , Spouse of the Blessed Virgin Mary .

791: what does a laboratory in a gynecologist office consist of *Medical laboratory*
A **medical laboratory** or clinical **laboratory** is a **laboratory** where tests are done on clinical specimens in order to get information about the health of a patient as pertaining to the diagnosis, treatment, and prevention of disease.

812: what countries allow gays to openly serve in the military *Sexual orientation and military service*
Nations that permit **gay** people to **serve openly** in the **military** include the 4 of the 5 members of the UN Security Council (United States, United Kingdom, France, and Russia), the Republic of China (Taiwan), Australia , **Israel** , **South Africa** , Argentina , and all NATO members excluding Turkey .

813: How Works Diaphragm Pump *Diaphragm pump*
A **diaphragm pump** (also known as a Membrane **pump**, Air Operated Double **Diaphragm Pump** (AODD) or Pneumatic **Diaphragm Pump**) is a positive displacement **pump** that uses a combination of the reciprocating action of a rubber , thermoplastic or teflon **diaphragm** and suitable valves either side of the **diaphragm** (check valve , butterfly valves, flap valves, or any other form of shut-off valves) to **pump** a fluid .

#815: how many percent is a basis point *Basis point*
The relationship between percentage changes and **basis points** can be summarized as follows: 1 percentage **point** change = 100 **basis points**, and 0.01 percentage **points** = 1 **basis point**.

819: what genre is bloody beetroots *The Bloody Beetroots*
The **Bloody Beetroots** is well-known for the black Venom mask he wears during performances.
The **Bloody Beetroots** is the pseudonym of Sir Bob Cornelius Rifo, the Italian electro house and dance-punk music producer , DJ and photographer .
"The **Bloody Beetroots** DJ set" contains Sir Bob Cornelius Rifo and Tommy Tea.

#838: how many asian indians live in usa *Indian American*
Indian Americans are citizens of the United States of **Indian** ancestry and comprise about 3.18 million **people**, or ~1.0% of the U.S. **population**, the **country's** third largest self-reported **Asian** ancestral **group** after Chinese **Americans** and Filipino **Americans** according to **American** Community Survey of 2010 data.

843: what caused the world war 2 *Causes of World War II*
The main **causes** of **World War II** were nationalistic issues, unresolved issues, and resentments resulting from **World War** I and the interwar period in Europe, in addition to the effects of the Great Depression in the 1930s.

867: how is root beer made? *Root beer*
Root beer is a carbonated , sweetened beverage , originally **made** using the **root** of a sassafras plant (or the bark of a sassafras tree) as the primary flavor.

872: what does informal logic mean *Informal logic*
Informal logic, intuitively, refers to the principles of **logic** and **logical** thought outside of a formal setting.

874: what country is turkey in *Turkey*
Turkey (), officially the Republic of **Turkey** , is a transcontinental **country** , **located** mostly on Anatolia in Western Asia and on East Thrace in Southeastern Europe .

876: what county is Augusta,GA located in? *Augusta, Georgia*
Augusta is the principal city of the **Augusta** – Richmond **County** Metropolitan Statistical Area , which as of 2010 had an estimated population of 556,877, making it both the second-largest city and the second-largest metro area in the **state** after Atlanta .

893: what does a cutter do *Cutter (baseball)*
In baseball , a **cutter**, or **cut** fastball, is a type of fastball which breaks slightly toward the pitcher 's glove side as it reaches home plate .
When a batter is able to hit a **cutter** pitch, it often results in soft contact and an easy out, due to the pitch's movement keeping the ball away from the bat's sweet spot .
The **cutter** is typically 2–5 mph slower than a pitcher's four-seam fastball .
Some pitchers use a **cutter** as a way to prevent hitters from expecting their regular fastballs.
In 2010, the average pitch classified as a **cutter** by PITCHf/x thrown by a right-handed pitcher was 88.6 mph; the average four-seamer was 92.1 mph.
An animated diagram of a **cutter**
A common technique used to throw a **cutter** is to use a four-seam fastball grip with the baseball set slightly off center in the hand.

895: what it takes aerosmith album *What It Takes (song)*
"What It **Takes**" is a power ballad by American hard rock band **Aerosmith** .

898: what county is erie colorado *Erie, Colorado*
Erie is a Statutory Town in Boulder and Weld **counties** in the U.S. **state** of **Colorado** .

902: what classes are considered humanities *Humanities*
The **humanities** that are also regarded as social **sciences** include history , anthropology , area **studies** , communication **studies** , cultural **studies** , law , economics and linguistics .

910: what food is in afghan *Afghan cuisine*
Accompanying these staples are dairy products (yogurt and whey), various **nuts** , and native vegetables, as well as fresh and dried **fruits**; **Afghanistan** is well known for its grapes .

922: what kind of cut is tri tip *Tri-tip*
The **tri-tip** is a **cut** of beef from the bottom sirloin primal **cut** .

923: what county is willmar mn in? *Willmar, Minnesota*
Willmar is a city in, and the **county** seat of, Kandiyohi **County** , Minnesota , United States .

#943: how many innings makes an official game *Official game*
Since most professional baseball **games** are nine **innings** long, the fifth **inning** is used as the threshold for an **official game**.

963: how tennessee became a state *Tennessee*
Tennessee was the last **state** to leave the **Union** and join the **Confederacy** at the outbreak of the U.S. Civil War in 1861, and the first **state** to be readmitted to the **Union** at the end of the war.

967: what county is coatesville indiana located in *Coatesville, Indiana*
Coatesville is a town in Clay Township , Hendricks **County** , **Indiana** , United States .

968: how does Delaware support its claim to being the first state? *Delaware*
Delaware was one of the 13 **colonies** participating in the **American** Revolution and on December 7, 1787, became the **first state** to ratify the Constitution of the United **States** , thereby becoming known as The **First State**.

973: what does fidelity do *Fidelity Investments*
FMR LLC (**Fidelity** Management and Research) or **Fidelity** Investments is an American multinational financial services corporation.
Fidelity Ventures is its venture capital arm.
Fidelity Investments manages a large family of mutual funds , provides fund distribution and investment advice services, as well as providing discount brokerage services , retirement services, wealth management , securities execution and clearance, life insurance and a number of other services.
Fidelity International Limited (FIL), was an international affiliate founded in 1969, serving most countries in the rest of the world.

976: what causes thunder sound *Thunder*
In turn, this expansion of air creates a sonic shock wave, similar to a sonic boom , which produces the **sound** of **thunder**, often referred to as a clap, crack, or peal of **thunder**.

990: what does continuous improvement mean *Continual improvement process*
A **continual improvement** process, also often called a **continuous improvement** process (abbreviated as CIP or CI), is an ongoing effort to **improve products**, services, or processes.

993: what area code is 810 *Area code 810*
Area code 810 is the telephone **area code** serving eastern Michigan .

996: what kind of earthquake was haiti *2010 Haiti earthquake*
The **2010 Haiti earthquake** was a catastrophic magnitude 7.0 Mw **earthquake** , with an epicenter near the town of Léogâne (Ouest Department), **approximately** west of Port-au-Prince , **Haiti's** capital.

997: what kind of horse was mr ed *Mister Ed*
The stars of the show were Mister **Ed**, a palomino **horse** who could "talk", played by gelding Bamboo Harvester and voiced by former Western star Allan Lane (who went uncredited for the entire length of the series), and his owner, an eccentric and enormously klutzy , yet friendly, architect **named** Wilbur Post (Alan Young).

999: who does chris humphries play for *Kris Humphries*
Kris Nathan **Humphries** (born February 6, 1985) is an American professional basketball **player** with the Brooklyn Nets .

#1010: what year was the 8 track invented *8-track tape*
Stereo **8** was created in 1964 by a consortium led by Bill Lear of Lear Jet Corporation, along with Ampex , Ford Motor **Company** , General Motors , Motorola , and RCA Victor Records (RCA).

1012: what are points on a mortgage *Point (mortgage)*
The longer you keep the property financed under the loan with purchased **points**, the more the money spent on the **points** will pay off.
Selling the property or refinancing prior to this break-even **point** will result in a net financial loss for the buyer while keeping the loan for longer than this break-even **point** will result in a net financial savings for the buyer.
Points, sometimes also called a "discount **point**", are a form of pre-paid interest .

1015: what are the official languages of nigeria *Languages of Nigeria*
The **official language** of **Nigeria**, English , the former colonial **language**, was chosen to facilitate the cultural and linguistic unity of the country.

1020: who played guitar on the kiss album, creatures of the night *Creatures of the Night*
It is also the band's last **album recorded** with Ace Frehley credited as an official member (until 1998's Psycho Circus), and its first **album** with Vinnie Vincent as the initially uncredited lead **guitarist** (Vincent would later be credited, but not featured pictorially on the cover, of 1985's reissue of the **album**).

1025: what are the different types of cross country skiing *Cross-country skiing*
Cross-country **skiing** is part of the Nordic **skiing** sport family, which includes **ski** jumping , Nordic combined (cross-country **skiing** and **ski** jumping), Biathlon (**skiing** and rifle marksmanship) and ski-orienteering (which includes map navigation along snow trails and tracks).

1026: what are use taxes? *Use tax*
A **use tax** in United States is complementary to sales **tax** which means, if you buy a **Taxable** product/services without paying any Sales **tax** to vendor, you owe **use tax**.

1030: what is an array in programming *Array programming*
In computer science , **array programming languages** (also known as vector or multidimensional **languages**) generalize operations on scalars to apply transparently to vectors , matrices , and higher dimensional **arrays**.

1032: who wrote a rose is a rose is a rose *Rose is a rose is a rose is a rose*
The sentence "**Rose** is a **rose** is a **rose** is a rose." was **written** by Gertrude Stein as part of the 1913 poem Sacred Emily, which appeared in the 1922 book Geography and Plays.

1043: what state is jakarta in? *Jakarta*
Jakarta (), officially known as the Special **Capital** Region of **Jakarta** (), is the **capital** and largest **city** of **Indonesia** .

1046: what song is sung in 6 foot 7 foot *6 Foot 7 Foot*
"**6 Foot 7 Foot**" (also styled as "6'7") is a **song** by American rapper Lil Wayne featuring label mate, Cory Gunz .

1052: who can file suspicious activity report *Suspicious activity report*
In United States financial regulation , a **suspicious activity report** (or SAR) is a **report** made by a financial institution to the Financial Crimes Enforcement Network (FinCEN), an **agency** of the United States Department of the Treasury , regarding **suspicious** or potentially **suspicious activity**.

1054: what are the sulfas *Sulfonamide (medicine)*
The original antibacterial sulfonamides (sometimes called **sulfa** drugs or sulpha drugs) are synthetic antimicrobial agents that contain the sulfonamide **group**.
It is important to make a distinction between **sulfa** drugs and other sulfur-containing drugs and additives, such as sulfates and sulfites, which are chemically unrelated to the sulfonamide **group**, and do not cause the same hypersensitivity reactions seen in the sulfonamides.

1060: who owns joe's crab shack *Joe's Crab Shack*
Joe's Crab Shack is an American chain of beach-themed seafood casual dining restaurants **owned** by Ignite Restaurant Group .

1063: what is a right wing christian *Christian right*
Christian right is a term used in the United **States** to describe right-wing **Christian political** groups that are characterized by their strong support of socially conservative policies.

1072: what is an assist in basketball *Assist (basketball)*
In **basketball** , an **assist** is attributed to a player who passes the ball to a teammate in a way that leads to a score by field goal , meaning that he or she was "**assisting**" in the **basket**.

1088: what school district is alvin in? *Alvin Independent School District*
Alvin Independent School District is a **school district** based in **Alvin** , Texas (USA).

1091: what are the houses of hogwarts *Hogwarts*
Hogwarts School of Witchcraft and Wizardry, shortened **Hogwarts**, is a fictional British **school** of **magic** for **students** aged eleven to eighteen, and is the primary setting for the first six books in J.K. Rowling 's Harry Potter series.

1094: What kind of personality test is the MBTI *Myers-Briggs Type Indicator*
The Myers-Briggs Type **Indicator** (**MBTI**) **assessment** is a psychometric **questionnaire** designed to measure **psychological** preferences in how people perceive the world and make decisions.

1100: where does a flea live *Flea*
Fleas are external parasites , **living** by hematophagy off the blood of mammals and birds .

1108: who wrote rhythm of the ancient mariner *The Rime of the Ancient Mariner*
The **Rime** of the **Ancient Mariner** (originally The **Rime** of the Ancyent Marinere) is the longest major poem by the English poet Samuel Taylor Coleridge , **written** in 1797–98 and published in 1798 in the first **edition** of Lyrical Ballads .

#1109: when did freddie mercury die *Freddie Mercury*
Freddie Mercury (born Farrokh Bulsara; Gujarati : ફરોખ બલ્સારા, Pharōkh Balsārā); 5 September 1946 – **24 November 1991**) was a British musician, singer and songwriter, best known as the lead vocalist and lyricist of the **rock** band Queen .

1111: what is a .PAC file *Proxy auto-config*
The URL of the **PAC file** is either configured manually or determined automatically by the Web Proxy Autodiscovery Protocol .
The browser fetches this **PAC file** before requesting other URLs.
A proxy auto-config (**PAC**) **file** defines how web browsers and other user agents can automatically choose the appropriate proxy server (access method) for fetching a given URL .
A **PAC file** contains a JavaScript function "FindProxyForURL(url, host)".

1117: what are spoilers for *Spoiler (automotive)*
Spoilers on the front of a vehicle are often called air dams, because in addition to directing air flow they also reduce the amount of air flowing underneath the vehicle which generally reduces aerodynamic lift and drag.
Spoilers for cars are often incorrectly confused with, or the term used interchangeably with, wings.
Spoilers are often fitted to race and high-performance sports cars , although they have become common on passenger vehicles as well.
Some **spoilers** are added to cars primarily for styling purposes and have either little aerodynamic benefit or even make the aerodynamics worse.
A **spoiler** is an automotive aerodynamic device whose intended design function is to '**spoil**' unfavorable air movement across a body of a vehicle in motion, usually described as turbulence or drag.

1120: who sings i am a man of constant sorrow — *Man of Constant Sorrow*
"**Man** of **Constant Sorrow**" (also known as "I Am A **Man** of **Constant Sorrow**") is a traditional American folk song first recorded by Dick Burnett , a partially blind fiddler from Kentucky .

1123: who kill franz ferdinand ww1 — *Assassination of Archduke Franz Ferdinand of Austria*
On 28 June 1914, Archduke **Franz Ferdinand** of Austria , heir presumptive to the Austro-Hungarian throne, and his wife, Sophie, Duchess of Hohenberg , were shot **dead** in Sarajevo , by Gavrilo Princip , one of a group of six Bosnian Serb assassins coordinated by Danilo Ilić .

1140: what is a mms message — *Multimedia Messaging Service*
Multimedia **Messaging** Service (**MMS**) is a standard way to send **messages** that include multimedia content to and from mobile phones .

1150: who sang proud mary — *Proud Mary*
"**Proud Mary**" is a rock song written by American singer-songwriter and multi-instrumentalist John Fogerty , and **recorded** by his band Creedence Clearwater Revival .

1154: where do crocodiles live — *Crocodile*
Crocodiles (**subfamily Crocodylinae**) or **true crocodiles** are **large** aquatic tetrapods that **live** throughout the tropics in Africa , Asia , the Americas and Australia .

1157: what relates to erosion — *Erosion*
Water and wind **erosion** are now the two primary causes of land degradation ; combined, they are responsible for 84% of degraded acreage, making excessive **erosion** one of the most significant global **environmental** problems.

1159: where in oregon is albany — *Albany, Oregon*
Albany is the 11th largest city in the U.S. state of **Oregon** , and is the **county** seat of **Linn County** .

#1168: when did abraham lincoln write the emancipation proclamation?? — *Emancipation Proclamation*
The **Emancipation Proclamation** was an **order** issued to all segments of the **Executive** branch (including the Army and Navy) of the United States by President **Abraham Lincoln** on January 1, 1863, during the American Civil War .

1172: what part of the pre-world war 1 arms race was the most intense? — *Causes of World War I*
The **causes** of World **War** I , which **began** in central Europe in late **July** 1914 and finished in 1918, included many factors, such as the conflicts and hostility of the four decades leading up to the **war**.

#1182: when president nixon resigns — *Richard Nixon*
Richard Milhous **Nixon** (January 9, 1913 – April 22, **1994**) was the 37th **President** of the United States , **serving** from 1969 to **1974**, when he became the only **president** to **resign** the office.

1184: where fourth of july came from — *Independence Day (United States)*
Independence Day, commonly known as the **Fourth** of **July**, is a federal **holiday** in the United **States** commemorating the adoption of the Declaration of Independence on **July** 4, 1776, declaring independence from the Kingdom of Great Britain .

1194: what people used mayan numeral system — *Maya numerals*
Maya **numerals** are a vigesimal (base - twenty) **numeral system used** by the Pre-Columbian Maya civilization .

1196: what is a newsgroup message — *Usenet newsgroup*
A **Usenet newsgroup** is a usually within the **Usenet** system, for **messages** posted from many users in different locations.

1204: who accompanied King louis the VII of France on the second crusade — *Second Crusade*
The **Second Crusade** was announced by **Pope** Eugene **III** , and was the **first** of the **crusades** to be led by **European kings**, namely **Louis VII** of **France** and **Conrad III** of **Germany** , with help from a number of other **European** nobles.

#1207: when did gary moore die — *Gary Moore*
Robert William **Gary Moore** (4 April 1952 – 6 February 2011), was a Northern Irish musician, most widely recognised as a singer and guitarist.

1209: what is a hosting company on a website — *Web hosting service*
A web **hosting service** is a type of **Internet hosting service** that allows individuals and organizations to make their **website** accessible via the World Wide Web .

1218: who owns exxon mobil — *ExxonMobil*
It is a direct descendant of John D. Rockefeller 's Standard Oil company, and was formed on November 30, 1999, by the **merger** of **Exxon** and **Mobil** .

1222: what is a rock quarry — *Quarry*
A **quarry** is a type of open-pit mine from which **rock** or minerals are extracted.

1223: what kind of school is MIT — *MIT Sloan School of Management*
The **MIT** Sloan **School** of **Management** (also known as **MIT** Sloan or Sloan) is the business **school** of the **Massachusetts** Institute of Technology , in Cambridge , **Massachusetts** , USA .

1225: what are banana plugs for — *Banana connector*
A **banana** connector (commonly **banana plug** for the male , **banana** socket or **banana** jack for the female) is a single-wire (one conductor) electrical connector used for joining wires to equipment .

#1229: when did the trojan war take place — *Trojan War*
The ancient **Greeks thought** that the **Trojan War** was a **historical event** that had **taken place** in the 13th or 12th **century BC**, and **believed** that Troy was **located** in modern-day Turkey near the Dardanelles .

1230: what is a letterbox movie — *Letterboxing (filming)*
The term refers to the shape of a **letter** box , a slot in a wall or door through which mail is delivered, being rectangular and wider than it is high.
Letterboxing is the practice of transferring film shot in a widescreen aspect ratio to standard-width video formats while preserving the film's original aspect ratio.
A 2.35:1 widescreen image **letterboxed** in a 1.33:1 screen

#1243: when did the civil war start and where — *American Civil War*
The **American Civil War** (ACW), also known as the **War** between the States or **simply** the **Civil War** (see naming), was a **civil war fought** from 1861 to **1865** between the United States (the "Union" or the "North") and several Southern **slave** states that declared their secession and formed the Confederate States of **America** (the "Confederacy" or the "South").

1244: What are the busiest airports in the world — *World's busiest airport*
The definition of the **world's busiest airport** has been specified by the **Airports** Council International in Geneva, Switzerland.

1248: where the streets have no name filming location — *Where the Streets Have No Name*
The song was notably performed on a Los Angeles **rooftop** for the **filming** of its music **video**, which won a Grammy Award for Best Performance Music **Video** .Recently the song has been used by the NFL's Baltimore Ravens as their entrance song in Super Bowl XLVII

1253: where did the mayflower land — *Mayflower*
The **Mayflower** was the ship that in 1620 transported 102 English Pilgrims , including a core group of Separatists , to New England.

1257: what languages are spoken in south africa — *Languages of South Africa*
The English version of the **South African** constitution refers to the **languages** by the names in those **languages**: isiZulu , isiXhosa , Afrikaans , Sepedi (referring to Northern Sotho), Setswana , English , Sesotho (referring to **Southern** Sotho), Xitsonga , Siswati , Tshivenda and isiNdebele (referring to **Southern** Ndebele).

1262: Who controlled Alaska before US? — *Alaska*
The name "**Alaska**" (Аляска) was already introduced in the **Russian** colonial period , when it was used only for the peninsula and is derived from the Aleut alaxsxaq, meaning "the mainland" or, more literally, "the object towards which the action of the sea is directed".
Alaska was purchased from **Russia** on March 30, 1867, for $7.2 million ($ adjusted for inflation) at approximately two cents per acre ($4.74/km2).
Alaska () is a state in the United States , situated in the northwest extremity of the North American continent , with the international boundary with Canada to the east, the Arctic Ocean to the north, and the Pacific Ocean to the west and south, with **Russia** further west across the Bering Strait .

1268: who sang black velvet — *Black Velvet (song)*
"**Black Velvet**" is a blues verse with a rock chorus written by Canadian songwriters Christopher Ward and David Tyson , recorded by Canadian **singer** songwriter Alannah Myles .

1274: who produced loyal to the game? — *Loyal to the Game*
Released in the United States on December 14, 2004 (December 12 in the United Kingdom), **Loyal** to the **Game** was **produced** by Eminem .

1275: WHAT ARE HERITABLE TRAITS — *Heritability*
heritability of a **trait** within a population is the proportion of observable differences in a **trait** between individuals within a population that is due to genetic differences.

1284: who does afge represent — *American Federation of Government Employees*
The American Federation of Government Employees (**AFGE**) is an American labor union **representing** over 650,000 employees of the federal government , about 5,000 employees of the District of Columbia , and a few hundred private sector employees, mostly in and around federal facilities.

1292: who wrote puff the magic dragon — *Puff, the Magic Dragon*
"**Puff**, the **Magic Dragon**" is a song **written** by Leonard Lipton and Peter Yarrow , and made popular by Yarrow's group Peter, Paul and Mary in a 1963 recording.

1301: what is a vetting process — *Vetting*
Vetting is the **process** of performing a background check on someone before offering them employment, conferring an award, etc.

#1303: when did the free soilers party start? — *Free Soil Party*
The **Free Soil Party** was a short-lived political **party** in the United States active in the 1848 and 1852 presidential elections, and in some state elections.

1308: where does the expression "knocking on wood" come from — *Knocking on wood*
Knocking on **wood**, or to touch **wood**, refers to the apotropaic tradition in western folklore of literally touching/**knocking** on **wood**, or merely stating that you are doing or intend same, in order to avoid " tempting fate " after making a favourable observation, a boast, or declaration concerning one's own death.

1310: who sang the nights the lights went out — *The Night the Lights Went Out in Georgia*
"The Night the **Lights Went** Out in Georgia" is a Southern Gothic song written by songwriter Bobby Russell and performed in 1972 by his then-wife Vicki Lawrence .

1315: what is a synthetic conduit — *Nerve guidance conduit*
A **nerve** guidance **conduit** (also referred to as an **artificial nerve conduit** or **artificial nerve graft**, as opposed to an autograft) is an **artificial** means of guiding axonal regrowth to facilitate **nerve** regeneration and is one of several clinical **treatments** for **nerve** injuries .

1316: what is a wwII theater — *European Theatre of World War II*
The **European Theatre** of World **War** II, also known as the **European War**, was a huge area of heavy fighting across **Europe** from Germany's invasion of Poland on **September** 1, 1939 until the end of the **war** with the German unconditional surrender on May 8, 1945 (V-E Day).

1318: what is amoxicillin for? — *Amoxicillin*
Amoxicillin is susceptible to degradation by β-lactamase -producing bacteria, which are resistant to a broad spectrum of β-lactam antibiotics, such as penicillin .
Amoxicillin is one of the most common antibiotics prescribed for children.
Amoxicillin (INN), formerly amoxycillin (BAN), and abbreviated amox, is a moderate-spectrum, bacteriolytic, β-lactam antibiotic used to treat bacterial infections caused by susceptible microorganisms .
Amoxicillin BP

1321: what state was the civil war in — *American Civil War*
The **American Civil War** (ACW), also known as the **War** between the **States** or **simply** the **Civil War** (see naming), was a **civil war** fought from **1861** to **1865** between the United **States** (the "Union" or the "North") and several **Southern slave states** that declared their **secession** and formed the Confederate **States** of **America** (the "Confederacy" or the "**South**").

1322: where did the early humans live? — *Human evolution*
According to the Recent African Ancestry theory , modern **humans evolved** in Africa possibly from **Homo heidelbergensis** , **Homo** rhodesiensis or **Homo** antecessor and migrated out of the continent some 50,000 to 100,000 years ago, replacing local populations of **Homo** erectus, **Homo** denisova , **Homo floresiensis** and **Homo** neanderthalensis.

1325: who invaded north africa during ww2 — *Operation Torch*
Operation Torch (initially called **Operation** Gymnast) was the British – American **invasion** of French **North Africa** in World War II during the **North African** Campaign , **started** on 8 November 1942.

#1326: WHAT YEARS WAS THE C5 VETTE PRODUCED — *Chevrolet Corvette (C5)*
The **Chevrolet** Corvette (**C5**) is a sports car **produced** by the **Chevrolet** division of General Motors for the 1997 through 2004 model **years** .

1328: What U.S. President's head has been featured on the nickel (five-cent coin) since 1938? *Nickel (United States coin)*
The **Buffalo nickel** was **introduced** in **1913** as part of a drive to increase the beauty of American **coinage**; in **1938**, the **Jefferson nickel** followed.

1329: who sings stand by me — *Stand by Me (song)*
"**Stand** by Me" is a song originally performed by Ben E. King and written by King, Jerry Leiber , and Mike Stoller , inspired by the spiritual "Lord **Stand** by Me," plus two lines rooted in Psalms 46:2–3.

#1331: When did F15s first fly — *McDonnell Douglas F-15 Eagle*
The **Eagle first flew** in July **1972**, and entered service in 1976.

1336: what are the arb medications — *ARB*
Angiotensin receptor blocker (antagonist), a **medication** for treating high blood pressure

1341: what are SATA Power Connector — *Serial ATA*
Serial ATA (**SATA**) is a computer bus interface that **connects** host bus adapters to mass storage **devices** such as hard disk **drives** and optical **drives** .

1344: What kind of poem is "This is Just to Say" — *This Is Just To Say*
"This Is Just To **Say**" (1934) is a famous imagist **poem** by William Carlos Williams .

1347: who played dumbledore in harry potter — *Albus Dumbledore*
Dumbledore is **portrayed** by Richard **Harris** in the **film** adaptions of **Harry Potter** and the Philosopher's **Stone** and **Harry Potter** and the **Chamber** of Secrets .

1348: what produces calcitonin — *Calcitonin*
Calcitonin (also known as thyrocalcitonin) is a 32-amino **acid** linear polypeptide hormone that is **produced** in humans primarily by the parafollicular cells (also known as C-cells) of the thyroid , and in many other animals in the ultimobranchial body .

1351: who hit the first home run at riverfront stadium? — *Riverfront Stadium*
On **June 30**, **1970**, the Reds hosted the Atlanta Braves in their grand opening, with Hank Aaron **hitting** the **first** ever **home run** at **Riverfront**.

1352: Who Makes Nexen Tires — *Nexen Tire*
Nexen Tire is a **tire** manufacturer, headquartered in Yangsan , South Gyeongsang Province , and Seoul , both in South Korea .

1353: who did richard nixon refer to as the silent majority — *Silent majority*
The term was popularized (though not first used) by U.S. President **Richard Nixon** in a November 3, 1969, speech in which he said, "And so tonight—to you, the great **silent majority** of my fellow Americans—I ask for your support."

1355: What are the lateral and median apertures of the brain? — *Lateral aperture*
It is an opening in each **lateral** extremity of the **lateral** recess of the fourth ventricle of the human **brain** , which also has a single **median aperture** .

1358: what part of the pig do pork chops come from — *Pork chop*
A **pork chop** is a **chop** of **pork** (a meat **chop**) cut perpendicularly to the spine of the **pig** and usually containing a rib or **part** of a vertebra, served as an individual portion.

1362: what are grits made from — *Grits*
Modern **grits** are commonly **made** of alkali-treated corn known as hominy .

1367: what state is new england in — *New England*
New England is a region in the northeastern corner of the United **States** consisting of the six **states** of **Maine** , **New** Hampshire , Vermont , Massachusetts , Rhode Island , and Connecticut .

1378: what are stanzas in poetry — *Stanza*
A **stanza** consists of a grouping of two or more lines , set off by a space, that usually has a set pattern of meter and rhyme.The **stanza** in **poetry** is analogous with the paragraph that is seen in prose , related thoughts are grouped into units.

1382: who shot franz ferdinand — *Assassination of Archduke Franz Ferdinand of Austria*
On 28 June 1914, **Archduke Franz Ferdinand** of **Austria** , heir presumptive to the Austro-Hungarian throne, and his wife, Sophie, Duchess of Hohenberg , were **shot** dead in Sarajevo , by Gavrilo **Princip** , one of a group of six Bosnian Serb assassins coordinated by Danilo Ilić .

1386: what are the three ossicles — *Ossicles*
The **ossicles** (also called auditory **ossicles**) are the **three** smallest bones in the human body, the malleus , the incus and the stapes .

1391: what is an army specialist — *Specialist (rank)*
Specialist (abbreviated "SPC") is one of the four junior enlisted ranks in the U.S. **Army** , just above Private First Class and equivalent in pay grade to **Corporal** .

1399: what to make with linen — *Linen*
Many products are **made** of **linen**: aprons, bags, towels (swimmers, bath, beach, **body** and wash towels), napkins, **bed linens**, **linen** tablecloths, runners, chair covers, and men's & women's wear.

1404: what south dakota county is wakonda in — *Wakonda, South Dakota*
Wakonda is a town in Clay **County** , **South Dakota** , United States .

1408: what are american people of japanese descent called — *Japanese American*
Japanese Americans have historically been among the three largest Asian **American** communities, but in recent decades, it has become the sixth largest group at roughly 1,304,286, including those of mixed-race or mixed-ethnicity.

1420: what is a bus adapter in a computer? — *Host adapter*
In **computer** hardware , a host controller, host **adapter**, or host **bus adapter** (HBA) connects a host system (the **computer**) to other network and storage devices.

1421: where does cashmere come from — *Cashmere wool*

Cashmere wool, usually simply known as **cashmere**, is a fiber obtained from **Cashmere** goats and other types of goat .

1423: Who invented egg rolls *Egg roll*
2__Egg_roll __ Varieties of **egg rolls** are found in mainland China , many Chinese-speaking regions of Asia, and Chinese immigrant communities around the world.

1424: what westerners include in pain and suffering *Pain and suffering*
Pain and **suffering** is the legal term for the physical and emotional stress caused from an injury (see also **pain** and **suffering**).

1429: who won the 2010 world cup *2010 FIFA World Cup*
In the **final**, **Spain** , the **European** champions , **defeated** third-time **finalists** the **Netherlands** 1–0 after extra **time** , with Andrés Iniesta 's **goal** in the 116th minute giving **Spain** their **first world title**, becoming the eighth nation to **win** the tournament, and the **first European** nation to **win** the tournament outside its home continent.

1431: who wrote the song in the mood *In the Mood*
"In the **Mood**" is a big band era #1 hit recorded by American bandleader Glenn Miller .

1434: what is active learning strategies *Active learning*
Active learning is an umbrella term that refers to several models of instruction that focus the responsibility of **learning** on **learners**.

1442: who sang I want to dance with somebody *I Wanna Dance with Somebody (Who Loves Me)*
"I Wanna **Dance** with Somebody (Who **Loves** Me)" is the first single from Whitney Houston 's second studio album Whitney .

1456: who sung the banana boat song *Day-O (The Banana Boat Song)*
"Day-O (The **Banana Boat** Song)" is a traditional Jamaican mento folk **song**, the best-known version of which was **sung** by Harry Belafonte and an alternate version interspersed with another Jamaican folksong , Hill and Gully Rider, by Dame Shirley Bassey .

#1459: what year did keeping up with the kardashians begin *Keeping Up with the Kardashians*
Keeping Up with the **Kardashians** (often referred to simply as The **Kardashians**) is an American reality television series that premiered on October 14, 2007, on E! .

1464: what is a store confectioner *Confectionery store*
A **confectionery store** (more commonly referred to as a sweet shop in the United Kingdom, a candy **store** in the North America, or a lolly shop in Australia) sells **confectionery** and is usually targeted to children.

#1465: when did qing dynasty begin *Qing Dynasty*
The **Qing Dynasty**, also **Empire** of the **Great Qing** or **Great Qing**, was the last imperial **dynasty** of **China** , ruling from **1644** to 1912 with a brief, abortive restoration in 1917.

1489: who wrote serenity prayer *Serenity Prayer*
The **Serenity Prayer** is the common name for an originally untitled **prayer** by the American theologian Reinhold Niebuhr (1892–1971).

1490: what are the Declaration and Resolves of the First Continental Congress about*Declaration and Resolves of the First Continental Congress*
The **Declaration** and **Resolves** of the **First Continental Congress** (also known as the **Declaration** of Colonial Rights, or the **Declaration** of Rights), was a statement adopted by the **First Continental Congress** on October 14, 1774, in response to the **Intolerable Acts** passed by the British Parliament .

1499: who created massey ferguson *Massey Ferguson*
The company was formed by a merger between **Massey** Harris and the **Ferguson** Company farm machinery manufacturer in 1953, **creating** the company **Massey** Harris **Ferguson**.

1506: where do the mohawks live *Mohawk people*
Their traditional homeland stretched southward of the **Mohawk** River , eastward to the Green Mountains of Vermont , westward to the border with the Oneida **Nation** 's traditional homeland **territory**, and northward to the St Lawrence River.

#1511: when did proof die *Proof (rapper)*
In 2006, **Proof** was shot and killed during an altercation at the CCC nightclub in Detroit.
DeShaun Dupree Holton (October 2, 1973 – April 11, 2006), better known by his stage name **Proof**, was an American rapper and actor from Detroit , Michigan .

1524: what are social security taxes *Social Security (United States)*
Tax deposits are formally entrusted to the **Federal** Old-Age and Survivors **Insurance Trust Fund**, the **Federal** Disability **Insurance Trust Fund**, the **Federal** Hospital **Insurance Trust Fund**, or the **Federal** Supplementary Medical **Insurance Trust Fund** which comprise the **Social Security Trust Fund** .

1527: what town is laurel hollow ny in *Laurel Hollow, New York*
Laurel Hollow is a village in the **Town** of Oyster Bay in Nassau County, New York in the United States .

1530: who won fifa world cup 2010 *2010 FIFA World Cup*
In the **final**, **Spain** , the **European** champions , **defeated** third-time **finalists** the **Netherlands** 1–0 after extra **time** , with Andrés Iniesta 's **goal** in the 116th minute giving **Spain** their **first world title**, becoming the eighth nation to **win** the tournament, and the **first European** nation to **win** the tournament outside its home continent.

#1531: when did egg mcmuffin get invented *McMuffin*
The **Egg McMuffin** is the signature breakfast sandwich; it was **invented** by the late McDonald's franchisee Herb Peterson in the late 1960s and was introduced nationwide in 1972.

1544: what the names of destiny's child *Destiny's Child*
Formed in 1990 in Houston , Texas, **Destiny's Child** members began their musical endeavors as Girl's Tyme comprising, among others, Knowles, Rowland, LaTavia Roberson and LeToya Luckett .
Destiny's Child was launched into mainstream recognition following the release of their best-selling second album, The Writing's on the Wall , which contained the number-one singles " Bills, Bills, Bills " and " Say My Name ".
Destiny's Child was an American R&B girl group whose final, and perhaps most recognizable, line-up comprised Beyoncé Knowles , Kelly Rowland and Michelle Williams .
Destiny's Child has sold more than 50 million records worldwide to date.
After years of limited success, they were signed to Columbia Records as **Destiny's Child**.

1551: What are procedure codes terms in coding? *Procedure codes*
Procedure codes are numbers or alphanumeric **codes** used to identify specific health interventions taken by medical professionals.

1562: what is a duvet cover used for *Duvet*
A **duvet** (or ; ; from the French **duvet** "down"), also known as a doona in Australian English or a continental **quilt** (or simply **quilt**) in British English , but this usage is no longer common, is a type of bedding , a soft flat bag filled with down , feathers , wool , silk or a synthetic alternative, and protected with a removable **cover**, analogous to a pillow and pillow case.

1563: who did mr bojangles *Mr. Bojangles (song)*
"**Mr**. **Bojangles**" is the title of a song originally written and recorded by American country music artist Jerry Jeff Walker for his 1968 album of the same title.

1574: what is an agents job role in film *Talent agent*
A **talent agent**, or booking **agent**, is a person who finds **jobs** for **actors** , authors , **film** directors , musicians , models , producers, professional athletes , writers and other people in various entertainment **businesses**.

1576: WHAT IS A LAW ENFORCEMENT MURDER BOOK *Murder book*
In **law enforcement** parlance, the term **murder book** refers to the case file of a **murder** investigation.

#1581: when album love always released *Love Always*
Love Always is the debut **album** of American R&B duo K-Ci & JoJo , **released** on June 17, 1997, by MCA Records .

1583: who wrote west side story *West Side Story*
West Side Story is an American **musical** with a book by Arthur **Laurents** , **music** by Leonard **Bernstein** , lyrics by Stephen **Sondheim** , and conception and choreography by Jerome Robbins .

1600: what states does interstate 70 travel through *Interstate 70*
Interstate 70 (I-70) is an **Interstate** Highway in the United **States** that runs from **Interstate** 15 near Cove Fort, **Utah** , to a Park and Ride near **Baltimore**, Maryland .

1604: where do mangos come from *Mango*
The **mango** is native to **South** Asia , from where it has been distributed worldwide to become one of the most cultivated fruits in the tropics .

1607: what video format will play in a DVD player *DVD-Video*
Discs **using** the DVD-Video specification **require** a **DVD drive** and an H.262/MPEG-2 Part **2** decoder (e.g., a **DVD player**, or a computer **DVD drive** with a **software DVD** player).

1610: what part of the government governs the US post office? *United States Postal Service*
The **United States Postal** Service (USPS), also known as the **Post Office** and U.S. Mail, is an independent **agency** of the **United States** federal **government** responsible for **providing postal** service in the **United States**.

1614: what motor does a 2001 monte carlo *Chevrolet Monte Carlo*
The **Monte Carlo** SS was **revived** from 2000 to **2007** and initially powered by 3.8-liter **V6** (supercharged in 2004–2005), later to be replaced by a 5.3-liter V8 for 2006–**2007**.

#1619: what year was President kennedy president? *John F. Kennedy*
John Fitzgerald "Jack" **Kennedy** (May 29, 1917 – **November** 22, 1963), often referred to by his initials JFK, was the 35th **President** of the United States , serving from 1961 until his **death** in **1963**.

1620: who owns popeyes chicken *Popeyes Louisiana Kitchen*
Often referred to as **Popeyes** and sometimes as **Popeyes Chicken** & Biscuits or **Popeyes Chicken** & Seafood, It was acquired by Sandy Springs, Georgia -based AFC Enterprises , originally America's Favorite **Chicken** Company, in 1993.

1624: who said give me liberty or give me death *Give me Liberty, or give me Death!*
"**Give** me **liberty**, or **give** me death!" is a quotation attributed to Patrick Henry from a speech he made to the Virginia Convention .

1630: what are the most known sports in america *Sports in the United States*
Four of the nation's five most **popular** team **sports** were developed in North **America**: **American football** , baseball , **basketball** and ice hockey , whereas **soccer** was developed in England.

1639: what is adem disease *Acute disseminated encephalomyelitis*
Acute disseminated encephalomyelitis (**ADEM**) is an immune mediated **disease** of the brain .

#1642: what year did john adams become president? *John Adams*
John Adams (**October** 30, 1735 (O.S. **October** 19, 1735) – July 4, 1826) was the second **president** of the **United** States (1797–1801), having earlier served as the **first vice president** of the **United** States .

1648: who created bing *Bing*
Bing (known previously as **Live Search**, Windows **Live Search**, and MSN **Search**) is a web **search** engine (advertised as a " decision engine ") from Microsoft .

1653: what are anti inflammatories *Anti-inflammatory*
Anti-inflammatory refers to the property of a substance or treatment that reduces **inflammation** .
Anti-inflammatory drugs make up about half of analgesics , remedying pain by reducing **inflammation** as opposed to opioids , which affect the central nervous system .

1656: who sang momma told me not to come *Mama Told Me Not to Come*
"Mama **Told** Me (Not to Come)" is a song by Randy Newman written for Eric Burdon 's first solo album in 1966.

1660: what are dialysis tubes *Dialysis tubing*
Dialysis tubing is a type of semi- or partially permeable membrane **tubing** made from regenerated cellulose or cellophane .

1662: who designed the statue of liberty *Statue of Liberty*
The **Statue** of **Liberty** (**Liberty** Enlightening the World; French: La **Liberté** éclairant le monde) is a neoclassical **sculpture** on **Liberty Island** in New York Harbor , **designed** by Frédéric Bartholdi and dedicated on **October** 28, 1886.

1667: what is am 2201 *AM-2201*
The toxicity of **AM-2201** is still a matter of debate and there may be long term side effects.
Reputed recreational use of **AM-2201** in the United States has led to it being specifically listed in a proposed 2011 amendment to the Controlled Substances Act , aiming to add a number of synthetic drugs into Schedule I .
AM-2201 (1-(5-fluoropentyl)-3-(1-naphthoyl)indole) is a research chemical that acts as a potent but nonselective full agonist for the cannabinoid receptor .

#1677: when did george washington died? *George Washington*
George Washington (– , 1799) was the **first President** of the **United States** (1789–1797), the commander-in-chief of the **Continental** Army during the American Revolutionary War , and one of the Founding Fathers of the **United States** .

1683: what are four thirds cameras *Four Thirds system*
The **Four Thirds** system is a standard created by Olympus and Kodak for digital single-lens reflex **camera** (DSLR) design and development.<ref name='DPReview.com 2001-02-13'>

1691: who wrote nature boy *Nature Boy*
"**Nature Boy**" is a song by Eden Ahbez , published in 1947.
Nat King Cole 's 1948 recording of the song was a major hit and "**Nature Boy**" has since become a pop and jazz standard , with dozens of major artists interpreting the song.

#1698: when did playstation 3 first launch *PlayStation 3 launch*
The **launch** of the **PlayStation 3** into the **Japanese market** on 10 **November** 2006 **marked** the second major seventh generation entertainment system to be released.

1708: who wrote the song cocaine *Cocaine (song)*
"**Cocaine**" is a **song written** and recorded by JJ Cale in 1976, but also known as a cover version recorded by Eric Clapton .

#1710: when did pearl harbor get bombed *Attack on Pearl Harbor*
The attack on **Pearl Harbor** (called Hawaii Operation or Operation AI by the Japanese Imperial General Headquarters (Operation Z in planning) and the Battle of **Pearl Harbor**) was a surprise military strike conducted by the Imperial Japanese Navy against the United States naval base at **Pearl Harbor** , Hawaii, on the morning of December 7, 1941 (December 8 in Japan).

#1714: when did marlins start *Miami Marlins*
The Miami **Marlins** began play in the 1993 season as the Florida **Marlins**.
Per agreement with the city and Miami-Dade County (which owns the park), the **Marlins** officially changed their name to the "Miami **Marlins**" on November 11, 2011.

1716: where do dungeness crab live *Dungeness crab*
The **Dungeness crab**, Metacarcinus magister (formerly Cancer magister), is a species of **crab** that inhabits eelgrass beds and water bottoms on the west coast of North America .

1720: who plays mary poppins in the movie *Mary Poppins (film)*
Mary Poppins is a 1964 musical film starring Julie Andrews , Dick **Van Dyke** , David Tomlinson , and Glynis Johns , **produced** by Walt **Disney** , and based on the **Mary Poppins books** series by P. L. **Travers** .

1722: What are the different translations for the Bible *Bible translations*
The full **Bible** has been **translated** into 518 **languages**, and 2798 **languages** have at least some portion of the **Bible**.

#1723: when does black friday start for christmas *Black Friday (shopping)*
Black Friday is the **day** following **Thanksgiving Day** in the United States , often regarded as the beginning of the **Christmas** shopping **season**.

#1730: when did Emperor Justinian die *Justinian I*
Justinian I () (,) (c. 482 – 14 November 565), commonly known as **Justinian** the **Great**, was Byzantine **Emperor** from 527 to 565.

1734: what movement was elizabeth cady stanton a part of *Elizabeth Cady Stanton*
Elizabeth Cady Stanton (November 12, 1815 – October 26, 1902) was an American social activist , abolitionist, and leading figure of the early women's rights **movement** .

1738: what radio station are the boston bruins on? *Boston Bruins Radio Network*
The **Boston Bruins Radio Network** is a 25-station (17 A.M., 8 F.M. plus 2 F.M. translators) CBS **Radio network** which carries live game broadcasts of the **Boston Bruins** .

#1739: when did ms .drgs go into effect *Diagnosis-related group*
The system is also referred to as "the **DRGs**", and its intent was to identify the "products" that a hospital provides.
DRGs may be further grouped into Major Diagnostic Categories (MDCs).
DRGs have been used in the US since 1982 to determine how much Medicare pays the hospital for each "product", since patients within each category are clinically similar and are expected to use the same level of hospital resources.
DRGs are assigned by a "grouper" program based on ICD (International Classification of Diseases) diagnoses, procedures, age, sex, discharge status, and the presence of complications or comorbidities .
DRGs are also standard practice for establishing reimbursements for other Medicare related reimbursements such as to home healthcare providers.

1741: what are the characteristics of an ethnic group *Ethnic group*
Ethnic identity is constantly reinforced through common **characteristics** which set the **group** apart from other **groups**.

1746: what mountains are on appalachian trail *Appalachian Trail*
The **Appalachian National** Scenic **Trail**, generally **known** as the **Appalachian Trail** or simply the A.T., is a marked hiking **trail** in the eastern United **States extending** between Springer Mountain in **Georgia** and Mount **Katahdin** in Maine .

1748: what are add ons *Add-on*
Add-ons for Firefox (addons.mozilla.org), the official Mozilla Foundation web site which serves as a repository of **add-ons** for Mozilla-based applications
Add-on (Mozilla) , a piece of software that enhances and customizes Mozilla-based applications
Add-on might mean:

1749: what would be the deliverables *Deliverable*
A **deliverable** also differs from a project document in that project document is typically part of a project **deliverable**, or a project **deliverable** may contain number of documents and physical things.
It may be either an outcome to be achieved (as in "The corporation says that becoming profitable this year is a deliverable.") or an output to be provided (as in "The **deliverable** for the completed project consists of a special-purpose electronic device and its controlling software.").
Deliverable is a term used in project management to describe a tangible or intangible object produced as a result of the project that is intended to be **delivered** to a customer (either internal or external).

A **deliverable** may be composed of multiple smaller **deliverables**.
A **deliverable** differs from a project milestone in that a milestone is a measurement of progress toward an output whereas the **deliverable** is the result of the process.
In technical projects, **deliverables** can further be classified as hardware , software , or design documents .
In contracted efforts, **deliverable** may refer to an item specifically required by contract documents, such as an item on a Contract Data Requirements List or mentioned in the Statement Of Work .
For a typical project, a milestone might be the completion of a product design while the **deliverable** might be the technical diagram of the product.
A **deliverable** could be a report, a document, a server upgrade or any other building block of an overall project.

1754: what school did oprah winfrey help — *Oprah Winfrey Leadership Academy for Girls*
The **Oprah Winfrey** Leadership Academy for Girls - South Africa is a female boarding **school** founded in January 2007 and located in Henley on Klip near Meyerton , south of Johannesburg , South Africa .

♯1757: when did scott pilgrim come out — *Scott Pilgrim vs. the World*
Scott Pilgrim vs. the **World** is a 2010 American **comedy** film directed by Edgar Wright , based on the graphic novel series **Scott Pilgrim** by Bryan **Lee** O'Malley .

1760: what is a google in math — *Googol*
In the binary numeral system , one would need 333 bits to represent a **googol**, i.e., 1 **googol** ≈ 2332.19, or exactly .
Other names for **googol** include ten duotrigintillion on the short scale , ten thousand sexdecillion on the long scale , or ten sexdecilliard on the Peletier long scale .
It is also the namesake of the internet company Google , with the name "Google" being a misspelling of "**googol**" by the company's founders.
A **googol** is the large number 10100; that is, the digit 1 followed by 100 zeroes :
A **googol** is approximately 70!
A **googol** has no particular significance in mathematics, but is useful when comparing with other very large quantities such as the number of subatomic particles in the visible universe or the number of hypothetically possible chess games.
1 **Googol** (1E100) is a small fraction more, than the biggest number a usual hand-held calculator can display and handle, which is 9.999...

1763: what structure is disulfide bonds — *Disulfide bond*
In chemistry , a **disulfide bond** (Br.E. disulphide **bond**) is a covalent **bond** , usually derived by the coupling of two thiol groups .

1768: who rules communism government — *Communist state*
It has a form of **government** characterized by single-party **rule** or dominant-party **rule** of a communist party (referred as Dictatorship of the Proletariat by its proponents) and a professed allegiance to a Leninist or Marxist–Leninist ideology as the guiding principle of the state.

1772: who makes skittles? — *Skittles (confectionery)*
Skittles is a brand of fruit-flavoured sweets , currently produced and marketed by the Wm. Wrigley Jr. Company , a division of Mars, Inc.

1778: what is a discipline of study — *List of academic disciplines*
An academic **discipline**, or field of **study**, is a branch of knowledge that is taught and researched at the college or university level.

1783: who wrote stand by me — *Stand by Me (song)*
"**Stand** by Me" is a song originally performed by Ben E. King and **written** by King, Jerry Leiber , and Mike Stoller , inspired by the spiritual "Lord **Stand** by Me," plus two lines rooted in Psalms 46:2–3.

1784: what are metaphors used for — *Metaphor*
This quote is a **metaphor** because the world is not literally a stage.
One of the most prominent examples of a **metaphor** in English literature is the All the world's a stage monologue from As You Like It :
Metaphor is a type of analogy and is closely related to other rhetorical figures of speech that achieve their effects via association, comparison or resemblance including allegory , hyperbole , and simile .
In simpler terms, a **metaphor** compares two objects or things without using the words "like" or "as".
A **metaphor** is a figure of speech that describes a subject by asserting that it is, on some point of comparison, the same as another otherwise unrelated object.

1788: what are the names of the ll divos — *Il Divo*
Il **Divo** is an English multinational operatic pop vocal group created by music manager, executive, and reality TV star Simon Cowell .
Il **Divo** is a group of four male singers: French pop singer Sébastien Izambard , Spanish baritone Carlos Marín , American tenor David Miller , and Swiss tenor Urs Bühler .

1792: What is a 28 day cycle? — *Menstrual cycle*
This article focuses on the human menstrual **cycle**, a "monthly" **cycle** that can vary around an average of ~**28 days** per **cycle**.

1797: what is a llc company? — *Limited liability company*
An **LLC** is a legal form of **company** that **provides** limited liability to its owners in the vast majority of **United States jurisdictions**.

1798: what are k cups — *K-Cup*
K-Cup portion packs are used with Keurig or other single **cup brewing systems** to **brew** a **cup** of coffee , tea , or hot chocolate .

1800: what is a chronograph watch — *Chronograph*
A **chronograph** is a specific type of **watch** that is used as a stopwatch combined with a display **watch**.

1801: what is a PCI port used for — *Conventional PCI*
Conventional PCI (**PCI** is an initialism formed from Peripheral Component Interconnect, part of the **PCI** Local Bus **standard** and often shortened to just **PCI**) is a local **computer** bus for attaching hardware devices in a **computer** .

♯1813: when did the movie deep blue sea come out — *Deep Blue Sea*
Deep Blue Sea is a 1999 science fiction horror film that stars Saffron Burrows , Thomas Jane , LL Cool J , Michael Rapaport , Stellan Skarsgård and Samuel L. Jackson .

1818: what are a and r reps — *Artists and repertoire*
Artists and **repertoire** (A&R) is the division of a record label or music publishing company that is responsible for talent scouting and overseeing the **artistic** development of recording **artists** and/or songwriters.

1819: what measurement is a furlong — *Furlong*
A **furlong** is a **measure** of distance in imperial units and U.S. customary units equal to one-eighth of a mile , equivalent to 220 yards , 660 feet , 40 rods , or 10 chains .

♯1825: when the wind blows james patterson — *When the Wind Blows (James Patterson novel)*
When the **Wind Blows** is a novel by **James Patterson** .

1829: what are the catholic gifts of the holy spirit — *Seven gifts of the Holy Spirit*
The **seven gifts** of the **Holy Spirit** is an enumeration of **seven spiritual gifts** originating with patristic authors, later elaborated by five intellectual **virtues** and four other groups of ethical characteristics.

1832: what are lobbying groups — *Lobbying*
Lobbying is done by many different types of people and organized **groups**, including individuals in the private sector , corporations , fellow legislators or government officials, or advocacy **groups** (interest groups).

1840: what nationality is kris jenner — *Kris Jenner*
Kristen Mary "**Kris**" **Jenner** (née Houghton, previously Kardashian; born November 5, 1955) is an American socialite, author and television personality.

♯1850: when did lucy stone died — *Lucy Stone*
Lucy Stone (August **13**, 1818 – October 19, **1893**) was a prominent American abolitionist and suffragist , and a vocal advocate and organizer promoting rights for women .

1851: what are stocks and bonds — *Bond (finance)*
Bonds and **stocks** are both securities , but the major **difference** between the two is that (capital) stockholders have an equity stake in the **company** (i.e. they are owners), whereas **bondholders** have a creditor stake in the **company** (i.e. they are lenders).

1852: what are the boundaries of the pelvic outlet — *Pelvic outlet*
The lower circumference of the lesser pelvis is very irregular; the space enclosed by it is named the inferior aperture or **pelvic outlet**.

♯1858: when did germans enter paris in ww2 — *Liberation of Paris*
The Liberation of **Paris** (also known as the Battle for **Paris**) took place during World War II from 19 **August** 1944 until the surrender of the occupying **German** garrison on **25 August**.

1859: who wrote the song feelin alright — *Feelin' Alright*
"**Feelin'** Alright?" also known as "**Feeling Alright**" is a **song written** by Dave Mason of the English rock band Traffic from their eponymous 1968 album, Traffic .

1860: what is a neuro tract — *Neural pathway*
A neural **pathway**, neural **tract**, or neural face, connects one part of the nervous system with another and usually consists of bundles of elongated, myelin -insulated neurons , known collectively as white matter .

♯1863: when did dr carter g woodson die — *Carter G. Woodson*
Carter Godwin **Woodson** (December 19, 1875April 3, 1950) was an African-American historian , author , journalist and the founder of the Association for the Study of African American Life and History .

1864: what are loan originations — *Loan origination*
Loan origination is the **process** by which a borrower **applies** for a new **loan** , and a lender **processes** that **application**.

♯1872: what year did aerosmith i dont want to miss a thing — *I Don't Want to Miss a Thing*
"I Don't **Want** to **Miss** a **Thing**" is a song performed by American rock **band Aerosmith** for the 1998 film Armageddon .

1886: who makes triumph motorcycles — *Triumph Motorcycles*
Triumph Motorcycles Ltd , a current British **motorcycle** manufacturer

---

1894: what produces primary xylem? — *Xylem*
The word **xylem** is derived from the Greek word ξύλον (xylon), meaning "**wood**"; the best-known **xylem** tissue is **wood** , though it is **found** throughout the plant.

1897: what is a ti 82? — *TI-82*
The **TI-82** was designed in 1993 as a stripped down, more user friendly version of the TI-85 , and as a replacement for the TI-81 .
The **TI-82** is a graphing calculator made by Texas Instruments .
Like the TI-81 , the **TI-82** features a 96x64 pixel display, and the core feature set of the TI-81 with many new features.

1899: what makes of the united states — *United States*
The **United States** of America (USA or U.S.A.), commonly called the **United States** (US or U.S.) or America, is a **federal** republic consisting of fifty **states** and a **federal district** .

♯1919: when did thomas jefferson become president — *Thomas Jefferson*
Thomas Jefferson (**April** 13, **1743** (**April** 2, **1743** O.**S**.) – July 4, 1826) was an **American** Founding **Father** , the principal author of the Declaration of Independence (1776) and the third **President** of the United States (1801–1809).

1920: who sang sun city — *Sun City (song)*
"**Sun City**" is a 1985 protest song written by Steven Van Zandt , produced by Van Zandt and Arthur Baker and recorded by Artists United Against **Apartheid** to convey opposition to the South African policy of **apartheid** .

1931: who wrote the song hallelujah — *Hallelujah (Leonard Cohen song)*
"**Hallelujah**" is a **song written** by **Canadian** singer-songwriter Leonard **Cohen** , originally released on his album Various Positions (1984).

1932: what are garnishments — *Garnishment*
A **garnishment** is a means of collecting a **monetary** judgment against a defendant by ordering a third party (the **garnishee**) to pay **money**, otherwise owed to the defendant , directly to the plaintiff .

1944: what are land parcels — *Parcel*
land lot , a piece of **land**;

1945: what state is Mn — *Minnesota*
Minnesota () is a **U.S. state** located in the Midwestern United **States** .

1947: what naturally occurring isotopes does cobalt have — *Isotopes of cobalt*
Naturally occurring cobalt (Co) is composed of 1 stable **isotope** , 59Co. 28 radioisotopes have been characterized with the most stable being 60Co with a half-life of 5.2714 years, 57Co with a half-life of 271.79 days, 56Co with a half-life of 77.27 days, and 58Co with a half-life of 70.86 days.

1948: Who Moved My Cheese Synopsis — *Who Moved My Cheese?*
It describes **change** in one's work and life, and four typical reactions to said **change** by two mice and two "littlepeople", during their hunt for **cheese** .

1949: who replaced nikita khrushchev — *Nikita Khrushchev*
Khrushchev's party colleagues removed him from power in **1964**, **replacing** him with **Leonid Brezhnev** as First Secretary and Alexei Kosygin as Premier.

1953: What Are Mnemonic Devices — *Mnemonic*
A **mnemonic** (, with a silent "m"), or **mnemonic device**, is any **learning** technique that aids **information** retention.

1962: What is a surveyors wheel — *Surveyor's wheel*
A **surveyor's wheel**, also called a clickwheel, hodometer, waywiser, trundle **wheel** , measuring **wheel**, or perambulator is a device for measuring distance.

♯1967: what year was girls just want to have fun release — *Girls Just Want to Have Fun*
"**Girls** Just **Want** to Have **Fun**" is a 1979 song originally written by Robert Hazard and made famous by singer Cyndi Lauper .

1972: what are the players in Quidditch? — *Quidditch*
Matches are **played** between two teams of seven **players** riding flying broomsticks, using four balls and six elevated ring-shaped goals, three on each side of the **Quidditch** pitch (field).

1975: who wrote the song a little more country than that> — *A Little More Country Than That*
"A **Little** More **Country** Than That" is the title of a **song written** by Joey + Rory 's Rory Lee Feek , Wynn Varble , and Don Poythress, and recorded by American **country** artist Easton Corbin .

1976: who makes blackberry — *BlackBerry*
The **BlackBerry** is a line of wireless handheld devices and services designed and marketed by Research In Motion **Limited** (RIM) operating as **BlackBerry** .

1979: what are arizona's symbols — *List of Arizona state symbols*
The newest adopted **symbol** of **Arizona** is the Colt Single Action Army in 2011.
The following is a list of **symbols** of the U.S. state of **Arizona**.
The first **symbol** was the motto, which was made official in 1864 for the **Arizona** Territory .
Fifteen of the state **symbols** are on display on the **Arizona** Capitol Museum .

1983: what is adoration catholic church — *Eucharistic adoration*
Eucharistic **adoration** is a practice in the Roman **Catholic Church** , and in a few Anglican and Lutheran **churches**, in which the Blessed Sacrament is exposed and **adored** by the faithful.

♯1984: when the body is systemic — *Systemic*
Systemic refers to something that is spread throughout, **system-wide**, affecting a group or **system** such as a **body**, economy, market or society as a whole.

1985: what are SLR cameras — *Single-lens reflex camera*
A single-lens reflex (**SLR**) **camera** is a **camera** that typically uses a mirror and prism system (hence "reflex", from the mirror's reflection) that permits the photographer to view through the lens and see exactly what will be captured, contrary to viewfinder **cameras** where the image could be significantly different from what will be captured.

1986: what is a vm server — *Virtual machine*
A virtual machine (**VM**) is a software implemented **abstraction** of the underlying hardware , which is presented to the application layer of the **system**.

♯1998: when did the titanic sink — *RMS Titanic*
RMS **Titanic** was a British passenger liner that **sank** in the North Atlantic Ocean on 15 April 1912 after colliding with an **iceberg** during her **maiden voyage** from **Southampton** , UK to **New York** City , US.

1999: where is the human thigh located? — *Human leg*
The **human leg** is the entire **lower extremity** or **limb** of the **human** body , including the foot , **thigh** and even the hip or gluteal region; however, the precise definition in **human** anatomy refers only to the section of the **lower limb** extending from the **knee** to the **ankle** .

2001: what is considered a large car — *Full-size car*
A full-size **car** is a marketing term used in North America for an automobile **larger** than a mid-size **car** .

2004: who is flo from progressive — *Flo (Progressive Insurance)*
In 2011, **Progressive** introduced an Australian counterpart to **Flo**, named Kitty, played by Australian actress Holly Austin. **Flo** is a fictional character who appears in commercials for **Progressive** Insurance .

2009: what is petit lemans — *Petit Le Mans*
The **Petit** Le Mans (French for little Le Mans) is a sports car endurance race held annually at Road Atlanta in Braselton, Georgia , USA.
The **Petit** Le Mans covers a maximum of (which is approximately 394 laps) or a maximum of 10 hours, whichever comes first; only once, in the rain-stopped 2009 race, has the leading team failed to complete .

2015: what is considered to be a disasters — *Disaster*
Developing countries suffer the greatest costs when a **disaster** hits – more than 95 percent of all deaths caused by **disasters** occur in developing countries, and losses due to natural **disasters** are 20 times greater (as a percentage of GDP) in developing countries than in industrialized countries.
Ruins from the 1906 San Francisco earthquake , remembered as one of the worst natural **disasters** in United States history
In contemporary academia, **disasters** are seen as the consequence of inappropriately managed risk .
Hazards that strike in areas with low vulnerability will never become **disasters**, as is the case in uninhabited regions.

A **disaster** is a natural or man-made (or technological) hazard resulting in an event of substantial extent causing significant physical damage or destruction, loss of life, or drastic change to the environment .
A **disaster** can be ostensively defined as any tragic event stemming from events such as earthquakes , floods , catastrophic accidents , fires , or explosions .

2017: What is firebird server *Firebird (database server)*
The database forked from Borland 's open source edition of InterBase in 2000, but since **Firebird** 1.5 the code has been largely rewritten.
Firebird is an open source SQL relational database management system that runs on Linux , Windows , and a variety of Unix .

2019: What is caused by the human immunodeficiency virus? *HIV*
Human immunodeficiency virus (HIV) is a lentivirus (slowly replicating retrovirus) that **causes** acquired **immunodeficiency** syndrome (AIDS), a condition in **humans** in which progressive failure of the immune system **allows** life-threatening opportunistic infections and cancers to thrive.

2026: what is the function of the vas deferens? *Vas deferens*
The **vas deferens** (plural: vasa deferentia), also called ductus **deferens** (Latin : "carrying-away vessel"; plural: ductus deferentes), is part of the male anatomy of many vertebrates ; they transport sperm from the epididymis in anticipation of ejaculation .

2027: What is Ischemia or infarction? *Ischemia*
Vascular **ischemia** of the toes with characteristic cyanosis .
Ischemia is generally caused by problems with blood vessels , with resultant damage to or dysfunction of tissue.
In medicine , **ischemia**, also spelled as ischaemia or ischæmia, (; from Greek language ισχαιμία, ischaimia; isch- root denoting a restriction or thinning or to make or grow thin/lean, haema blood) is a restriction in blood supply to tissues , causing a shortage of oxygen and glucose needed for cellular metabolism (to keep tissue alive).

2035: where is humboldt ks *Humboldt, Kansas*
Humboldt is a city situated along the Neosho River in the southwest part of Allen County , located in southeast Kansas , in the Central United States .

2036: what was nixon accused of *Watergate scandal*
The Watergate **scandal** was a political **scandal** that occurred in the United States in the 1970s as a result of the June 17, 1972 break-in at the Democratic National Committee headquarters at the Watergate office complex in Washington, D.C., and the **Nixon administration's** attempted cover-up of its involvement.

2043: who is victoria jackson from saturday night live *Victoria Jackson*
Victoria Jackson (born August 2, 1959) is an American comedian, actress, satirist, singer and internet blogger best known as a cast member of the NBC television sketch comedy series **Saturday Night Live** (SNL) from 1986 to 1992.

2044: what was the post modernist era in literature? *Postmodern literature*
Postmodern **literature** is **literature** characterized by heavy reliance on techniques like **fragmentation**, paradox, and questionable narrators, and is often (though not exclusively) defined as a style or **trend** which emerged in the post–World War II **era**.

2045: who is the group enigma *Enigma (musical project)*
The Romanian-born Cretu conceived the **Enigma** project while working in Germany, but based his recording studio A.R.T. Studios in Ibiza, Spain , from the early 1990s until May 2009, where he has recorded all of **Enigma's** studio releases to date.

2049: what is the federal death tax *Estate tax in the United States*
The estate **tax** in the United **States** is a **tax** imposed on the **transfer** of the "**taxable** estate " of a deceased **person**, whether such **property** is **transferred** via a will , according to the **state** laws of intestacy or otherwise made as an incident of the **death** of the owner, such as a **transfer** of **property** from an intestate estate or trust, or the payment of **certain** life insurance benefits or financial account sums to beneficiaries.

♯2052: when was raphael born *Raphael*
Raffaello Sanzio da **Urbino** (April **6** or March 28, 1483 – April **6**, 1520), better known simply as **Raphael**, was an Italian painter and architect of the High Renaissance .

2059: where is the seed located in an artichoke *Artichoke*
The edible matter is **buds** that form within the **flower** heads before the **flowers** come into bloom.

2065: WHAT IS PARESTHESIAS OF HANDS *Paresthesia*
The manifestation of **paresthesia** may be transient or chronic.
Paresthesia (or), is a sensation of tickling, tingling, burning, pricking, or numbness of a person's skin with no apparent long-term physical effect.

2072: who is paul avery to the zodiac killings *Paul Avery*
Paul Avery (April 2, 1934December 10, 2000) was an American police reporter, best known for his stories on the infamous serial **killer** known as the **Zodiac** , and later for his work on the Patricia Hearst kidnapping.

2074: what is sump pump used for *Sump pump*
A **sump pump** is a **pump used** to remove water that has accumulated in a water collecting **sump** basin, commonly found in the basement of homes.

♯2075: when was bloody kansas in *Bleeding Kansas*
Bleeding **Kansas**, **Bloody Kansas** or the Border War, was a series of violent political confrontations involving anti-slavery Free-Staters and pro-slavery " Border Ruffian " elements, that took place in the **Kansas** Territory and the neighboring towns of Missouri between 1854 and 1861.

2076: What Is Benzene Used For *Benzene*
It is mainly **used** as a precursor to heavy **chemicals**, such as **ethylbenzene** and cumene , which are **produced** on a billion kilogram scale.

2079: what is homebrew for wii? *Wii homebrew*
Wii homebrew refers to the reuse of Nintendo 's **Wii** game console to run software that has not been authorized by Nintendo .
In more general terms, **Wii Homebrew** refers to the use of the **Wii** 's hardware, accessories and software for purposes outside those intended by the manufacturer.

2081: who was bush's national security advisor *Stephen Hadley*
Stephen John Hadley (born February 13, 1947) was the 21st U.S. Assistant to the President for **National Security** Affairs (commonly referred as **National Security Advisor**), serving under President George W. **Bush** .

2088: what is another name for the large intestine *Large intestine*
The **large intestine** (or bowel, colon) is the last part of the digestive system in vertebrate animals .
The **large intestine** is about long, which is about one-fifth of the whole length of the intestinal canal .
The **large intestine** consists of the cecum , rectum and anal canal .
In Terminologia Anatomica the **large intestine** includes the cecum, colon, rectum, and anal canal.

2090: where is the valley at in los angeles *San Fernando Valley*
The San Fernando **Valley** (locally known as "The **Valley**") is an urbanized **valley** located in the **Los Angeles** metropolitan area of southern California , United States , defined by the mountains of the Transverse Ranges circling it.

♯2094: when is susan smith eligible for parole *Susan Smith*
According to the South Carolina Department of Corrections , **Smith** will be **eligible** for **parole** on November 4, 2024, after serving a minimum of thirty years.

♯2098: when was everybody hates chris made *Everybody Hates Chris*
Everybody **Hates Chris** is an American television period sitcom that depicts the teenage experiences of comedian **Chris** Rock (who is also the narrator) while growing up in the Bedford-Stuyvesant neighborhood of Brooklyn, New York .

2099: what is lung effusion *Pleural effusion*
Pleural **effusion** is excess fluid that accumulates between the two pleural layers , the fluid-filled **space** that surrounds the **lungs**.

2104: who is shem in the bible *Shem*
Shem (; Sēm; Arabic : Sām; Ge'ez : □□, Sēm; "renown; prosperity; name") was one of the sons of Noah in the Hebrew **Bible** as well as in Islamic literature .
However, the New American Standard **Bible** gives, "Also to **Shem**, the father of all the children of Eber, and the older brother of Japheth, children were born."

2105: what is go daddy.com? *Go Daddy*
Go Daddy or **Go Daddy** Group **Inc**. is a privately held **company** that is primarily an internet domain **registrar** and web hosting **company**.

2106: where is andy whitfield from? *Andy Whitfield*
Andy Whitfield (died 11 September 2011) was a Welsh Australian actor and model.

2107: where is the country andorra located *Andorra*
Andorra (; ,), officially the Principality of **Andorra** (), also called the Principality of the Valleys of **Andorra**, (), is a landlocked microstate in Southwestern Europe , **located** in the eastern Pyrenees mountains and bordered by **Spain** and France .

2110: what is the highest point in oahu *Oahu*
The **highest point** is Mt. Ka'ala in the Waianae Range, rising to above sea level.

2113: what is spelt flour *Spelt*
Spelt was an important staple in parts of Europe from the Bronze Age to medieval times ; it now survives as a relict crop in Central Europe and northern Spain and has found a new market as a health food.
Spelt is sometimes considered a subspecies of the closely related species common wheat (T. aestivum), in which case its botanical name is considered to be Triticum aestivum subsp. spelta.
Spelt, also known as dinkel wheat, or hulled wheat, is an ancient species of wheat from the fifth millennium BC.

2118: where is kennywood in pittsburgh *Kennywood*
Kennywood is an amusement park located in West Mifflin , **Pennsylvania** , a suburb of **Pittsburgh** .

2125: Who is the rap singer in right round with kesha in the background? *(dancer)* *Right Round*
"**Right Round**" is a song performed by **American rapper** Flo Rida , **released** as the lead **single** from his second studio album, R.O.O.T.S. (2009).

2128: what is rock of ages about *Rock of Ages (musical)*
Rock of **Ages** is a **rock** / jukebox **musical** , with a book by Chris D'Arienzo, built around classic **rock** hits from the 1980s, especially from the famous glam metal bands of the decade.

2134: what is the singer Stevie Wonders full name *Stevie wonder*
Stevland Hardaway Morris (**born** May 13, 1950 as **Stevland Hardaway** Judkins), **known** by his stage **name Stevie Wonder**, is an American **singer**, songwriter, and multi-instrumentalist, a child prodigy who developed into one of the most creative **musical** figures of the late 20th century.

2140: what is general chu chicken *General Tso's chicken*
General Tso's chicken (sometimes Governor **Tso's chicken**, **General** Gau's **chicken**, **General** Tao's **chicken**, **General** Tsao's **chicken**, **General** Tong's **chicken**, **General** Tang's **chicken** or simply General's **Chicken**) is a sweet , slightly spicy , deep-fried **chicken** dish that is popularly served in North American Chinese restaurants.

2142: where is david ortiz from *David Ortiz*
David Américo **Ortiz** Arias (born November 18, 1975), nicknamed "Big Papi", is a Dominican-American professional baseball designated hitter with the Boston Red Sox of Major League Baseball (MLB).

2143: what is an irregular heartbeat pvc *Premature ventricular contraction*
A premature ventricular contraction (PVC), also known as a premature ventricular **complex**, ventricular premature contraction (or **complex** or **complexes**) (VPC), ventricular premature beat (VPB), or ventricular extrasystole (VES), is a relatively common event where the **heartbeat** is initiated by Purkinje fibres in the ventricles rather than by the sinoatrial node , the normal **heartbeat** initiator.

2150: what is flour made from *Flour*
Flour is a powder which is **made** by grinding cereal grains , or other seeds or roots (like **Cassava**).

2156: what is cubic ft *Cubic foot*
To calculate **cubic** feet multiply length X width X height.
The term **cubic** foot is an Imperial and US customary (non- metric) unit of volume , used in the United States and the United Kingdom.
It is defined as the volume of a **cube** with sides of one foot (0.3048 m) in length .

♯2157: when was kirstie alley on cheers *Kirstie Alley*
Kirstie Louise **Alley** (born January 12, 1951) is an American actress and comedian known for her role in the TV series **Cheers** , in which she played Rebecca Howe from 1987–1993, winning an Emmy Award and a Golden Globe Award as the Outstanding Lead Actress in a Comedy Series in 1991.

2161: what is another name for cpu *Central processing unit*
A central processing unit (CPU), also referred to as a central processor unit, is the **hardware** within a computer that carries out the instructions of a computer program by performing the basic arithmetical, logical, and input/output operations of the system.

2169: what was the city of Mithridates *Mithridates VI of Pontus*
Mithridates VI or Mithradates VI (), from Old Persian Mithradatha, "gift of Mithra "; 134–63 BC, also known as Mithradates the Great (Megas) and Eupator Dionysius, was king of Pontus and Armenia Minor in northern Anatolia (now Turkey) from about 120–63 BC.
Mithridates is remembered as one of the Roman Republic 's most formidable and successful enemies, who engaged three of the prominent generals from the late Roman Republic in the Mithridatic Wars : Lucius Cornelius Sulla , Lucullus and Pompey .

2174: what is extreme right wing *Far-right politics*
The far **right** is commonly associated with persons or groups who hold **extreme** nationalist , xenophobic , racist , religious fundamentalist or reactionary views.
The far-right (also known as the **extreme right**) refers to the highest degree of rightism in right-wing politics .

2177: what is high emotional intelligence? *Emotional intelligence*
Emotional intelligence (EI) is the ability to identify, assess, and control the **emotions** of oneself, of others, and of groups.

2178: what is the role of heredity *Heredity*
Heredity is the passing of **traits** to offspring from its parents or ancestor.

♯2186: what is prefix phone number *Telephone prefix*
A telephone **prefix** is the first set of digits of a telephone **number**; in the North American Numbering Plan countries (country code 1), it is the first three digits of a seven-digit **phone number**.

2187: what is metformin used for *Metformin*
Metformin (BP , pronounced , ; originally sold as **Glucophage**) is an **oral** antidiabetic drug in the biguanide class.

2190: what is it like at mount rushmore? *Mount Rushmore*
The **Mount Rushmore** National Memorial is a sculpture carved into the granite face of **Mount Rushmore** near Keystone , South Dakota , in the United States .

2195: where is keith whitley from *Keith Whitley*
Jackie **Keith Whitley** (July 1, 1954Stambler, Irwin, and Grelun Landon (2000). - Country Music: The Encyclopedia. - New York: St. Martin's Press. - p.533. - ISBN 978-0-312-26487-1.—Carlin, Richard (2003). - Country Music: A Biographical Dictionary. - New York: Routledge - p.427. - ISBN 978-0-415-93802-0.—Larkin, Colin (1995). - The Guinness Encyclopedia of Popular Music. - New York: Stockton Press - P.4462. - ISBN 978-0-85112-662-3.—Stanton, Scott (2003). - The Tombstone Tourist: Musicians. - New York: Pocket Books. - p.395. - ISBN 978-0-7434-6330-0.—Hicks, Jack. - "Singer **Keith Whitley's** Memory Alive Through Songs, Love in Home Town". - The Kentucky Post . - September 25, 1991.—"Country Music Star **Keith Whitley** Dead at 33". - Lexington Herald-Leader . - May 10, 1989.——Hurst, Jack. - "**Whitley's** Last Days". - Chicago Tribune . - May 14, 1989.—"Alcohol Kills Country Singer **Keith Whitley**". - United Press International . - (c/o The San Francisco Chronicle). - May 10, 1989. — May 9, 1989), known professionally as **Keith Whitley**, was an American country music singer.

2200: who is basketball star antoine walker *Antoine Walker*
Antoine Devon **Walker** (born August 12, 1976) is an American former professional **basketball** player.

2202: where are the internal and external iliac arteries *External iliac artery*
The **external iliac arteries** are two major **arteries** which bifurcate off the common **iliac arteries** anterior to the sacroiliac joint of the pelvis.

2207: where is the Rhine river located on a map *Rhine*
The **Rhine** (; ;) is a European **river** that runs from the Swiss canton of Grisons in the southeastern Swiss Alps through **Germany** and eventually flows into the North Sea coast in the **Netherlands** and is the twelfth longest **river** in Europe , at about , with an average discharge of more than .

2208: what is sanskrit shri *Sri*
Sri (Devanagari : □□□, IAST ; Śrī), also transliterated as Sree or **Shri** or Shree is a word of **Sanskrit** origin, used in the Indian subcontinent as polite form of address equivalent to the English "Mr." in written and spoken language, or as a title of veneration for deities (usually translated as "Holy").

2214: WHere is a famous alluvial plain *Mississippi Alluvial Plain*
The Mississippi River **Alluvial Plain** is an **alluvial plain** created by the Mississippi River on which lies parts of seven U.S. states , from southern Louisiana to southern Illinois .

2215: Where is South Beach in Miami *South Beach*
South Beach, also nicknamed **SoBe**, is a neighborhood in the city of **Miami Beach** , Florida , United States , located due east of **Miami** city proper between Biscayne Bay and the Atlantic Ocean .

2233: what is bones job *Bones (TV series)*
The show is based on forensic anthropology and forensic archaeology , with each **episode** focusing on an FBI case file concerning the mystery behind human remains brought by FBI Special Agent Seeley Booth (David **Boreanaz**) to the forensic anthropologist Dr. Temperance "**Bones**" Brennan (**Emily Deschanel**).

♯2235: when was malcolm x assassinated *Malcolm X*
Malcolm X (; May 19, 1925February 21, 1965), born **Malcolm Little** and also **known** as El-Hajj Malik El-Shabazz (), was an African-American Muslim minister and human rights activist.

2239: where is penn state stadium *Beaver Stadium*
Beaver **Stadium** is an outdoor college football **stadium** in University Park , Pennsylvania , United **States** , on the campus of The Pennsylvania **State** University .

2240: what is in a tequila sunrise? *Tequila Sunrise (cocktail)*
The **Tequila Sunrise** is a cocktail made in two different ways, the original (**tequila** , crème de cassis , lime juice and soda water) and the more popular concoction (**tequila** , orange juice , and grenadine syrup).

2244: where are facial sinuses — *Paranasal sinuses*
Paranasal **sinuses** are a group of four paired air-filled spaces that surround the nasal cavity (maxillary **sinuses**), above the eyes (frontal **sinuses**), between the eyes (ethmoid **sinuses**), and behind the ethmoids (sphenoid **sinuses**).

2246: what is lean manufacturing and who developed — *Lean manufacturing*
Lean manufacturing, **lean** enterprise, or **lean** production, often simply, "Lean," is a production practice that considers the expenditure of resources for any goal other than the creation of value for the end customer to be wasteful, and thus a target for elimination.

2253: what is the capital city of california. — *Sacramento, California*
Sacramento is the **capital city** of the U.S. **state** of **California** and the seat of **government** of Sacramento **County** .

2255: What was "Freedom Summer"? — *Freedom Summer*
Freedom Summer (also known as the Mississippi **Summer** Project) was a campaign in the United States launched in June 1964 to attempt to register as many African American voters as possible in Mississippi , which had historically excluded most blacks from voting.

2264: what was the Name the first electronic handheld calculator — *Sumlock ANITA calculator*
The ANITA Mark VII and ANITA Mark VIII **calculators** were launched simultaneously in late 1961 as the world's **first** all- **electronic** desktop **calculators** .

♯2265: when was washington elected president — *George Washington*
George Washington (– , 1799) was the **first President** of the **United States** (1789–1797), the commander-in-chief of the **Continental** Army during the American Revolutionary War , and one of the Founding Fathers of the **United States** .

2279: what is customary at shiva — *Shiva (Judaism)*
This state lasts for seven **days**, during which family members traditionally gather in one home (preferably the home of the deceased) and receive visitors.

2281: who is the guy in the wheelchair who is smart — *Stephen Hawking in popular culture*
Professor Stephen Hawking , known for being a theoretical physicist , has appeared in many works of popular culture .

2283: what is in a hot toddy — *Hot toddy*
A **hot toddy**, also **hot** totty, and **hot** tottie, is typically a mixed drink made of liquor and water with sugar and spices and served **hot**.

2284: what is honey bee propolis — *Propolis*
Propolis is a resinous mixture that **honey bees** collect from tree buds, sap flows, or other botanical sources.

♯2285: when was steven tyler born — *Steven Tyler*
Steven Tyler (**born Steven** Victor Tallarico; March 26, 1948) is an American **singer**, songwriter, and multi-instrumentalist, best known as the frontman and lead **singer** of the **Boston** -based rock band Aerosmith , in which he also plays the harmonica, and occasional piano and percussion.

2286: what is name of national anthem song of switzerland — *Swiss Psalm*
The Swiss Psalm (, , ,) is the **national anthem** of **Switzerland** .

2287: what is blood urea — *Blood urea nitrogen*
Normal human adult **blood** should contain between 7 to 21 mg of **urea** nitrogen per 100 ml (7–21 mg/ dL) of **blood**.

2291: what is corpus christi holiday — *Corpus Christi (feast)*
The Feast of **Corpus Christi** (Latin for Body of **Christ**), also known as **Corpus** Domini, is a Latin Rite liturgical solemnity celebrating the tradition and belief in the body and blood of Jesus **Christ** and his Real Presence in the Eucharist .

2296: what is the difference between multistage and cluster sampling — *Cluster sampling*
In this technique, the total population is divided into these groups (or **clusters**) and a **simple** random **sample** of the groups is selected.
Cluster sampling is a **sampling** technique used when "natural" but relatively homogeneous groupings are evident in a statistical population .

2297: where is jamestown north carolina — *Jamestown, North Carolina*
Jamestown is a town in Guilford County , **North Carolina** , United States , and is a suburb of the nearby cities of Greensboro and High Point.

2302: what is stepwise linear regression — *Stepwise regression*
In statistics , **stepwise regression** includes **regression** models in which the choice of predictive variables is carried out by an automatic procedure.

♯2314: when was andy griffith born — *Andy Griffith*
Andy Samuel Griffith (June 1, 1926 – July 3, 2012) was an American actor, television producer, Grammy Award -winning Southern-gospel singer, and writer.

2315: what is the definition of a hung jury — *Hung jury*
A **hung jury** or deadlocked **jury** is a **jury** that cannot, by the required voting threshold, agree upon a verdict after an extended period of deliberation and is unable to change its votes.

2321: what is the capacity of the cowboy stadium — *Cowboys Stadium*
The **stadium** seats 85,000, making it the third **largest stadium** in the NFL by seating **capacity** .

2326: what is the color puce — *Puce*
The **colors** in the boxes at right are two of the various shades and varieties of **puce**.
Puce (often misspelled as "puse", "peuse" or "peuce") is defined in the United States as a brownish-purple **color**.

2333: where was jfk buried — *State funeral of John F. Kennedy*
After the Requiem Mass at St. Matthew's Cathedral , the late president was **buried** at **Arlington** National Cemetery in Virginia.

2337: what is renaissance english — *English Renaissance*
The **English Renaissance** was a cultural and **artistic** movement in **England** dating from the late 15th and early 16th centuries to the early 17th century.

2338: Who is the husband of Betty Ford — *Betty Ford*
Elizabeth Ann Bloomer Warren "**Betty**" **Ford** (April **8**, 1918 – **July 8**, 2011), was First Lady of the United States from 1974 to 1977 during the presidency of her **husband** Gerald **Ford** .

2344: what was Coco Chanel's real first name? — *Coco Chanel*
Gabrielle "**Coco**" Bonheur **Chanel** (August 19, 1883 – **January** 10, **1971**) was a French fashion designer and founder of the **Chanel** brand.

2345: what is grist mill stone — *Gristmill*
A **gristmill** (also: **grist mill**, corn **mill** or flour **mill**) grinds grain into flour .

2352: where is big pokey from — *Big Pokey*
Milton Powell (born December 4, 1977), better known by his stage name **Big Pokey**, is a rap artist from Houston, Texas and is one of the original members of the Screwed Up Click .
Big Pokey joined up with DJ Screw in the early 1990s and started releasing songs on DJ Screw's many mixtapes .

2353: what is baklava recipe — *Baklava*
Baklava (, , or ; also Baklawa) is a rich, sweet pastry made of layers of phyllo pastry filled with chopped nuts and sweetened with syrup or honey .

2354: what is the rule of the 9s — *Total body surface area*
In adults, the "**rule** of nines" is used to determine the total percentage of area burned for each major section of the body.

2357: what is the sign for degrees — *Degree symbol*
The **degree** symbol (°) is a typographical symbol that is used, among other things, to represent **degrees** of arc (e.g. in geographic coordinate systems), hours (in the medical field), or **degrees** of temperature .

2361: who is carlos pena on big time rush — *Big Time Rush*
Big Time Rush (BTR) is an American television series created by Scott Fellows about the Hollywood misadventures of four hockey players from Minnesota —Kendall, James, **Carlos**, and Logan, after they are selected to form a boy band .

2370: WHO IS HENRY SAMPSON JR. — *Henry Sampson (inventor)*
Henry T. Thomas **Sampson**, **Jr**. (born in **Jackson**, Mississippi in 1934) is an African-American inventor.

2383: what is the fundamental theorem of calculus used for — *Fundamental theorem of calculus*
The **fundamental theorem** of **calculus** is a **theorem** that links the concept of the derivative of a **function** with the concept of the **integral** .

2385: what is nicki minaj real name — *Nicki Minaj*
Onika Tanya Maraj (born December 8, 1982), known by her stage **name Nicki Minaj** (), is a Trinidadian-born American rapper, **singer**, songwriter and **television** personality.

2390: what is the disease osteonecrosis of the jaw? — *Osteonecrosis of the jaw*
Osteonecrosis of the **jaw** (ONJ) is a **severe** bone **disease** that affects the maxilla and the mandible .

2395: who is suicide tna — *Suicide (character)*
Suicide is a fictional character from **TNA** Impact!

2397: where is hickory located nc — *Hickory, North Carolina*
Hickory is the principal city in the Hickory-Lenoir-Morganton MSA , in which the population at the 2010 Census was 365,497.
Hickory is a city in Catawba County , with parts also in Burke County and Caldwell County .

2398: what is Polyester in packaging PET — *Polyethylene terephthalate*
Polyethylene terephthalate (sometimes written poly(**ethylene** terephthalate)), commonly abbreviated **PET**, **PETE**, or the obsolete PETP or PET-P, is a thermoplastic **polymer resin** of the **polyester** family and is used in synthetic **fibers** ; beverage, food and other liquid containers ; thermoforming applications; and engineering **resins** often in combination with glass **fiber**.

2403: who is director jack smight — *Jack Smight*
Jack Smight (March 9, 1925 - September 1, 2003) was an American theatre and film **director** .

2404: what is dmitri mendeleev 4 — *Dmitri Mendeleev*
Dmitri Ivanovich **Mendeleev** (; 8 February 1834 – 2 February 1907 O.S. 27 January 1834 – 20 January 1907) was a Russian chemist and inventor.

2405: where is modesto california — *Modesto, California*
Modesto (Spanish for "**modest**"), officially the **City** of **Modesto**, is the county seat and largest **city** of Stanislaus County , **California** , United States .

2407: what is the female symbol called — *Gender symbol*
The two standard gender **symbols** denoting male ♂ and **female** ♀ are derived from astrological **symbols** , denoting the classical planets Mars and Venus , respectively.

2409: what is leasehold; real estate — *Leasehold estate*
A **leasehold estate** is an ownership of a temporary **right** to **hold** land or **property** in which a lessee or a tenant **holds rights** of **real property** by some form of title from a lessor or landlord .

♯2425: when was Mary Anderson born — *Mary Anderson (inventor)*
Mary Anderson (1866–1953) was a real estate developer, rancher, viticulturist and inventor of the windshield wiper blade.

2426: WHO WAS IN THE MOVIE I CONFESS WITH MONTGOMERY CLIFT — *I Confess (film)*
I **Confess** is a 1953 drama **film** directed by Alfred **Hitchcock** , and starring **Montgomery Clift** as Fr. Michael William Logan, a Catholic priest , Anne Baxter as Ruth Grandfort, and Karl Malden as Inspector Larrue.

2430: what is black hairy tongue disease — *Black hairy tongue*
Black hairy tongue (BHT, also termed lingua villosa nigra) refers to a condition of the **tongue** where the filiform papillae elongate with **black** or brown discoloration, giving a **black** and **hairy** appearance.

2434: what is ethernet gigabit — *Gigabit Ethernet*
In computer networking , **gigabit Ethernet** (GbE or 1 GigE) is a term describing various technologies for transmitting **Ethernet** frames at a rate of a **gigabit** per second (1,000,000,000 bits per second), as defined by the IEEE 802.3-2008 standard.

2438: where is the tibia location — *Tibia*
The **tibia** is named for the Greek aulos flute, also known as a **tibia**.
The **tibia** (), shinbone or shankbone, is the larger and stronger of the two bones in the leg below the knee in vertebrates (the other being the fibula), and connects the knee with the ankle bones .
The **tibia** is found next to the fibula on the medial side of the leg, closer to center-line.
The **tibia** is connected to the fibula by an interosseous membrane , forming a type of joint called a syndesmosis with very little movement.

2442: where is bloemfontein in africa — *Bloemfontein*
Bloemfontein (, ; Dutch for "fountain of flowers") is the capital city of the Free State Province of South **Africa** ; and, as the judicial capital of the nation, one of South **Africa's** three national capitals – the other two being Cape Town , the legislative capital, and Pretoria , the administrative capital.

2447: What is the prostate for? — *Prostate*
The **prostate** (from Greek προστάτης – **prostates**, literally "one who stands before", "protector", "guardian") is a compound tubuloalveolar exocrine **gland** of the male reproductive system in most mammals .

2450: what is in the preamble of the constitution — *Preamble to the United States Constitution*
The **Preamble** to the United States **Constitution** is a brief introductory statement of the **Constitution's** fundamental purposes and guiding principles.

2462: where is the sermon on the mount in the bible — *Sermon on the Mount*
To most believers in Jesus, the **Sermon** on the **Mount** contains the central tenets of Christian discipleship .
The **Sermon** on the **Mount** (anglicized from the Matthean Vulgate Latin section title: Sermo in monte) is a collection of sayings and teachings of Jesus , which emphasizes his moral teaching found in the Gospel of Matthew (chapters 5 , 6 and 7).
Sermon on the **Mount** by Carl Bloch

2468: what is article 5 about — *Article Five of the United States Constitution*
Article Five of the United States Constitution describes the process whereby the Constitution may be altered.

2471: what is the political tea party — *Tea Party movement*
The **Tea Party** movement is an American **political** movement that advocates strict adherence to the United States Constitution , reducing U.S. government spending and taxes , and reduction of the U.S. **national** debt and federal budget deficit .

2474: What is the function of albumin? — *Human serum albumin*
Albumin transports hormones, fatty acids, and other compounds, buffers pH, and maintains osmotic pressure , among other **functions**.

2476: who is ezekiel in the bible — *Book of Ezekiel*
The Book of **Ezekiel** is the third of the Latter Prophets in the Hebrew **Bible** , following the books of Isaiah and Jeremiah and preceding the Book of the Twelve .

2483: where are the ten commandments in the bible — *Ten Commandments*
The **Ten Commandments** appear twice in the Hebrew **Bible** , in the books of **Exodus** and Deuteronomy .

2504: where is loxahatchee fl at — *Loxahatchee, Florida*
Loxahatchee is located within the Indian Trails Improvement District and the **Loxahatchee** Groves Water Control District.
In 2006, the state of Florida approved incorporation of part of **Loxahatchee** as the Town of **Loxahatchee** Groves to become the 38th municipality in Palm Beach County.
The name **Loxahatchee** means " turtle river."
Loxahatchee is also the name of the Post Office that serves this area, which is under the zip code of 33470.
Loxahatchee is a community located in Palm Beach County , Florida , United States .
Lion Country Safari , a drive-through safari park , is located in **Loxahatchee**.

2506: what is chicken chow mein — *Chow mein*
Chow mein (; **Chinese** : 炒麵, "**fried** noodles") is a **Chinese** term for a dish of stir-fried noodles , of which there are many varieties.

2511: what is the oath of allegiance of the united states — *Oath of Allegiance (United States)*
The **United States Oath** of **Allegiance** (officially referred to as the "**Oath** of Allegiance," 8 C.F.R. Part 337 (2008)) is an **oath** that must be taken by all immigrants who wish to become **United States** citizens .

2512: what is that dot on the forehead of indian women
hfill *Bindi (decoration)*
A bindi (, from Sanskrit bindu, meaning "a drop, small particle, **dot**"), or a pottu/bottu (in Tamil, Telugu, and Malayalam) is a **forehead** decoration worn in South Asia (particularly India , Bangladesh , Nepal , Sri Lanka and Mauritius) and Southeast Asia .

2524: where is Desmanthus Illinoensis found — *Desmanthus illinoensis*
Desmanthus illinoensis (commonly known as Illinois bundleflower, prairie-mimosa or prickleweed) is a plant in many areas of the south central US.

2531: what is the origin of fiona — *Fiona*
The name **Fiona** has been particularly popular for girls of Asian or Pacific Islander descent in New York City , where **Fiona** ranked as the ninth most popular name for girls from those groups in 2007.
There is also a modern tendency to equate the authentic Scottish Gaelic feminine name Fionnghal with **Fiona**.
The name was subsequently used as a pseudonym by William Sharp (1855–1905), who authored several romantic works under the name "**Fiona** Macleod".
The name **Fiona** was invented, and first used, by the Scottish poet James Macpherson (1736–96), author of the Ossian poems, which he claimed were translations from ancient Gaelic sources (sources, when challenged, he never produced).
The name **Fiona** is rising in popularity particularly in German -speaking countries.
The name **Fiona** is also sometimes used as an Anglicisation of the Irish-language name Fiona, although these names are otherwise unrelated (as Fiona is derived from an element meaning "vine").
Fiona was tied for third place in the ranking of most popular names for baby girls born in the tiny principality of Liechtenstein in 2008.
Fiona was the 49th most popular name for baby girls born in 2008 in Germany .
Fiona is a feminine given name .

2533: what is auto quarter panel — *Quarter panel*
On an **automobile**, a **quarter panel** (or rear wing as it is known in British English) is usually considered the body **panel** which goes between the rear door and the trunk and typically wraps around the wheel well; however, the similar front section Fender (vehicle) , between the door and the hood, is sometimes incorrectly referred to as a **quarter panel**.

♯2536: when was pearl harbor bombed — *Attack on Pearl Harbor*
The attack on **Pearl Harbor** (called Hawaii Operation or Operation AI by the Japanese Imperial General Headquarters (Operation Z in planning) and the Battle of **Pearl Harbor**) was a surprise military strike conducted by the Imperial Japanese

Navy against the United States naval base at **Pearl Harbor** , Hawaii, on the morning of December 7, 1941 (December 8 in Japan).

2537: what is the la tour de france — *Tour de France*
The **Tour de France** () is an annual multiple stage bicycle race primarily held in **France** , while also occasionally making passes through nearby countries.

2540: what is associates arts degree — *Associate degree*
An **associate degree** is an undergraduate academic **degree** awarded by community colleges , junior colleges , technical colleges, and bachelor's **degree** -granting colleges and universities upon completion of a course of study usually lasting two years.

2543: what is vat tax? — *Value added tax*
A **VAT** is like a sales **tax** in that ultimately only the end **consumer** is **taxed**.
A value added **tax** (**VAT**) is a form of **consumption tax** .

2544: what is high sticking in hockey — *High-Sticking*
-sticking is the name of two infractions in the sport of ice **hockey** that may occur when a player intentionally or inadvertently plays with his or her **stick** above the height of the shoulders or above the cross bar of a **hockey** goal.

#2553: when was the trogan war — *Trojan War*
The ancient **Greeks thought** that the Trojan **War** was a **historical event** that had taken **place** in the 13th or 12th **century BC**, and **believed** that Troy was **located** in modern-day Turkey near the Dardanelles .

2556: where is the palatine canal — *Greater palatine canal*
The greater **palatine canal** (or pterygopalatine **canal**) is a passage in the skull that transmits the greater **palatine** artery , vein, and nerve between the pterygopalatine fossa and the oral cavity .

2557: what is the lowest temperature ever recorded in antarctica — *Climate of Antarctica*
Antarctica has the **lowest** naturally occurring **temperature** ever **recorded** on the ground on Earth: −89.2 °C (−128.6 °F) at Vostok Station .

2560: what is oregon institute of technology like — *Oregon Institute of Technology*
Oregon Institute of **Technology**, also known as **Oregon** Tech or OIT, is one of seven Universities in the **Oregon** University System , and the only public **institute** of **technology** in the Pacific Northwest .

#2562: When Is Passover Over — *Passover*
In Judaism , a **day** commences at dusk and lasts until the following dusk, thus the first **day** of **Passover** only begins after dusk of the 14th of Nisan and ends at dusk of the 15th **day** of the month of Nisan.

2580: WHERE IS ROUGH AND READY, CA — *Rough and Ready, California*
Rough and **Ready** is a census-designated place in Nevada County, California , United States .

2584: where is rashard lewis from??? — *Rashard Lewis*
Rashard Quovon **Lewis** (born August 8, 1979 in Pineville, Louisiana) is an American professional basketball player who currently plays for the Miami Heat of the NBA .

2590: what is doxycycline hyclate used for — *Doxycycline*
Doxycycline is a member of the tetracycline antibiotics group, and is commonly **used** to treat a variety of infections .

2592: what is sodium hypochlorite solution — *Sodium hypochlorite*
Sodium hypochlorite solution, commonly known as bleach or liquid bleach, is frequently used as a disinfectant or a bleaching agent.

#2597: when was jacques cousteau born — *Jacques Cousteau*
Jacques-Yves **Cousteau** (; commonly known in English as **Jacques Cousteau**; 11 **June** 1910 – 25 **June** 1997) was a French naval officer, explorer , **conservationist** , filmmaker, innovator, scientist, photographer, author and researcher who studied the sea and all forms of life in water.

2607: where was the fugitive slave law made — *Fugitive Slave Act of 1850*
The **Fugitive Slave Law** or **Fugitive Slave** Act was passed by the United States Congress on September 18, 1850, as part of the Compromise of 1850 between Southern **slave-holding** interests and Northern Free-Soilers .

2612: what is eic earned income credit — *Earned Income Tax Credit*
The United States **federal earned income** tax **credit** or **earned income credit** (**EITC** or **EIC**) is a refundable tax **credit** for low- and medium-income individuals and couples, primarily for those who have qualifying children.

2620: where is valley village ca — *Valley Village, Los Angeles*
Village is a district in the San Fernando **Valley** region of Los Angeles, California .
Valley Village ParkValley

2630: What is sign of cancer — *Cancer (astrology)*
Cancer (♋) is an astrological **sign** , which is associated with the constellation **Cancer** .

2643: what is a gsm cell phone — *GSM*
GSM (Global System for Mobile Communications, originally '), is a **standard** set developed by the European Telecommunications **Standards** Institute (ETSI) to describe protocols for second generation (2G) digital cellular networks used by mobile **phones** .

2644: what was the cash and carry lend lease — *Lend-Lease*
The **Lend Lease** act was an act where the United States had supported its allies.

2646: what is el morro in puerto rico — *Castillo San Felipe del Morro*
Castillo San **Felipe del Morro** also known as Fort San **Felipe del Morro** or **Morro** Castle, is a 16th-century citadel located in San Juan, **Puerto Rico** .

2651: where are poison dart frog seen — *Poison dart frog*
Poison dart frog (also dart-poison **frog**, **poison frog** or formerly **poison** arrow **frog**) is the common name of a group of **frogs** in the family Dendrobatidae which are native to Central and South America .

2659: where were the Winter Olympics in 2006 — *2006 Winter Olympics*
The **2006 Winter Olympics**, **officially known** as the XX **Olympic Winter** Games, was a **winter** multi-sport event which was celebrated in Turin , Italy from February 10, **2006**, through February 26, **2006**.

2671: WHAT IS NON BINDING? — *Non-binding arbitration*
Subsequent to a non-binding arbitration, the parties remain free to pursue their claims either through the courts, or by way of a **binding** arbitration, although in practice a settlement is the most common outcome.
Non-binding arbitration is a type of arbitration in which the arbitrator makes a determination of the rights of the parties to the dispute, but this determination is not **binding** upon them, and no enforceable arbitration award is issued.

2675: who was mr big on sex and the city — *Mr. Big (Sex and the City)*
John James "**Mr**. **Big**" Preston is a recurring fictional **character** in the HBO **series Sex** and the **City** , portrayed by Chris Noth .

2682: what is the actresses name that played in walk that line? — *Walk the Line*
The **film** was nominated for five Academy Awards including **Best Actor** (Joaquin Phoenix), **Best Actress** (Reese Witherspoon, which she won), and **Best** Costume Design (Arianne Phillips).

2686: where is university of nelson mandela metropolitan located — *Nelson Mandela Metropolitan University*
Nelson Mandela Metropolitan University (NMMU) is a South African tertiary education institution with its main administration in the coastal city of Port Elizabeth .

#2687: when was How the west was won filmed? — *How the West Was Won (film)*
How the **West** Was **Won** is a 1962 American epic - **Western film**.

#2703: when was the great fire in chicago — *Great Chicago Fire*
The **Great Chicago Fire** was a conflagration that burned from Sunday, **October** 8, to early Tuesday, **October** 10, 1871, killing hundreds and destroying about in **Chicago** , Illinois .

2717: what is pci Interface — *Conventional PCI*
Conventional PCI (**PCI** is an initialism formed from Peripheral Component Interconnect, part of the **PCI** Local Bus **standard** and often shortened to just **PCI**) is a local **computer** bus for attaching hardware devices in a **computer** .

2718: what is jagger bombs — *Jägerbomb*
The **Jägerbomb** () is a bomb shot cocktail that was originally mixed by dropping a shot of Jägermeister into a glass of beer and in recent years evolved by the Bagheri brothers (UK) with Red Bull or other energy drinks.
A long drink mixed with Jägermeister and Red Bull is called "**JägerBull**" as it is adopted from Jägermeister and RedBull.

2724: what is the name of mountains along california — *California Coast Ranges*
The other three coastal **California mountain** ranges are the Transverse Ranges , Peninsular Ranges and the Klamath **Mountains** .

2733: who is the author of tree grows in brooklyn — *A Tree Grows in Brooklyn (novel)*
A **Tree Grows** in **Brooklyn** is a 1943 **novel** written by Betty Smith .

2734: where is al jazeera based — *Al Jazeera*
Al Jazeera (' , literally "The Island", abbreviating " The [Arabian] Peninsula "), also known as Aljazeera and JSC (**Jazeera Satellite** Channel), is a broadcaster owned by the privately held **Al Jazeera Media Network** and headquartered in Doha , Qatar .

2756: what is the location of coldwater ms — *Coldwater, Mississippi*
Coldwater is a small town in Tate County , Mississippi .

2757: what is the function of the hard palate — *Hard palate*
The **hard palate** is a thin horizontal bony plate of the skull , located in the roof of the mouth.
Also on the anterior portion of the roof of the **hard palate** is the Rugae which are the irregular ridges in the mucous membrane that help facilitate the movement of food backwards towards the pharynx.

2761: where is Chayanne from? — *Chayanne*
Elmer Figueroa Arce (born June 28, 1968), best known under the stage name **Chayanne**, is a Puerto Rican Latin pop singer and actor.
As a solo artist, **Chayanne** has released 21 solo albums and sold over 15 million albums worldwide.

2768: what is eggnog made of — *Eggnog*
Eggnog, or egg nog, is a sweetened dairy -based beverage traditionally **made** with milk and/or cream , sugar , and whipped eggs (which gives it a frothy texture).

2773: what is gravy made of — *Gravy*
The **gravy** may be further colored and flavored with **gravy** salt (a simple mix of salt and caramel food colouring) or **gravy** browning (**gravy** salt dissolved in water) or ready-made cubes and powders can be used as a substitute for natural meat or **vegetable** extracts.

2786: what is cta used for — *Computed tomography angiography*
Computed tomography angiography (**CTA**) is a computed tomography technique **used** to visualize arterial and venous vessels throughout the body.

2795: what is vitamin b12 used for — *Vitamin B12*
Vitamin B12, **vitamin B12** or **vitamin** B-12, also **called** cobalamin, is a water-soluble **vitamin** with a key role in the normal functioning of the brain and nervous system , and for the formation of blood .

2797: What is the the North American Free Trade Agreement? — *North American Free Trade Agreement*
The **North American Free Trade Agreement** (NAFTA) is an **agreement** signed by Canada , Mexico , and the United States , creating a trilateral **trade** bloc in **North America** .

2798: what is tmz stand for — *TMZ (website)*
The name **TMZ stands** for the historic " studio zone " or 30-mile zone radius from the intersection of West Beverly Boulevard and North La Cienega Boulevard in Los Angeles.

2800: what is impingement of the shoulder — *Impingement syndrome*
Shoulder impingement syndrome, also called painful arc syndrome, supraspinatus syndrome, swimmer's **shoulder**, and thrower's **shoulder**, is a clinical syndrome which occurs when the tendons of the rotator cuff muscles become irritated and inflamed as they pass through the subacromial space, the passage beneath the acromion .

2818: who was the first one to invent medicine — *History of medicine*
The ancient Egyptians had a system of **medicine** that was very advanced for its time and influenced **later** medical **traditions**.

2822: Who was Daniel J Daly? — *Daniel Daly*
Sergeant Major **Daniel** Joseph "**Dan**" **Daly** (November 11, 1873 – April 27, 1937) was a United States Marine and one of only nineteen men (including seven Marines) to have received the Medal of Honor twice .

2829: what is the plot of the shawshank redemption? — *The Shawshank Redemption*
Adapted from the Stephen **King** novella Rita Hayworth and **Shawshank Redemption** , the film tells the story of Andy Dufresne, a banker who spends nearly two decades in **Shawshank** State Prison for the murder of his wife and her lover **despite** his claims of innocence.

2830: what is an information technology manager — *Information technology management*
Managing this responsibility within a company entails many of the basic management functions, like budgeting , staffing, and organizing and controlling, along with other aspects that are unique to **technology**, like change management, software design , network planning, tech support etc.
IT management is the discipline whereby all of the **technology** resources of a firm are managed in accordance with its needs and priorities.

2832: what is the scientific name of a cardinal bird — *Northern Cardinal*
The Northern **Cardinal** (**Cardinalis cardinalis**) is a North American **bird** in the genus **Cardinalis** ; it is also known colloquially as the redbird or common **cardinal**.

2838: where are kenworth trucks built — *Kenworth*
Kenworth is an American manufacturer of medium and heavy-duty Class 8 **trucks** based in Kirkland, **Washington** , **United States** , a suburb of Seattle, **Washington**.

#2844: when is the feast of St. Rita — *Rita of Cascia*
The Roman Catholic **Church** , under the pontificate of Pope Leo XIII officially **canonized Rita** on May 24, 1900, while her **feast** day is celebrated every May 22.

2846: where is j.cole from — *J. Cole*
Jermaine Lamarr **Cole** (born January 28, 1985), better known by his stage name **J**. **Cole**, is an American hip-hop recording artist and record producer from Fayetteville, North Carolina .

2856: what is the gdp for greenland 2010? — *Economy of Greenland*
GDP per capita is similar to the average European economies but the economy is critically dependent upon substantial support from the Danish government, which supplies about half the revenues of the home rule government who in turn employ about 8,000 **Greenlanders** out of a labor force of 40,156 (Jan. 2012).

2857: What is hydrogen in — *Hydrogen*
Most of the **hydrogen** on Earth is in molecules such as water and organic compounds because **hydrogen** readily forms covalent compounds with most non-metallic elements.
In 1766–81, Henry Cavendish was the first to recognize that **hydrogen** gas was a discrete substance, and that it produces water when burned, a property which later gave it its name: in Greek, **hydrogen** means "water-former".

2858: where is UWA on world list of universities — *University of Western Australia*
One of **Australia's** best and most prestigious **universities**, **UWA** is highly ranked **internationally** in various **publications**; the 2011 QS **World University** Rankings **placed UWA** at 73rd **internationally**.

#2878: when was purple haze by jimi hendrix made? — *Purple haze*
"**Purple Haze**" is a song written and **recorded** by **Jimi Hendrix** in 1967, released as the second single by The **Jimi Hendrix** Experience in both the United Kingdom and the United States .

2880: where was martin luther born — *Martin Luther*
Martin Luther (; 10 November 1483 – 18 **February 1546**) was a German monk , former Catholic priest , **professor** of theology and seminal figure of a reform movement in sixteenth century Christianity , **subsequently** known as the **Protestant** Reformation .

2891: what is the mortality rate of sepsis — *Septic shock*
The **mortality rate** from **septic** shock is approximately 25–50%.

2893: where is kos from? — *K-os*
The alias "**k-os**", spelled with a lower case "k", was intended to be less aggressive than the pseudonyms of other rappers whose names were all upper case, such as KRS-One .
k-os usually performs with a live band, something that is uncommon in the hip hop genre.
k-os received his first musical exposure with the single "Musical Essence", released in 1993.
Kevin Brereton (born February 20, 1972), better known by his stage name **k-os** (; "chaos"), is a Canadian rapper , singer , songwriter and record producer .
It is an acronym for " Knowledge of Self ," although in a later interview he said that it originally stood for "Kevin's Original Sound." **k-os'** music incorporates a wide variety of music genres, including rap , funk , rock , and reggae .
A musician as well as a producer, **k-os** has written and produced nearly every part of all four of his albums.

2904: what is the sign called? — *At sign*
The at **sign** is also **commonly called** the at symbol, ampersat, apetail or commercial at in **English**—and less **commonly** a wide range of other terms.

2905: who is the book the catcher in the rye by? — *The Catcher in the Rye*
The **Catcher** in the **Rye** is a 1951 novel by J. D. **Salinger** .

2910: what is prince williams last name — *Prince William, Duke of Cambridge*
Prince William, **Duke** of **Cambridge** (**William** Arthur Philip Louis; **born** 21 June 1982), is the elder son of Charles, **Prince** of Wales , and Diana, **Princess** of Wales , and third-eldest grandchild of **Queen Elizabeth** II and **Prince** Philip, **Duke** of Edinburgh .

2914: what is IBRIX — *IBRIX Fusion*
The software was produced, sold, and supported by **IBRIX** Incorporated of Billerica, Massachusetts .
Subsequent to the acquisition, the software components of **IBRIX** have been combined with ProLiant servers to form the X9000 series of storage systems.
IBRIX Fusion is a scalable parallel file system combined with integrated logical volume manager , availability features and a management interface.
HP announced on July 17, 2009 that it had reached a definitive agreement to acquire **IBRIX**.

2915: what is bilirubin total — *Bilirubin*
Bilirubin is excreted in bile and urine , and elevated levels may indicate certain diseases.
Bilirubin (formerly referred to as hematoidin) is the yellow breakdown product of normal heme catabolism .

2918: what is quasi judicial agency — *Quasi-judicial body*

A quasi-judicial body is an entity such as an arbitrator or tribunal board, generally of a public administrative **agency**, which has powers and procedures resembling those of a court of law or judge , and which is obligated to objectively determine facts and draw conclusions from them so as to provide the basis of an official action.

2924: who is the founder of twitter *Twitter*
Twitter was created in March 2006 by Jack Dorsey and by July, the social networking site was launched.
Twitter is an online social networking service and microblogging service that enables its users to send and read text-based messages of up to 140 characters , known as "tweets".
Twitter Inc. is based in San Francisco , with additional servers and offices in New York City , Boston , and San Antonio . Since its launch, **Twitter** has become one of the ten most visited websites on the Internet, and has been described as "the SMS of the Internet ."

2932: who is bj's wholesale club *BJ's Wholesale Club*
BJ's Wholesale Club, Inc., commonly referred to simply as **BJ's**, is a membership-only warehouse **club** chain operating on the United States East Coast , as well as in the state of Ohio .

2936: what is the name of the wizard of oz *Wizard of Oz (character)*
The **Wizard** of **Oz**, known during his reign as The Great and Powerful **Oz**, is the epithet of Oscar Zoroaster Phadrig Isaac Norman Henkel Emmannuel Ambroise Diggs, a fictional character in the Land of **Oz** , created by American author L. Frank Baum .

2940: what is homebrew for wii *Wii homebrew*
Wii homebrew refers to the reuse of Nintendo 's **Wii** game console to run software that has not been authorized by Nintendo .
In more general terms, **Wii Homebrew** refers to the use of the **Wii** 's hardware, accessories and software for purposes outside those intended by the manufacturer.

2952: what is korean money called *South Korean won*
The won () (sign : ₩; code : KRW) is the currency of **South** Korea .

2955: what is three phase electrical *Three-phase electric power*
Three-phase electric power is a common **method** of alternating-current **electric power** generation , transmission , and distribution .

2957: What is the significance of Good Friday? *Good Friday*
It is also known as **Holy Friday**, Great **Friday**, Black **Friday**, or **Easter Friday**, though the latter properly **refers** to the **Friday** in **Easter** week .

2959: where is the great basin located on a us map *Great Basin*
It is noted for both its arid conditions and its **Basin** and **range** topography that varies from the North American low point at Badwater **Basin** to the highest point of the contiguous United States , less than away at the summit of Mount Whitney .

2966: what was the date of pearl harbor *Attack on Pearl Harbor*
The attack on **Pearl Harbor** (called Hawaii Operation or Operation AI by the Japanese Imperial General Headquarters (Operation Z in planning) and the Battle of **Pearl Harbor**) was a surprise military strike conducted by the Imperial Japanese Navy against the **United** States naval base at **Pearl Harbor** , Hawaii, on the morning of December 7, **1941** (December 8 in Japan).

2969: what is disney's magic kingdom *Magic Kingdom*
Magic Kingdom Park, also known as **Magic Kingdom**, is the first of four theme **parks** built at the Walt **Disney** World **Resort** in Bay Lake , **Florida** .

2972: what is the kanji for language *Kanji*
Kanji (;) are the adopted logographic Chinese characters (hanzi) that are used in the modern Japanese **writing** system along with hiragana (ひらがな,), katakana (カタカナ,), Hindu-Arabic numerals , and the occasional **use** of the Latin alphabet .

2973: what is in fruitcake *Fruit cake*
Fruit cake (or **fruitcake**) is a cake made with chopped candied fruit and/or dried fruit , nuts , and spices , and (optionally) soaked in spirits .

2978: who was the congressman who was caught with an escort in ny *Eliot Spitzer prostitution scandal*
On March 10, 2008, The New York Times reported that New York Governor Eliot Spitzer had patronized an elite **escort** service run by Emperors Club VIP .

2982: what is lockton affinity *Lockton Companies*
Lockton Affinity: **Lockton Affinity**, an affiliate of **Lockton** Companies, meets the insurance needs of **affinity** groups, franchises, professional **organizations**, and associations of all sizes.

2991: what is the name of chris cornell's band? *Chris Cornell*
Chris Cornell (born Christopher John Boyle; July 20, 1964) is an American rock **musician** best **known** as the lead vocalist and rhythm **guitarist** for Soundgarden and as the **former** lead vocalist for Audioslave .

2994: what is preciosa crystal? *Preciosa (corporation)*
Preciosa is the luxury brand name for the range of precision-cut lead **crystal** glass and related products produced by **Preciosa** a.s. of Jablonec nad Nisou , Czech Republic .

2998: where is the chupacabra found *Chupacabra*
The **Chupacabra** or **Chupacabras** (, from chupar "to suck" and cabra "goat", literally "goat sucker") is a legendary cryptid rumored to inhabit parts of the Americas .

2999: What is the purpose of North American Free Trade Agreement *North American Free Trade Agreement*
The **North American Free Trade Agreement** (NAFTA) is an **agreement** signed by Canada , Mexico , and the United States , creating a trilateral **trade** bloc in **North America** .

3004: who are all of the jonas brothers *Jonas Brothers*
Formed in 2005, they have gained popularity from the **Disney Channel** children's television network and consists of three **brothers** from Wyckoff, New Jersey ; Paul Kevin **Jonas** II , Joseph Adam **Jonas** and Nicholas Jerry **Jonas** .

3012: where is the brisket from *Brisket*
The **brisket** muscles include the superficial and deep pectorals.
The beef **brisket** is one of the nine beef prime cuts.
Brisket is a cut of meat from the breast or lower chest of beef or veal .
American cuts of beef including the **brisket**
According to the Random House Dictionary of the English Language , Second Edition, the term derives from the Middle English **brusket** which comes from the earlier Old Norse , meaning cartilage .

3016: what is human chorionic *Human chorionic gonadotropin*
In molecular biology , **human chorionic** gonadotropin (hCG) is a hormone **produced** by the fertilized egg after conception .

3021: what is busiest airport in US *List of the busiest airports in the United States*
Hartsfield–Jackson Atlanta International **Airport** is the **busiest** single **airport** in the United States.

3027: where is osaka japan *Osaka*
is a city in the Kansai **region** of **Japan's** main island of Honshu , a designated city under the Local Autonomy Law , the capital city of **Osaka** Prefecture and also the largest part of the Keihanshin metropolis, which comprises three major cities of **Japan**, Kyoto , **Osaka** and Kobe .

3030: what is surname and given name *Given name*
A **given name**, in Western contexts often referred to as a **first name**, is a **personal name** that specifies and differentiates between members of a group of individuals, especially in a **family**, all of whose members usually share the same **family name** (surname).

3037: What is an economic feature? *Economics*
Other broad distinctions include those between positive **economics** (describing "what is") and normative **economics** (advocating "what ought to be"); between **economic** theory and applied **economics** ; between rational and behavioral **economics** ; and between **mainstream economics** (more "orthodox" and dealing with the "rationality-individualism-equilibrium nexus") and heterodox **economics** (more "radical" and dealing with the "institutions-history-social structure nexus").

3043: what is section eight housing *Section 8 (housing)*
Section 8 of the **Housing** Act of 1937 (), often simply known as **Section** 8, as repeatedly amended, authorizes the payment of rental **housing** assistance to private landlords on behalf of approximately 3.1 million low-income **households**.

Example 4.15.4. In Table S12 below, we list the questions where our top-scoring answers have all been officially labeled as incorrect by the WikiQA team, including 167 "quantitative" questions, out of 662 total. It should be noted that some questions in this table are simply too tricky to be handled by literal interpretations. For example, our top-scoring answer to Q1838 differs from the official one (shown in blue):

1838: who sang cool jerk *Cool Jerk*
In the feature film Home Alone 2: Lost in New York , Uncle Frank (Gerry Bamman) **sings** "**Cool Jerk**" in the shower.
"Cool Jerk" is a popular song written by Donald Storball and originally performed by The Capitols .

Since our algorithm fails to reinterpret "who sang" as "who originally performed", our answer is outshone by the official correct answer. Doubtlessly, this particular entry in Table S12 illustrates the limitations of our automated question comprehension.

Example 4.15.5. It is still arguable whether dozens of entries in Table S12 represent miscarriage of justice by the WikiQA team. In the cases selected below, we believe that our top-scoring answers are not any worse than the official ones (painted in blue):

♯129: how many stripes on the flag *Flag of the United States*
The national **flag** of the United States of America , often simply referred to as the American **flag**, consists of **thirteen** equal horizontal **stripes** of **red** (**top** and **bottom**) **alternating** with **white**, with a **blue** rectangle in the canton (referred to specifically as the "union") bearing fifty small, **white**, **five-pointed stars** arranged in nine offset horizontal rows of **six stars** (**top** and **bottom**) **alternating** with rows of **five stars**.
The 50 stars on the flag represent the 50 states of the United States of America and the 13 stripes represent the thirteen British colonies that declared independence from the Kingdom of Great Britain and became the first states in the Union.

214: what does salesforce.com do? *Salesforce.com*
Salesforce.com Inc. is a global enterprise software company **headquartered** in San Francisco , California .
Though best known for its customer relationship management (CRM) product, Salesforce has also expanded into the "social enterprise arena" through acquisitions.

♯557: how many spanish speaking people are there in the us *Spanish language in the United States*
There are 45 million Hispanophones who **speak Spanish** as a first or second **language**, as well as six million **Spanish language** students, composing the **largest** national Spanish-speaking community outside of **Mexico** and making **Spanish** the Romance **language** and the Indo-European **language** with **largest** number of **native speakers** in the world.
According to the 2010 American Community Survey conducted by the U.S. Census Bureau , Spanish is the primary language spoken at home by almost 37 million people aged five or older, a figure more than double that of 1990.

1831: where do you find iodine *Iodine*
Because of this function, radioisotopes of **iodine** are concentrated in the thyroid gland along with nonradioactive **iodine**.
Iodine is found on Earth mainly as the highly water-soluble iodide ion, I^-, which concentrates it in oceans and brine pools .

2463: who is E from entourage *Eric Murphy*
Murphy is a fictional **character** on the comedy-drama television series **Entourage** .
Eric "E."

2532: what is the official language of america? *Languages of the United States*
The situation is quite varied at the state and territorial levels, with some states mirroring the federal policy of **adopting** no **official language** in a **de** jure capacity, others **adopting English** alone, others **officially adopting English** as well as local **languages**, and still others **adopting** a policy of **de** facto bilingualism.
The most commonly used language is English .

2841: what is the type of democracy in which all citizens have the right to make government decisions *Democracy*
One **form** of **democracy** is **direct democracy** , in which all eligible **citizens** have **direct** and **active** participation in the **decision making** of the **government**.
Democracy is a form of government in which all eligible citizens have an equal say in the decisions that affect their lives.

In particular, the penultimate official answer listed above (accompanying Q2532) is logically irrelevant,[48] if not factually misleading; the last official answer listed above (accompanying Q2841) does not even address the "type of democracy" in the question (*i.e.* direct, representative *etc.*) at all. Despite our disagreement over certain judgements made by the WikiQA team, we are not going to alter their official labels for correct/incorrect answers, so as to ensure a fair comparison in [1, Fig. 7b] and Fig. S2.

Table S12. List of WikiQA misses

4: how a water pump works — *Pump*

4: how a water pump works — *Pump*

Pumps can be classified into three major groups according to the method they use to move the fluid: direct lift, displacement, and gravity **pumps**.

♯11: how big is bmc software in houston, tx — *BMC Software*

Headquartered in **Houston** , Texas , **BMC** develops, markets and sells **software** used for multiple functions, including IT service management, data center **automation**, **performance** management, virtualization lifecycle management and cloud computing management.

♯17: how much are the harry potter movies worth — *Harry Potter*

Also due to the **success** of the books and films, **Harry Potter** has been used for a theme **park**, The **Wizarding World** of **Harry Potter** in **Universal Parks** & Resorts 's Islands of Adventure .

♯20: how old was sue lyon when she made lolita — *Lolita (1962 film)*

The film stars **James Mason** as Humbert Humbert, **Sue Lyon** as Dolores Haze (**Lolita**) and Shelley Winters as Charlotte Haze with Peter Sellers as Clare Quilty.

21: how are cholera and typhus transmitted and prevented — *Cholera*

Cholera is an infection in the small intestine caused by the bacterium Vibrio **cholerae** .

28: how are aircraft radial engines built — *Radial engine*

The **radial** configuration was very commonly used in **aircraft engines** before turbine **engines** became predominant.

♯30: how deep can be drill for deep underwater — *Deepwater drilling*

There are basically two kinds of mobile **deepwater drilling** rigs: semi-submersible **drilling** rigs and **drillships** .

♯32: how long was frank sinatra famous — *Frank Sinatra*

Sinatra left Capitol to found his own record label, Reprise Records in 1961 (finding success with albums such as Ring-a-Ding-Ding! , **Sinatra** at the Sands and **Francis Albert Sinatra** & **Antonio** Carlos **Jobim**), toured internationally, was a founding member of the Rat Pack and fraternized with celebrities and statesmen, including John F. **Kennedy** .

♯42: how much is jk rowling worth — *J. K. Rowling*

Joanne "Jo" **Rowling** (), (born 31 July 1965), pen name J. **K. Rowling**, is a British novelist, best known as the author of the Harry Potter fantasy **series**.

♯43: how big is auburndale florida — *Auburndale, Florida*

Auburndale is a city in Polk County, **Florida** , United States .

♯46: how old is the singer bob seger — *Bob Seger*

As a locally successful Detroit-area artist, he performed and recorded as **Bob Seger** and the Last Heard and **Bob Seger** System throughout the 1960s.

♯48: how long was i love lucy on the air — *I Love Lucy*

I **Love Lucy** is an American **television sitcom** starring Lucille Ball , Desi Arnaz , Vivian Vance , and William **Frawley** .

♯59: HOW MUCH IS CENTAVOS IN MEXICO — *Mexican peso*

The name was originally used in reference to **pesos** oro (gold weights) or **pesos** plata (silver weights).

66: how did armando christian perez become famous — *Pitbull (entertainer)*

In 2004, he released his debut album titled M.I.A.M.I. (short for Money Is A Major Issue) under TVT Records .

♯75: how old were the twin towers when destroyed — *World Trade Center*

At the time of their completion, the original 1 World Trade Center (the North **Tower**) and 2 World Trade Center (the South Tower), known collectively as the "**Twin Towers**", were the **tallest buildings** in the world.

♯89: how many presidents of the us — *List of Presidents of the United States*

Upon the death, resignation, or removal from **office** of an incumbent **President**, the Vice **President** assumes the **office**.

♯100: what happens to the light independent reactions of photosynthesis? — *Light-independent reactions*

This **process happens** when **light** is available **independent** of the kind of **photosynthesis** (C3 carbon fixation , C4 carbon fixation , and Crassulacean Acid Metabolism); CAM plants store malic acid in their vacuoles every night and release it by day in order to make this **process** work.

107: what countries are under the buddhism religion — *Buddhism by country*

Government policies in these **countries** may encourage the under-reporting or non-reporting of **religious** adherence, resulting in official totals that may drastically underestimate the number of **religious** practitioners in these **countries**.

108: how did wild bill's father die — *Wild Bill Hickok*

James Butler Hickok (May 27, 1837 – August 2, 1876), better known as **Wild Bill** Hickok, was a folk hero of the American Old West .

♯109: how many land rovers have landed on mars — *Mars rover*

It currently manages the **Mars** Exploration **Rover** mission's active Opportunity **rover** and inactive Spirit , and, as part of the **Mars** Science Laboratory mission, the Curiosity **rover**.

112: what bird family is the owl — *Owl*

Owls hunt mostly small mammals , insects , and other **birds**, although a few species specialize in **hunting** fish .

113: how did harmon killebrew get strong — *Harmon Killebrew*

Harmon Clayton **Killebrew** (; June **29**, 1936 – May 17, 2011), nicknamed "**Killer**" and "Hammerin' **Harmon**", was an American professional baseball first baseman , third baseman , and left fielder .

116: What did the augurs use to interpret the will of the gods? — *Augur*

An **augur** holding a lituus , the curved wand often **used** as a symbol of **augury** on Roman coins

♯117: How much did Waterboy grossed — *The Waterboy*

The **Waterboy** is a 1998 American sports/comedy film directed by Frank Coraci (who played Robert 'Roberto' Boucher, Sr.), starring Adam **Sandler** , Kathy Bates , Fairuza Balk , Henry Winkler , Jerry Reed , Larry Gilliard, Jr. , Blake Clark , Peter Dante and Jonathan Loughran , and produced by Robert Simonds and Jack Giarraputo .

118: What county is Farmington Hills, MI in? — *Farmington Hills, Michigan*

Although the two cities have separate services and addresses, **Farmington** and **Farmington Hills** are often thought of as the same community.

121: what does a groundhog look for on groundhog day — *Groundhog Day*

Groundhog Day, already a widely recognized and popular tradition, received widespread attention as a result of the 1993 film **Groundhog Day** , which was set in Punxsutawney, PA (though filmed primarily in Woodstock, IL) and portrayed Punxsutawney Phil .

♯129: how many stripes on the flag — *Flag of the United States*

The national **flag** of the United States of America , often simply referred to as the American **flag**, consists of **thirteen** equal horizontal **stripes** of **red** (**top** and **bottom**) **alternating** with **white**, with a **blue** rectangle in the canton (referred to specifically as the "union") bearing fifty small, **white**, **five-pointed stars** arranged in nine offset horizontal rows of **six stars** (**top** and **bottom**) **alternating** with rows of **five stars**.

132: how is whooping cough distinguished from similar diseases — *Pertussis*

In some countries, this **disease** is called the 100 days' **cough** or **cough** of 100 days.

136: what county is galveston in texas — *Galveston County, Texas*

League City is the largest city in **Galveston County** in terms of population; URL_http between 2000 and 2005 it surpassed **Galveston** as the **county's** largest city.

137: what cities are in the bahamas — *List of cities in the Bahamas*

This is a list of **cities** in the **Bahamas** .

♯139: how many schools are in the big ten — *Big Ten Conference*

The **Big Ten Conference**, formerly Western **Conference** and **Big** Nine **Conference**, is the oldest **Division** I college athletic **conference** in the United States .

141: what composer used sound mass — *Sound mass*

In contrast to more traditional musical textures , **sound mass composition** "minimizes the importance of individual pitches in preference for texture , timbre , and dynamics as primary shapers of gesture and impact."

144: How Do You Play Spades — *Spades*

Its major difference as compared to other **Whist** variants is that, instead of **trump** being decided by the **highest** bidder or at random, the **Spade suit** is always **trump**, hence the name.

145: what continent is australia — *Australia (continent)*

The **continent** of **Australia** lies on a continental shelf overlain by shallow seas which divide it into several landmasses — the Arafura Sea and Torres Strait between **mainland Australia** and New Guinea, and Bass Strait between **mainland Australia** and Tasmania.

146: what causes heart disease — *Cardiovascular disease*

Cardiovascular **disease** refers to any **disease** that affects the cardiovascular system , principally **cardiac disease**, vascular **diseases** of the brain and kidney , and peripheral **arterial disease**.

147: what countries did immigrants come from during the immigration — *History of immigration to the United States*

During the nation's history , the growing **country** experienced successive waves of **immigration** which rose and fell over time, particularly from Europe, with the cost of transoceanic transportation sometimes paid by travelers becoming indentured servants after their arrival in the New World.

♯150: how many people live in atlanta georgia — *Atlanta*

Atlanta is the primary transportation hub of the Southeastern United States , via highway, railroad, and air , with Hartsfield–Jackson **Atlanta** International Airport being the world's **busiest** airport since 1998.

159: what does it mean to be a commonwealth state — *Commonwealth*

Most notably, the **Commonwealth** of Nations , an association primarily of former members of the British Empire , is often referred to as simply "the **Commonwealth**".

♯163: how long can you be in the supreme court — *Supreme Court of the United States*

It has ultimate (and largely discretionary) appellate **jurisdiction** over all **federal courts** and over **state court cases** involving issues of **federal law**, and original **jurisdiction** over a small range of **cases**.

165: what can sql 2005 do — *Microsoft SQL Server*

There are at least a dozen different **editions** of **Microsoft SQL** Server aimed at different audiences and for different workloads (ranging from small applications that store and retrieve **data** on the same computer, to millions of users and computers that access huge amounts of **data** from the Internet at the same time).

♯167: how many people were killed in the oklahoma city bombing — *Oklahoma City bombing*

The **Oklahoma City bombing** was a domestic terrorist **bomb** attack on the Alfred P. Murrah **Federal Building** in downtown **Oklahoma City** on April 19, 1995.

172: What group took home the award for best rock album at the Australian Recording Industry Association (ARIA) Music Awards? — *ARIA Music Awards*

The **Australian Recording Industry Association Music Awards** (commonly known as **ARIA Music Awards** or **ARIA Awards**) is an **annual** series of **awards** nights celebrating the **Australian music industry**, put on by the **Australian Recording Industry Association** (ARIA).

180: what did benedict arnold die — *Benedict arnold*

In the winter of 1782, **Arnold** moved to London with his second wife, **Margaret** "Peggy" **Shippen Arnold** .

♯185: how many ports are there in networking — *Port (computer networking)*

In the client-server model of application architecture, **ports** are used to provide a multiplexing service on each server-side **port** number that **network** clients connect to for service initiation, after which communication can be reestablished on other connection-specific **port** numbers.

195: how is canada's health care — *Comparison of the health care systems in Canada and the United States*

Total **government spending** per capita in the U.S. on **health care** was 23% **higher** than **Canadian government spending**, and U.S. **government expenditure** on **health care** was just under 83% of total **Canadian spending** (**public** and **private**) though these statistics don't take into account population **differences**.

♯197: how many xbox 360 games are there — *List of Xbox 360 games*

For a list of downloadable **Xbox** Live Arcade **games**, see the List of **Xbox** Live Arcade **games** .

207: what country did wine originate in — *History of wine*

Following the decline of Rome and its industrial-scale **wine** production for **export**, the Christian Church in medieval Europe also became a firm supporter of **wine**, necessary for celebration of the Catholic Mass. Whereas **wine** was forbidden in medieval Islamic cultures, its use in Christian libation was widely tolerated.

♯208: how many district are in west bengal — *List of districts of West Bengal*

Geographically, **West Bengal** is divided into a variety of **regions**— Darjeeling Himalayan hill **region** , Terai and Doors **region**, North **Bengal plains** , Rarh **region** , **Western** plateau and high lands , coastal **plains**, Sunderbans and the Ganges Delta .

♯210: how many towns in new hampshire — *List of cities and towns in New Hampshire*

However, **towns** currently are able to change their form of government by simple voter approval of a **new** municipal charter, with several of the more populous **towns** having already done so.

214: what does salesforce.com do? — *Salesforce.com*

Salesforce.com Inc. is a global enterprise software company **headquartered** in San Francisco , California .

217: what does 1 year mean to the planet — *Earth*

The Earth's **axis** of **rotation** is **tilted** 23.4° away from the perpendicular of its orbital **plane** , producing **seasonal** variations on the **planet's** surface with a **period** of one tropical **year** (365.24 **solar** days).

219: what is feedback mechanism in plants during respiration — *Cellular respiration*

While the overall **reaction** is a combustion **reaction**, no single **reaction** that comprises it is a combustion **reaction**.

220: how fire extinguisher works — *Fire extinguisher*

A **fire extinguisher**, flame **extinguisher**, or simply an **extinguisher**, is an active **fire** protection device used to **extinguish** or control small **fires**, often in emergency situations.

225: what countries are in central america? — *Central america*

Central America () is the **central** geographic region of the **Americas** .

227: how does a steam engine work — *Steam engine*

In general usage, the term **steam engine** can **refer** to either the integrated **steam** plants (including boilers etc.) such as railway **steam** locomotives and portable **engines** , or may **refer** to the piston or turbine machinery alone, as in the beam **engine** and stationary **steam engine** .

237: what county is tacoma washington in — *Tacoma, Washington*

Since the 1990s, developments in the downtown core include the **University** of **Washington Tacoma** ; **Tacoma** Link , the first modern electric light rail service in the **state**; the **state's highest** density of **art** and **history museums**; and a restored urban waterfront, the Thea Foss Waterway .

238: what county is chula vista ca in — *Chula Vista, California*

Located in the city is one of America's few year-round United States Olympic Training **centers** and popular tourist destinations include Cricket Wireless Amphitheater, the **Chula Vista** marina, and the **Chula Vista Nature Center**.

240: what did isaac newton do — *Isaac Newton*

In addition to his work on the calculus, as a **mathematician Newton** contributed to the study of power series , generalised the binomial theorem to non-integer exponents, and developed **Newton's** method for approximating the roots of a function .

♯244: how many US soldiers killed in Iraq war — *Casualties of the Iraq War*

Casualties of the conflict in **Iraq** since **2003** (beginning with the **2003 invasion** of **Iraq** , and continuing with the ensuing occupation of **Iraq** , as well as the activities of the various **armed** groups operating in the country) have come in many forms, and the accuracy of the information available on different types of **Iraq War casualties** varies greatly.

[48] Think about this: Punjabi is the most commonly used language in Pakistan, whose official languages are English and Urdu.

#255: how many professional hockey teams in canada — *National Hockey League*
The **National Hockey League** (**NHL**;) is an "unincorporated not-for-profit **association**" which operates a **major professional ice hockey league** of 30 franchised **member clubs**, of which **seven** are **currently located** in **Canada** and 23 in the **United States**.

#267: how many games did brett favre start in a row — *Brett Favre*
Favre became the Packers' **starting** quarterback in the fourth **game** of the **1992 season** , stepping in for injured quarterback Don **Majkowski** , and **started** every **game** through the 2007 **season** .

273: what does a vote to table a motion mean? — *Table (parliamentary procedure)*
In the rest of the English-speaking **world** such as the United Kingdom , to **table means** to move to **place** [the topic] upon the **table** (or to move to **place** on the table): a **proposal** to begin consideration (or reconsideration) of a **proposal**.

276: What does the class mean for SDHC cards? — *Secure Digital*
Host devices that comply with newer **versions** of the specification provide backward **compatibility** and accept older SD **cards**, but this article explains several factors that can prevent the use of a newer SD **card**:

#295: how many apple store are there in total? — *Apple Store*
The **Apple Store** is a chain of retail **stores** owned and operated by **Apple** Inc. , dealing in computers and consumer electronics.

300: how is jerky made — *Jerky*
Some makers still use just salt and sun-dry fresh sliced meat to **make jerky**.

#305: what happened on the moon during the period of Late Heavy Bombardment? — *Late Heavy Bombardment*
The **Late Heavy Bombardment** (commonly referred to as the **lunar** cataclysm, or **LHB**) is a **hypothetical** event around 4.1 to 3.8 billion years ago (Ga).

307: How did Edgar Allan Poe die? — *Edgar allan poe*
Edgar Allan Poe (born **Edgar Poe**; January 19, 1809 – October 7, 1849) was an American author, poet, editor and literary critic, considered part of the American Romantic Movement .

#311: how many seasons heroes — *Heroes (TV series)*
The second **season** of **Heroes** attracted an average of 13.1 **million viewers** in the U.S., and marked NBC's sole series among the top 20 ranked programs in total **viewership** for the 2007–**2008 season**.

#318: how old r Dylan and Cole Sprouse — *Dylan and Cole Sprouse*
They are twins and are collectively referred to as **Dylan** and **Cole Sprouse** or the **Sprouse** brothers, usually abbreviated as **Sprouse** Bros.

#321: how many days does the chinese new year last — *Chinese New Year*
Chinese New Year

324: what does s.h.i.e.l.d stand for — *S.H.I.E.L.D.*
S.H.I.E.L.**D**. is a fictional **espionage** and law-enforcement agency in the **Marvel Comics** Universe .

#326: how many consoles has xbox 360 sold — *Xbox 360*
Several major features of the **Xbox 360** are its integrated **Xbox** Live **service** that allows players to compete **online** ; **download** arcade **games**, **game** demos, trailers, TV shows, music and movies; and its Windows Media Center multimedia capabilities.

328: WHAT COUNTRY IS MEXICO IN — *Mexico*
Mexico ranks sixth in the **world** and first in the Americas by **number** of UNESCO **World** Heritage Sites with 31 , and in 2007 was the **tenth** most **visited country** in the **world** with 21.4 million international arrivals per year.

#330: how is slugging percentage calculated — *Slugging percentage*
The next year he **slugged** .846, and these records went unbroken until 2001 , when Barry Bonds achieved 411 bases in 476 at-bats, bringing his **slugging percentage** to .863, unmatched since.

335: what does the federal reserve do — *Federal Reserve System*
The **Federal Reserve** System's structure is composed of the presidentially appointed Board of **Governors** (or **Federal Reserve** Board), the **Federal Open** Market **Committee** (FOMC), twelve **regional Federal Reserve** Banks located in major cities throughout the nation, numerous **privately owned** U.S. **member** banks and various advisory councils.

#340: how many books are included in the protestant Bible? — *Books of the Bible*
The first part of Christian **Bibles** is the Old Testament , which contains, at minimum, the twenty-four **books** of the Hebrew **Bible** divided into thirty-nine **books** and **ordered** differently from the Hebrew **Bible**.

#341: how many stars on the first american flag — *Betsy Ross flag*
Although the Betsy Ross story is accepted by most **Americans**, some **flag** historians and revisionists do not accept the Betsy Ross design as the **first American flag**.

#345: what became of rich on price is right — *Rich Fields*
Richard Wayne "**Rich**" Fields is an American broadcaster, spokesman, announcer and meteorologist , best known for being the announcer of the American version of The **Price** Is **Right** from 2004–2010.

#352: what happened to stevie ray vaughan — *Stevie Ray Vaughan*
As the younger **brother** of **Jimmie Vaughan** , **Vaughan** started playing the **guitar** at age seven and formed several bands that **occasionally** performed in local nightclubs.

358: what caused ww — *Causes of World War I*
Most **historians** and popular commentators include **causes** from more than one category of **explanation** to provide a rounded account of the **causes** of the war.

#362: how many seasons were there of the wire — *The Wire*
Despite only **receiving** modest **ratings** and never **winning** major **television awards**, The **Wire** has been **described** by many **critics** as one of the greatest **TV dramas** of all time.

364: what kind of legal remedy is it to ask someone to fulfill there promise — *Contract*
Contract **law** varies **greatly** from one **jurisdiction** to another, including differences in **common law** compared to **civil law** , the impact of received **law** , particularly from England in **common law** countries, and of **law** codified in regional legislation.

365: how is ASP.NET different from .NET — *ASP.NET*
ASP.**NET** is built on the Common Language Runtime (CLR), allowing programmers to write ASP.**NET** code using any supported .**NET** language .

#370: how many people live in memphis tennessee — *Memphis, Tennessee*
A resident of **Memphis** is referred to as a **Memphian** , and the **Memphis** region is known, particularly to **media** outlets, as " **Memphis** & The Mid-South ".

#372: how many muscles in the human body — *List of muscles of the human body*
The **muscles** of the **human body** can be categorized into a number of groups which include **muscles** relating to the head and neck, **muscles** of the torso or trunk, **muscles** of the upper limbs, and **muscles** of the lower limbs.

373: how did seminole war end — *Seminole Wars*
The **Seminole Wars**, also known as the **Florida Wars**, were three **conflicts** in **Florida** between the **Seminole** — the collective name given to the amalgamation of various groups of native **Americans** and Black people who settled in **Florida** in the early 18th century — and the United States **Army** .

#385: how many nature oceans are on earth — *Ocean*
This animation uses **Earth** science data from a variety of sensors on NASA **Earth** observing satellites to measure physical oceanography parameters such as **ocean** currents, **ocean** winds, **sea** surface height and **sea** surface temperature.

388: what county is jacksonville florida in — *Jacksonville, Florida*
Jacksonville is the principal **city** in the **Jacksonville** metropolitan **area** , with a population of 1,345,596 in **2010**.

391: what does a liquid oxygen plant look like — *Liquid oxygen*
Liquid oxygen — abbreviated LOx, LOX or Lox in the aerospace , submarine and gas **industries** — is one of the physical forms of elemental **oxygen** .

402: What does the term "mens rea" mean — *Mens rea*
Therefore, **mens rea** refers to the mental element of the **offence** that accompanies the actus **reus**.
In Australia , for example, the elements of the federal **offences** are now designated as "fault elements" or "mental elements" (**mens rea**) and "physical elements" or "external elements" (actus reus).

406: what do UA's detect — *Drug test*
Major uses of drug testing are to **detect** the **presence** of performance enhancing **steroids** in **sport** or for drugs prohibited by laws, such as cannabis , cocaine and heroin.

414: What does Human sperm consist of? — *Semen*
The process that results in the discharge of semen is called **ejaculation** .

416: how did the penguins acquire sidney crosby — *Sidney Crosby*
The **Penguins returned** to the Finals against Detroit the following year and won in seven **games**; **Crosby** became the youngest captain in NHL history to win the Stanley Cup .

#420: how much total wealth in USA — *Wealth in the United States*
In addition, **wealth** is unevenly distributed, with the **wealthiest** 25% of US households **owning** 87% of the **wealth** in the **United States**, which was $54.2 **trillion** in 2009.

429: what does a plus-minus sign mean — *Plus-minus sign*
The **sign** is normally pronounced "**plus** or minus".
The **plus-minus sign** () is a mathematical symbol commonly used either

433: how was the moon formed — *Moon*
Red and orange tinted **Moon**, as seen from Earth during a lunar **eclipse**, where the Earth comes between the **Moon** and **Sun**

435: what did ronald reagan do as president — *Ronald Reagan*
Ronald Wilson **Reagan** (; February 6, 1911 – **June** 5, 2004) was the 40th **President** of the United States (1981–1989).

439: what does judgment as a matter of law mean — *Judgment as a matter of law*
JMOL motions may also be made after the verdict is returned, where they are called "renewed" motions for **judgment** as a **matter** of **law** (RJMOL), but the motion is still commonly known by its former name, **judgment** notwithstanding verdict , or j.n.o.v. (from the English **judgment** and the Latin non obstante veredicto).

#441: how many presidents have been assassinated — *List of United States presidential assassination attempts and plots*
Assassination attempts and plots on **Presidents** of the United States have been numerous: more than 20 attempts to kill sitting and former **presidents**, as well as the Presidents-elect , are known.

443: what does oklahoma produce — *Oklahoma*
Oklahoma City and **Tulsa** serve as **Oklahoma's** primary **economic** anchors, with nearly two thirds of **Oklahomans** living within their metropolitan statistical areas .

#444: how many redwall books are there — *Redwall*
It is the title of the first **book** of the series, published in 1986, as well as the name of the Abbey featured in the **book** and the name of an animated TV series **based** on three of the novels (**Redwall** , Mattimeo , and Martin the Warrior), which first aired in 1999.

446: what date did the american civil war start — *American Civil War*
The **American Civil War** (ACW), also known as the **War** between the States or **simply** the **Civil War** (see naming), was a **civil war fought** from 1861 to **1865** between the United States (the "Union" or the "North") and several Southern **slave** states that declared their secession and formed the Confederate States of **America** (the "Confederacy" or the "South").

452: how jameson irish whiskey is made — *Jameson Irish Whiskey*
With annual sales of over 31 million bottles, **Jameson** is by far the best selling **Irish whiskey** in the world, as it has been sold internationally since the early 19th century when John **Jameson** along with his son (also named John) was producing more than a million gallons annually.

#464: How many consecutive games did Ken Jennings win? — *Ken Jennings*
Jennings is noted for holding the record for the longest **winning** streak on the U.S. syndicated **game** show **Jeopardy!**

468: what county is jennings, la — *Jennings, Louisiana*
Jennings is the principal city of the **Jennings** Micropolitan Statistical Area , which includes all of **Jefferson Davis Parish**.

469: how does nanotechnology affect health — *Nanotechnology*
A more generalized description of **nanotechnology** was subsequently established by the National **Nanotechnology** Initiative , which defines **nanotechnology** as the manipulation of matter with at least one dimension sized from 1 to 100 nanometers .

476: what does it take to start a lodge in freemason — *Masonic Lodge*
By exception the three surviving **lodges** that formed the world's first known Grand **Lodge** in London (today called the United Grand **Lodge** of England) have the unique privilege to operate as time immemorial i.e. without such warrant; only one other **lodge** operates without a warrant - this is the Grand Stewards' **Lodge** in London, although it is not also entitled to the "time immemorial" title.

480: how does black pepper grow — *Black pepper*
Peppercorns, and the ground **pepper** derived from them, may be described simply as **pepper**, or more precisely as **black pepper** (cooked and **dried** unripe fruit), green **pepper** (**dried** unripe fruit) and white **pepper** (**dried** ripe seeds).

482: how south african leaders are elected — *President of South Africa*
The President of the Republic of **South Africa** is the head of state and head of government under **South Africa's** Constitution .

485: how was the phone invented — *Invention of the telephone*
This article covers the early years from 1844 to 1898, from conception of the **idea** of an electric voice-transmission device, to failed attempts to use "make-and-break" current, to successful experiments with electromagnetic telephones by Alexander Graham Bell and **Thomas Watson** , and finally to commercially successful telephones in the late 19th century.

491: what forms seasons — *Season*
Hot regions have two or three **seasons**; the **rainy** (or wet, or monsoon) **season** and the **dry season**, and in some **tropical** areas, a cool or mild **season**.

492: how did mohammed gandhi die — *Mahatma Gandhi*
Mohandas Karamchand Gandhi (; 2 October 1869 – 30 January 1948), commonly known as **Mahatma Gandhi**, was the preeminent leader of **Indian** nationalism in British-ruled **India** .

#495: how many users do twitter have — *Twitter*
Twitter is an online social networking **service** and microblogging **service** that enables its **users** to **send** and read text-based **messages** of up to 140 characters , known as "**tweets**".

513: how did the vietnam war end — *Vietnam War*
They viewed the **conflict** as a colonial **war** , **fought** initially against France, backed by the U.S., and later against South **Vietnam**, which it regarded as a U.S. puppet **state** .

#516: how much more time does chemo give to people with renal cancer — *Renal cell carcinoma*
Renal cell **carcinoma** (RCC, also known as **hypernephroma**) is a **kidney cancer** that originates in the lining of the proximal convoluted tubule , the very small tubes in the **kidney** that transport GF (glomerular filtrate) from the glomerulus to the descending limb of the nephron.

521: how does a solid state drive work — *Solid-state drive*
Hybrid **drives** or **solid state** hybrid **drives** (SSHD) combine the features of SSDs and HDDs in the same **unit**, containing a large hard **disk drive** and an SSD cache to improve performance of frequently accessed data.

#529: how long did the roman empire last — *Roman Empire*
The **Roman Empire** () was the post- **Republican** period of the ancient **Roman** civilization , characterised by an autocratic form of government and large territorial **holdings** around the Mediterranean in **Europe** , Africa , and Asia .

#533: how much does U.S. pay on health care per person — *Comparison of the health care systems in Canada and the United States*
Total government spending per capita in the U.**S**. on **health care** was 23% higher than Canadian **government spending**, and U.**S**. **government expenditure** on **health care** was just under 83% of **total** Canadian **spending** (**public** and **private**) though these statistics don't take into account population differences.

537: how is rfid tag powered — *Radio-frequency identification*
Radio-frequency **identification** (**RFID**) is the wireless non-contact use of radio-frequency electromagnetic **fields** to transfer data, for the purposes of automatically **identifying** and **tracking tags** attached to objects.

#556: how long have kanab ambersnail been endangered? — *Kanab ambersnail*
The **Kanab ambersnail**, scientific name Oxyloma haydeni kanabensis or Oxyloma kanabense, is a critically **endangered** subspecies or species of small, air-breathing land snail , a terrestrial pulmonate gastropod mollusc in the family Succineidae , the amber snails.
Now considered a Critically **Endangered** species on the IUCN Red List of Threatened Species due to a series of factors (including influence), the **Kanab ambersnail** has been reintroduced to three springs above the historic high water level along the Colorado River .

#557: how many spanish speaking people are there in the us — *Spanish language in the United States*
There are 45 million Hispanophones who **speak Spanish** as a first or second **language**, as well as six million **Spanish language** students, composing the **largest** national Spanish-speaking community outside of **Mexico** and making **Spanish** the Romance **language** and the Indo-European **language** with **largest** number of **native speakers** in the world.

558: what does (sic) mean? — *Sic*
Sic (noise artist) , styled as [sic], stage name of Jennifer Morris, a Canadian noise artist

560: how does sedimentary rock form — *Sedimentary rock*
The **sedimentary rock** cover of the continents of the **Earth's crust** is extensive, but the **total** contribution of **sedimentary rocks** is estimated to be only 8% of the **total** volume of the **crust**.

562: what did st.patrick do — *Saint Patrick*
The text, however, distinguishes between "Old **Patrick**" (thought to mean Palladius) and "**Patrick**, archapostle of the Scots ," who died in 492.

563: how do you know if something is the golden ratio — *Golden ratio*
Many 20th century artists and architects have **proportioned** their works to approximate the **golden ratio**—especially in the form of the **golden** rectangle , in which the **ratio** of the longer side to the shorter is the **golden ratio**—believing this **proportion** to be aesthetically pleasing (see Applications and observations below).

570: How is a computer used? — *Computer*

Simple **computers** are small enough to fit into mobile devices , and mobile **computers** can be powered by small batteries . Personal **computers** in their various forms are icons of the Information Age and are what most people think of as "computers."

♯581: how often does ham station need to ID? *Station identification*
Station identification (ident or channel **ID**) is the **practice** of radio or television **stations** or networks identifying themselves on air, typically by means of a call sign or brand name (sometimes known, particularly in the United States, as a "sounder" or "stinger", more generally as a **station** or network ID).

♯583: how many days are in a leap year *Leap year*
The term **leap year** gets its name from the fact that while a fixed **date** in the Gregorian calendar normally advances one **day** of the **week** from one **year** to the next, in a **leap year** the **day** of the **week** will advance two **days** (from March onwards) due to the year's extra **day** inserted at the end of **February** (thus "**leaping** over" one of the **days** in the week).

597: what county is bethlehem pa in *Bethlehem, Pennsylvania*
There are three general sections of the city: North **Bethlehem**, **South Bethlehem** and **West Bethlehem**.

598: what countries legalize marijuana *Legality of cannabis*
While **federal law** in the **United States** bans all sale and **possession** of **cannabis**, enforcement varies widely at the **state** level and some **states** have established medicinal **marijuana** programs in that contradict **federal law**; two **states** (Colorado and **Washington**) have repealed their **laws** prohibiting the recreational **use** of **cannabis** and replaced them with a regulatory regime, also contrary to **federal** statute.

♯603: how many albums has dmx sold to this date *DMX (rapper)*
In 1999, **DMX** released his best-selling **album** ...And Then There Was **X** , which featured the hit **single** " Party Up (Up in Here) ".

610: what do mucous membranes secrete *Mucous membrane*
The term **mucous membrane** refers to where they are found in the body and not every **mucous membrane** secretes mucus.

611: what company is cricket wireless by *Cricket Wireless*
Cricket Communications, Inc., (d.b.a. **Cricket Wireless**) founded in 1999, provides **wireless** services to over 7 million customers in the United States.

♯615: how many numbers on a credit card *Bank card number*
Payment card numbers are found on **payment cards** , such as **credit cards** and debit **cards** , as well as stored-value **cards** , gift **cards** and other similar **cards**.

617: how is public policy created *Public policy*
The U.S. professional association of **public policy** practitioners, researchers, scholars, and students is the Association for **Public Policy** Analysis and Management .
As an academic discipline , **public policy** is studied by professors and students at **public policy** schools of major universities throughout the country.

♯619: how many wives did henry the 8th have *Henry VIII of England*
Henry's struggles with Rome led to the separation of the **Church** of **England** from **papal** authority, the Dissolution of the Monasteries , and his own establishment as the Supreme Head of the **Church** of **England** .

620: what country are bongo drums from? *Bongo drum*
The **drums** are of different size: the larger **drum** is called in Spanish the hembra (female) and the smaller the macho (male).

♯621: how many books in bible *Books of the Bible*
The first part of Christian **Bibles** is the Old Testament , which contains, at minimum, the twenty-four **books** of the Hebrew **Bible** divided into thirty-nine **books** and **ordered** differently from the Hebrew **Bible**.

639: what does freedom of speech cover *Freedom of speech*
The **right** to **freedom** of expression is recognized as a **human right** under **Article 19** of the Universal Declaration of **Human Rights** and recognized in international **human rights** law in the International Covenant on Civil and Political **Rights** (ICCPR).

647: what county is san jose in? *San Jose, California*
San Jose was founded on November 29, 1777, as El Pueblo de **San José** de Guadalupe, the **first** civilian town in the **Spanish** colony of Nueva **California** , which later became Alta **California** .

651: what cars have smart key systems *Smart key*
A **smart key** is an electronic **access** and authorization **system** which is available as an option or standard in several **cars**.

676: how did John F. Kennedy die? *John F. Kennedy*
John Fitzgerald "Jack" **Kennedy** (May 29, 1917 – **November** 22, 1963), often referred to by his initials JFK, was the 35th President of the United States , serving from 1961 until his **death** in **1963**.

♯677: how many members are in the house of representatives *United States House of Representatives*
The Speaker of the **United States House** of **Representatives** , who **presides** over the chamber, is elected by the **members** of the **House**, and is therefore traditionally the **leader** of the **House** Democratic **Caucus** or the **House** Republican **Conference** , whichever party has more voting **members**.

678: what can silk be used for *Silk*
Many **silks** are mainly **produced** by the **larvae** of **insects** undergoing complete metamorphosis , but some adult **insects** such as webspinners **produce silk**, and some **insects** such as raspy crickets **produce silk** throughout their lives.

682: what does uncle sam represent to the american people *Uncle Sam*
The first use of **Uncle Sam** in literature was in the 1816 allegorical book "The Adventures of **Uncle Sam** in Search After His Lost Honor" by Frederick Augustus Fidfaddy, Esq. An **Uncle Sam** is mentioned as early as 1775, in the original " Yankee Doodle " lyrics of the Revolutionary War.

684: what do cyberstalkers do *Cyberstalking*
Cyberstalking is a criminal offense that comes into play under state anti-stalking **laws**, slander **laws**, and harassment **laws**.

685: how do insulin syringes work *Syringe*
The word "**syringe**" is derived from the Greek συριγξ syrinx = "tube" via back-formation of a new singular from its Greek-type plural "**syringes**" (συριγγες syringes).

692: how did david carradine die *David Carradine*
David Carradine (born John Arthur **Carradine**; **December** 8, 1936 – June 3, **2009**) was an American actor and martial artist, best known for his leading role as a warrior monk , Kwai Chang Caine , in the 1970s television series Kung Fu .

700: what affects the money supply *Money supply*
Second, if the velocity of **money** , i.e., the ratio between **nominal GDP** and **money supply**, changes, an **increase** in the **money supply** could have either no **effect**, an exaggerated **effect**, or an unpredictable **effect** on the growth of **nominal GDP**.

703: how is today special? *Today's Special*
It was **set** in a department store , based on the flagship location of the now defunct Simpson's in Toronto .

707: what does auld lang syne mean *Auld Lang Syne*
Consequently "For **auld lang syne**", as it appears in the first line of the chorus, might be loosely translated as "for (the sake of) **old times**".

715: how was the president involved in the gulf war *Gulf War*
The **war** is also known under other **names** , such as the Persian **Gulf War**, First **Gulf War**, **Gulf War** I, or the First Iraq **War**, before the **term** "Iraq **War**" became identified instead with the 2003 Iraq **War** (also referred to in the U.S. as "**Operation** Iraqi Freedom").

♯716: how many times has a player hit for the cycle *Hitting for the cycle*
One **NPB player** has also **hit** for the **cycle** in an **NPB All-Star game** .

723: what does it mean if i'm flat footed? *Flat feet*
Three studies (see citations below in military section) of military recruits have shown no evidence of later increased **injury**, or **foot** problems, due to **flat feet**, in a population of people who reach military service age without prior **foot** problems.

726: what age group is generation x *Generation X*
Generation X, commonly abbreviated to Gen **X**, is the **generation** born after the Western post-World War II baby boom .

729: how did James Dean die? *James Dean*
James Byron **Dean** (February 8, 1931 – September 30, 1955) was an American actor.

734: what glows in the dark *Glow-in-the-dark*
Glow in the **Dark** Tour , a 2008 concert tour by Kanye West

737: what did sparta do around 650 bc *Sparta*
Sparta (Doric **Greek** : ; Attic **Greek** :), or Lacedaemon, was a **prominent** city-state in **ancient Greece** , situated on the banks of the Eurotas River in Laconia, in south-eastern Peloponnese .

738: what fantasy american football means *Fantasy football (American)*
Due to the growth of the Internet, **Fantasy football** has vastly increased in **popularity**, particularly because **Fantasy football** providers such as ESPN, Yahoo, CBS, and the NFL itself are able to keep track of statistics entirely online, eliminating the need to check box scores and newspapers regularly to keep track of players.

♯741: how many muscles in the body *List of muscles of the human body*
The **muscles** of the human **body** can be categorized into a number of groups which include **muscles** relating to the head and neck, **muscles** of the torso or trunk, **muscles** of the upper limbs, and **muscles** of the lower limbs.

747: what does the green mean on the mexican flag *Flag of Mexico*
While the **meaning** of the **colors** has changed over time, these three **colors** were adopted by **Mexico** following **independence** from Spain during the country's War of **Independence** , and subsequent First **Mexican** Empire .

♯749: how often do elk have sex *Elk*
Male elk have large antlers which are shed each year.

♯757: how many players on a side for a football game *American football*
American football is the most **popular sport** in the **United States** today, and the **National Football League** (NFL) is its most **popular league**.

760: what does the temporal lobe part of the brain do *Temporal lobe*
The **temporal lobe** is a region of the cerebral cortex that is located beneath the lateral fissure on both cerebral hemispheres of the mammalian **brain** .

♯761: how many lungs does a human have *Human lung*
The right **lung** consists of three **lobes** while the left **lung** is slightly smaller consisting of only two **lobes** (the left **lung** has a "cardiac notch" allowing space for the heart within the chest).

768: what is the population of center tx *Center, Texas*
The Rio Theater in **Center**
It is named for its location near the **center** of Shelby County, not for its location in Texas as it is located near the Louisiana border.
First Christian Church at 124 Cora Street in **Center** is one of the oldest congregations in the community.
First Baptist Church at 117 Cora Street in **Center** is located next to the downtown section.
Chamber of Commerce Building in **Center**.
Center is a city in Shelby County , Texas , United States .

770: what causes photo red eye *Red-eye effect*
The red-eye **effect** in photography is the common appearance of **red** pupils in color photographs of **eyes** .

771: what creates a cloud *Cloud*
These include strato- for low **clouds** with limited convection that form mostly in uneven layers, cumulo- for complex highly-convective storm **clouds**, nimbo- for thick layered **clouds** of some complexity that can produce moderate to heavy precipitation, alto- for middle **clouds**, and cirro- for high **clouds**; the latter two of which may be of simple or moderately complex structure.

♯772: how many students go to santa barbara *University of California, Santa Barbara*
The **University** of **California**, **Santa Barbara** (**commonly** referred to as UC **Santa Barbara** or UCSB) is a public research **university** and one of the ten general campuses of the **University** of **California system**.

774: what culture is mariah carey *Mariah Carey*
Under the guidance of Columbia Records executive Tommy Mottola , **Carey** released her self-titled debut studio album **Mariah Carey** in 1990; it went multi-platinum and spawned four consecutive number one singles on the U.S. Billboard Hot 100 chart.

779: how was color introduced in film? *Color motion picture film*
Color motion picture film refers both to unexposed **color** photographic **film** in a format suitable for use in a **motion picture** camera , and to finished **motion picture film** , ready for use in a **projector** , which bears images in **color**.

781: what artist have song with ashanti? *Ashanti (entertainer)*
Ashanti is most famous for her eponymous debut album , which **featured** the hit **song** " Foolish ", and sold over 503,000 copies in its first week of release throughout the U.S. in April **2002**.

783: what came first army or air force *United States Army Air Forces*
Although other **nations** already had separate **air forces independent** of the **army** or navy (such as the British Royal **Air Force** and the German Luftwaffe), the AAF remained a part of the **United States Army** until the **United States Air Force came** into being in September **1947**.

784: what day is 2011 super bowl? *Super Bowl XLV*
Unlike other matchups, this game featured two title-abundant franchises: coming into the game, the **Packers** held the most NFL **championships** with 12 (9 league **championships** prior to the **Super Bowl** era and 3 **Super Bowl** championships), while the Steelers held the most **Super Bowl championships** with 6.

785: what does the FOIA apply to *Freedom of Information Act (United States)*
The Federal Government's Freedom of Information **Act** should not be confused with the different and varying Freedom of Information **Acts** passed by the individual states .
Many of those state **acts** may be similar but not identical to the federal **act**.

♯786: how many countries are member of the eu? *European Union*
With a combined **population** of over 500 **million** inhabitants, or 7.3% of the world **population**, the **EU**, in **2012**, generated a nominal gross domestic product (GDP) of 16.584 **trillion** US dollars, representing approximately 20% of the global GDP when measured in terms of purchasing power parity , and represents the **largest** nominal GDP and GDP PPP in the world.

♯790: how many vehicles are registered in the us *Passenger vehicles in the United States*
This number, along with the average age of **vehicles**, has increased steadily since 1960, indicating a growing number of **vehicles** per capita.

792: what cards do you need in poker to get a royal flush *List of poker hands*
The ranking of a particular hand is increased by including multiple **cards** of the same **card** rank, by all five **cards** being from the same suit , or by all five **cards** being of consecutive rank.

799: what does xylem transport *Xylem*
The word **xylem** is derived from the Greek word ξύλον (xylon), meaning "**wood**"; the best-known **xylem** tissue is **wood** , though it is **found** throughout the plant.

♯800: how many mvps has kobe bryant won *Kobe Bryant*
In 2006, **Bryant scored** a career-high 81 points against the Toronto Raptors , the second most points **scored** in a **single game** in **NBA** history, second only to Wilt Chamberlain's 100-point **game** in 1962.

804: what do jehovah witnesses believe *Jehovah's Witnesses*
Jehovah's Witnesses is a millenarian restorationist Christian denomination with nontrinitarian **beliefs** distinct from mainstream Christianity .Sources for descriptors:• Millenarian: • Restorationist: • Christian: • Denomination: The organization reports worldwide membership of over 7.78 million adherents involved in evangelism , convention attendance of over 12 million, and annual Memorial attendance of over 19 million.

♯806: how old old is xp operating system *Windows XP*
According to web analytics **data generated** by Net Applications , Windows **XP** was the most **widely** used **operating system** until August 2012, when Windows 7 overtook it.

♯811: how many pawns in chess *Pawn (chess)*
It is also **common** to refer to a rook **pawn** , **meaning** any **pawn** on the a- or h-file, a knight **pawn** (on the b- or g-file), a bishop **pawn** (on the c- or f-file), a queen **pawn** (on the d-file), a king **pawn** (on the e-file), and a central **pawn** (on either the d- or e-file).

814: how is hydrogen produced *Hydrogen production*
Hydrogen production is the family of industrial methods for generating **hydrogen**.

833: what country is madrid spain in *Madrid*
As the **capital** city of **Spain**, seat of government , and **residence** of the **Spanish** monarch , **Madrid** is also the political, **economic** and cultural centre of **Spain**.

835: What causes rogue waves *Rogue wave*
Rogue waves (also known as freak **waves**, monster **waves**, killer **waves**, extreme **waves**, and abnormal **waves**) are relatively large and spontaneous ocean surface **waves** that occur far out at sea, and are a threat even to large ships and ocean liners .

♯837: what happened to montgomery clift *Montgomery Clift*
Edward **Montgomery Clift** (October 17, 1920July 23, 1966) was an American film and stage actor The New York Times' obituary noted his portrayal of "moody, sensitive young men".

♯850: what happened to george o'malley on grey's anatomy? *George O'Malley*
Introduced as a **surgical** intern at the fictional Seattle Grace Hospital, **O'Malley** worked his way up to resident level, while his **relationships** with his colleagues Meredith **Grey** (Ellen Pompeo), Cristina Yang (Sandra Oh), Izzie **Stevens** (Katherine Heigl) and Alex Karev (Justin Chambers) formed a focal point of the series.

♯852: how many people die from myasthenia gravis per year *Myasthenia Gravis*
Myasthenia gravis (from Greek μύς "muscle", "weakness", and "serious"; abbreviated MG) is an autoimmune neuromuscular **disease** leading to fluctuating muscle weakness and fatigability .

♯854: how many rooms in borgata hotel *The Borgata*
The **Borgata Hotel** Casino and **Spa** is a luxury **hotel** , casino , and **spa** in Atlantic City, New Jersey , United States .

859: what country has the most muslims in the world *Islam by country*

In the Middle East, the non-Arab **countries** of Turkey and Iran are the largest **Muslim-majority countries**; in Africa, Egypt and Nigeria have the most populous **Muslim** communities.

861: what country is dubai in *Dubai*

The Sheikhdom of **Dubai** was formally established in 1833 by Sheikh Maktoum bin Butti Al-Maktoum when he persuaded around 800 members of his tribe of the Bani Yas , living in what was then the Second Saudi State and now part of Saudi Arabia , to follow him to the **Dubai** Creek by the Abu Falasa clan of the Bani Yas.

♯863: how much of earth is covered ocean water *Ocean*

The Mars **ocean** hypothesis **suggests** that nearly a third of the surface of Mars was once **covered** by **water**, though the **water** on Mars is no longer **oceanic**, and a runaway greenhouse effect may have boiled away the global **ocean** of Venus.

♯870: how much did yankee stadium cost *Yankee Stadium*

The **first** game at the new **Yankee Stadium** was a pre-season exhibition game against the Chicago Cubs **played** on **April** 3, 2009, which the **Yankees won** 7–4.

♯879: What happened to "The Glades" tv series *The Glades (TV series)*

The **Glades** was renewed by A&E for a third **season** on October 18, 2011, which aired from June 3 to August 12, 2012.

♯887: how many albums has eminem sold in his career *Eminem*

His next two records The **Marshall Mathers LP** , and The **Eminem** Show , also **won** Best Rap **Album Grammy Awards**, making **Eminem** the first **artist** to **win** Best Rap **Album** for three **consecutive** LPs.

891: what do pigs eat *Pig*

Pigs include the domestic **pig** , its ancestor the wild boar , and several other wild relatives.

♯900: What happened during the Starving Time in Jamestown? *Starving Time*

The **Starving Time** at **Jamestown** in the Colony of Virginia was a period of **starvation** during the winter of 1609–1610 in which all but 60 of 500 colonists died.

905: what does the president of the usa do *President of the United States*

It also prohibits a **person** from being elected to the **presidency** more than once if that **person** previously had **served** as **president**, or acting **president** , for more than two years of another **person's** term as **president**.

911: how kimberlite pipes form *Volcanic pipe*

Volcanic **pipes** are geological structures **formed** by the violent, supersonic eruption of deep-origin volcanoes .

♯913: how many seasons of grey's anatomy are there *Grey's Anatomy*

Dr. Preston Burke (Isaiah Washington) **departs** at the conclusion of the third **season** , and is **replaced** by **Dr**. Erica Hahn (Brooke Smith), who leaves the show during the fifth **season** , and later **Dr**. Teddy Altman (Kim Raver), who **departs** at the **end** of the eighth **season** .

914: how is schizophrenia diagnosed? *Schizophrenia*

Despite the etymology of the term from the Greek roots skhizein (σχίζειν, "to **split**") and phrēn, phren- (φρήν, φρεν-; "mind"), **schizophrenia** does not imply a "**split personality**", or "multiple **personality** disorder" (which is known these days as dissociative identity disorder)—a condition with which it is often confused in public perception.

931: what kind of company is Microsoft? *Microsoft*

As of 2013, **Microsoft** is market dominant in both the PC operating **system** and office suite markets (the latter with **Microsoft** Office).

938: how does weather happen *Weather*

Studying how the **weather works** on other planets has been helpful in understanding how **weather works** on Earth.

♯940: how many countries have english as an official language *List of countries where English is an official language*

According to the Constitution of India , " Hindi in the Devanagari script" is the **official language** of the union; and **English** the 'subsidiary **official language**'; however, **English** is mandated for the authoritative texts of all federal laws and Supreme Court decisions and (along with Hindi) is one of the two **languages** of the Indian Parliament .

941: what countries are in cono sur *Southern Cone*

High life expectancy , the **highest** Human Development Index of Latin America, **high Standard** of living , significant participation in the global markets and the emerging economy of its members make the **Southern Cone** the most prosperous macro-region in **South** America.

949: what branch of the military is delta force *Delta Force*

The Central Intelligence Agency 's highly secretive Special Activities Division (SAD) and more **specifically** its elite Special **Operations** Group (SOG) often works with – and recruits – **operators** from **Delta Force**.

957: what causes the seasons *Season*

Hot regions have two or three **seasons**; the **rainy** (or wet, or monsoon) **season** and the **dry season**, and in some **tropical** areas, a cool or mild **season**.

959: what does 3g network mean *3G*

3G finds application in **wireless** voice **telephony** , mobile Internet access, fixed **wireless** Internet access, video calls and mobile TV .

960: what does Mazel tov! mean *Mazel tov*

"**Mazel tov**" or "**mazal tov**" (Hebrew / Yiddish : מזל טוב," Hebrew: "**mazal tov**"; Yiddish: "**mazel tov**"; lit.

961: How Do You Get Hepatitis C *Hepatitis C*

Hepatitis C is an infectious disease affecting primarily the liver , caused by the **hepatitis C** virus (HCV).

962: how does flexible spending account work *Flexible spending account*

The most common **type** of **flexible spending account**, the medical expense FSA (also medical FSA or **health** FSA), is similar to a **health** savings **account** (HSA) or a **health** reimbursement **account** (HRA).

♯966: how many baseball teams usa *Major league baseball*

Major **League Baseball** (**MLB**) is a professional **baseball league** , consisting of **teams** that play in the **National League** and the **American League** .

♯972: how many amendments in the US constitution *List of amendments to the United States Constitution*

To date, no convention for proposing **amendments** has been called by the states, and only once—in 1933 for the ratification of the twenty-first **amendment** —has the convention method of ratification been employed.

979: what day is the federal holiday for Martin Luther King Jr. *Martin Luther King, Jr. Day*

Martin Luther King, **Jr**. **Day** is a United States **federal holiday** marking the birthday of Rev. Dr. **Martin Luther King**, **Jr**..

985: what area code is 949 *Area code 949*

The **area code** in red is **Area Code 949**; all others in blue are California **area codes**.

♯994: how many babies are in a typical raccoon litter *Raccoon*

The **raccoon** (, Procyon lotor), sometimes spelled **racoon**, also known as the common **raccoon**, North **American raccoon**, northern **raccoon** and colloquially as coon , is a medium-sized mammal **native** to North **America**.

998: What does Rapture meaning in a theological sense? *Rapture*

Denominations such as Roman Catholics , Orthodox **Christians** , Lutheran **Christians**, and Reformed **Christians** believe in a **rapture** only in the **sense** of a **general** final **resurrection**, when **Christ** returns a single time.

1002: what are corporation balance *Balance sheet*

In financial accounting , a **balance** sheet or statement of financial position is a summary of the financial **balances** of a sole proprietorship , a business partnership , a **corporation** or other business organization, such as an LLC or an LLP .

1003: what is .17 hmr caliber *.17 HMR*

It descended from the .22 Magnum by necking down the .22 Magnum case to take a .17 **caliber** (4.5 mm) bullet, and it is more costly to shoot than traditional .22 **caliber** rimfire cartridges.

1007: what made the civil war different from others *American Civil War*

The **American Civil War** (ACW), also known as the **War** between the States or **simply** the **Civil War** (see naming), was a **civil war** fought from 1861 to **1865** between the United States (the "Union" or the "**North**") and several **Southern slave** states that declared their secession and formed the Confederate States of **America** (the "Confederacy" or the "**South**").

1008: where to buy potato bread made without wheat *Potato bread*

Potato bread is a form of **bread** in which **potato** replaces a portion of the regular **wheat** flour .

1014: who killed julius caesar *Julius caesar*

Gaius Julius Caesar (, July 100 BC – 15 March 44 BC) was a Roman general , statesman , Consul and notable author of Latin prose.

1024: who wrote white christmas *White Christmas (song)*

I just **wrote** the best **song** I've ever **written** — heck, I just **wrote** the best **song** that anybody's ever written!"

1027: WHAT IS A FY QUARTER *Fiscal year*

Many universities have a **fiscal year** which ends during the summer, both to align the **fiscal year** with the school **year** (and, in some cases involving public universities, with the state government's **fiscal** year), and because the school is normally less **busy** during the summer months.

1028: who pulmonary hypertension *Pulmonary hypertension*

In medicine , **pulmonary hypertension** (PH) is an increase of **blood pressure** in the **pulmonary** artery , **pulmonary** vein , or **pulmonary** capillaries, together known as the **lung** vasculature , leading to shortness of breath , dizziness , fainting , and other symptoms, all of which are exacerbated by exertion.

1033: what is a notary for *Notary public*

With the exceptions of Louisiana , Puerto Rico , Quebec , whose private law is based on civil law , and British Columbia , whose notarial tradition stems from scrivener **notary** practice, a **notary** public in the rest of the United **States** and most of Canada has powers that are far more limited than those of civil-law or other **common-law notaries**, both of whom are qualified lawyers admitted to the bar: such **notaries** may be referred to as notaries-at-law or lawyer **notaries**.

♯1038: when did the cold war start *Cold War*

The tensest times were during the **Berlin** Blockade (1948–1949), the **Korean War** (1950–1953), the Suez Crisis (1956), the **Berlin** Crisis of 1961 , the Cuban missile crisis (1962), the Vietnam **War** (1959–1975), the Yom Kippur **War** (1973), the Soviet **war** in **Afghanistan** (1979–1989), the Soviet downing of **Korean** Air Lines Flight 007 (1983), and the "Able Archer" NATO military exercises (1983).

1039: what state is pike's peak in? *Pikes Peak State Park*

It gets its name from the Iowa incarnation of **Pikes Peak**, a particularly high point overlooking the gorge of the Upper Mississippi, and like **Pikes Peak** in Colorado , is named for Zebulon **Pike** .

1049: what is a book index *Index (publishing)*

In a traditional back-of-the-book **index** the headings will include names of people, places and events, and concepts selected by a person as being relevant and of interest to a possible reader of the **book**.

1050: who sings the song never ending story *The NeverEnding Story (song)*

"The NeverEnding **Story**" (titled "The NeverEnding **Story** (L'histoire sans fin)" in the French version) is the title **song** from the English version of the 1984 film The NeverEnding **Story** .

1053: who created facebook *Facebook*

Facebook is an online social networking service , whose **name** stems from the colloquial **name** for the book given to **students** at the start of the academic year by some **university** administrations in the United States to help **students** get to know each other.

1057: what are risk for infections *Risk of infection*

Although anyone can become **infected** by a pathogen, patients with this diagnosis are at an elevated **risk** and extra **infection** controls should be considered.

1061: where do cruises dock in new york city *New York Passenger Ship Terminal*

With an upsurge in **cruise** ship traffic and the terminal's ability to handle comfortably only three large ships at a time, two **new** terminals have opened in the harbor — the Cape Liberty **Cruise** Port opened in 2004 in Bayonne, **New** Jersey (used by Royal Caribbean **Cruise** Line , Celebrity **Cruises** and Azamara **Cruises**), and the Brooklyn **Cruise** Terminal (used by the Queen Mary 2 and other ships of the Carnival Corporation **cruise** brands) opened in 2006 in Brooklyn, **New York** .

1064: who played the lead roles in the movie leaving las vegas *Leaving Las Vegas*

After limited **release** in the United States on October 27, 1995, **Leaving Las Vegas** made its nationwide **release** on February 9, 1996, receiving strong praise from critics and audiences.

1065: what is a CMM machine *Coordinate-measuring machine*

A **machine** which takes readings in six degrees of freedom and displays these readings in mathematical form is known as a **CMM**.

♯1067: what percentage of water in in the body *Body water*

In physiology , **body water** is the **water** content of the human **body** .

1068: what type of batteries are 357 (LR44) *LR44 battery*

Silver-oxide **batteries type** SR44 may provide extra capacity compared to **LR44 types** but have slightly different voltage characteristics.

1069: where did hurricane katrina begin *Hurricane Katrina*

The **hurricane strengthened** to a Category **5 hurricane** over the warm Gulf water, but weakened before making its second landfall as a Category 3 **hurricane** on the morning of Monday, **August** 29 in southeast **Louisiana**.

1073: what is 1 mil guaranies in united states dollars *Mill (currency)*

In the **United States**, it is a notional unit equivalent to of a **United States dollar** (a tenth of a cent).

1075: where did the persian war take place *Greco-Persian Wars*

The **allied Greeks** followed up their success by destroying the rest of the **Persian fleet** at the **Battle** of Mycale , before expelling **Persian garrisons** from **Sestos** (479 BC) and **Byzantium** (**478** BC).

1078: what is a day care for? *Day care*

Some childminders **care** for children from several **families** at the same time, either in their own **home** (commonly known as "**family day care**" in Australia) or in a **specialized** child **care** facility.

1079: who discovered neptune the planet *Discovery of Neptune*

Unfortunately, Le Verrier's triumph also led to a tense international dispute over priority, as, shortly after the **discovery**, George Airy , at the time British Astronomer **Royal** , announced that Adams had also predicted the **discovery** of the **planet**.

1081: what separates me from you album *What Separates Me from You*

The **album** debuted on the US Billboard 200 at **number** 11 with 58,000 first week sales, becoming A Day to Remember's personal best, as Homesick peaked at **number** 21.

♯1084: when did thomson make the plum-pudding model *Plum pudding model*

Still, **Thomson's model** (along with a similar Saturnian ring **model** for atomic electrons , also put forward in 1904 by Nagaoka after James Clerk Maxwell 's **model** of Saturn's rings), were earlier harbingers of the later and more successful solar-system-like Bohr **model** of the atom.

1085: what part of the pig is bacon *Bacon*

Meat from other animals, such as beef , lamb , chicken , goat , or turkey , may also be cut, cured, or otherwise prepared to resemble **bacon**, and may even be **referred** to as "**bacon**".

1086: where did erisa come from *Employee Retirement Income Security Act*

ERISA is sometimes used to refer to the full body of **laws** regulating employee benefit plans, which are found mainly in the Internal Revenue Code and **ERISA** itself.

1087: what percent of illegal immigrants are from mexico and europe *Illegal immigration to the United States*

Illegal immigration to the **United** States is the act of foreign **nationals** entering the **United** States , without **government** permission and in violation of **United** States nationality law , or staying beyond the termination date of a visa, also in violation of the law.

♯1092: when us subprime mortgage market collapse *Subprime mortgage crisis*

As **adjustable-rate mortgages began** to reset at **higher interest** rates (**causing higher monthly** payments), **mortgage delinquencies** soared.

1098: what type of game is heavy rain *Heavy Rain*

It won 2010's **Game** of the Year from CNN and **Gaming** Union, and Best PS3 **Game** of **2010** by GameSpy and IGN .

1099: what are superannuation contributions? *Superannuation in Australia*

For example, **employers** are required to pay a proportion of an **employee's** salaries and wages (currently 9%) into a **superannuation** fund, but people are encouraged to put aside **additional** funds into **superannuation**.

1102: what is a group of deer called *Deer*

Species in the Cervidae **family** include white-tailed **deer** , mule **deer** such as black-tailed **deer** , **elk** , moose , red **deer** , **reindeer** (caribou), **fallow deer** , roe **deer** and chital .

1103: what state was john mccain a senator in during the 2008 election *United States presidential election, 2008*

This **election** was also notable for being the first time in U.**S**. history that both **major party** candidates were sitting U.**S**. **Senators**, only the third time (after 1920 and 1960) that any sitting U.**S**. **Senator** was **elected president**, and only the second time that the **winning President** and **Vice President** (Obama and Biden) were both sitting U.**S**. **Senators**.

♯1104: when does the electoral college votes *Electoral College (United States)*

Maine and Nebraska use the "**congressional** district method", selecting one **elector** within each **congressional** district by **popular vote** and selecting the remaining **two electors** by a statewide **popular vote**.

♯1105: when barack obama was born *Barack Obama*

He began his **presidential campaign** in **2007**, and in **2008**, after a close **primary campaign** against Hillary Rodham Clinton , he won sufficient delegates in the Democratic Party **primaries** to receive the **presidential** nomination.

♯1128: what year did disney's animal kingdom lodge open *Disney's Animal Kingdom Lodge*

Disney's Animal Kingdom Lodge is located in the **Animal Kingdom** Resort Area , adjacent to **Disney's Animal Kingdom** .

1130: what are two languages in Nigeria? *Languages of Nigeria*

Nigeria's linguistic diversity is a microcosm of Africa as a whole, encompassing three major African **languages** families : Afroasiatic , Nilo-Saharan , and Niger–Congo .

1133: what war led to Pearl Harbor *Events leading to the attack on Pearl Harbor*

Rather than seize and fortify the islands, and wait for the inevitable US counterattack, Japan's military **leaders** instead decided on the pre-emptive **Pearl Harbor** attack, which they assumed would negate the American forces needed for the liberation and reconquest of the islands.
A series of events **led** to the attack on **Pearl Harbor** .

1141: where in the world are smallpox common *Smallpox*

The disease was originally known in English as the "pox" or "red **plague**"; the term "**smallpox**" was first used in **Britain** in the 15th **century** to distinguish **variola** from the "great pox" (syphilis).

#1142: what year did the beatles came out with the song i wanna hold your hand — *I Want to Hold Your Hand*
"I Want to **Hold** Your **Hand**" is a **song** by the English rock band the **Beatles** .

1143: what is a gasser car — *Gasser (car)*
The **gasser** is the predecessor of the modern Funny **Car** .

1145: what tensions preceded the berlin blockade — *Berlin Blockade*
During the multinational **occupation** of post–World War II **Germany** , the **Soviet** Union **blocked** the **Western Allies** ' railway, road, and canal access to the sectors of **Berlin** under **Allied control**.

1153: who sang that song-a change is going to come — *A Change Is Gonna Come*
The **song** has gained in popularity and critical acclaim in the decades since its release, and is #12 on Rolling Stone 's 500 Greatest **Songs** of All Time .

#1155: what percentage of the human body is water — *Body water*
In physiology , **body water** is the **water** content of the **human body** .

1158: what is a form of legal ownership — *Ownership*
Given a short-sighted **owner**, however, a **private** property system can make these **tragedies** worse—for example, a **private owner** of a piece of oil-rich property, depending on his worldview, might be more interested in short-term financial gain than incremental use with an eye toward other's concerns (e.g., those of future generations, the disenfranchised, etc.).

1162: who wrote what's my name rihanna — *What's My Name? (Rihanna song)*
An accompanying music **video**, directed by Philip Andelman, portrays a romantic encounter between **Rihanna** and Drake in a grocery store along with romantic **scenes** between the pair and **Rihanna walking** through Manhattan 's **Lower** East Side .

#1163: what year did isaac newton die — *Isaac Newton*
In addition to his work on the **calculus**, as a **mathematician Newton** contributed to the study of power series , generalised the binomial **theorem** to non-integer exponents, and developed **Newton's** method for approximating the roots of a function .

1164: what is a constant in math? — *Mathematical constant*
What it means for a **constant** to arise "naturally", and what makes a **constant** "interesting", is ultimately a matter of taste, and some mathematical constants are notable more for historical reasons than for their intrinsic mathematical interest.

1178: who owns land rover — *Land Rover*
The ongoing commercial success of the original **Land Rover** series models, and latterly the Range **Rover** in the 1970s in the midst of BL's well documented business troubles prompted the establishment of a separate **Land Rover company** but still under the BL umbrella, remaining part of the subsequent **Rover Group** in 1988, under the **ownership** of British Aerospace after the remains of British **Leyland** were broken up and privatised.

1180: what states have legalized prostitution — *Prostitution in the United States*
As with other countries, **prostitution** in the United **States** can be divided into three broad categories: **street prostitution**, **brothel prostitution**, and escort **prostitution**.

1181: what states are on the east coast — *East Coast of the United States*
The **East Coast** of the United **States**, also known as the **Eastern** Seaboard or the Atlantic Seaboard and commonly shortened to **East Coast**, refers to the easternmost **coast** of the United **States** along the Atlantic Ocean .

#1183: when did sertraline come on the market — *Sertraline*
Sertraline is primarily **prescribed** for major **depressive** disorder in adult outpatients as well as obsessive–compulsive , panic , and social anxiety disorders in both adults and children.

#1186: when did secretariat win — *Secretariat (horse)*
Secretariat's grandsire, Nasrullah , is also the great-great-grandsire of 1977 Triple Crown **winner** Seattle Slew .

1187: what are batteries made up of — *Automotive battery*
Battery recycling of **automotive batteries** reduces the need for resources required for manufacture of new **batteries**, diverts toxic **lead** from landfills, and prevents risk of improper disposal.

1189: what is a full job time?? — *Full-time*
Full-time jobs are often considered careers .

1190: what are the three primary colors in the subtractive color model — *Subtractive color*
A **subtractive color model** explains the mixing of a limited set of dyes , inks , paint pigments or natural **colorants** to create a wider range of **colors** , each the result of partially or completely **subtracting** (that is, absorbing) some wavelengths of light and not others.

1191: what are layers of the ionosphere — *Ionosphere*
It is distinguished because it is **ionized** by solar radiation.

1201: who plays as big bird — *Big Bird*
Big Bird is a protagonist of the children's television show Sesame Street .

1205: who founded walmart — *Walmart*
Walmart remains a family-owned **business** , as the company is controlled by the Walton family , who own a 48 percent stake in **Walmart**.

1206: what are some legal uses of meth — *Methamphetamine*
Both dextromethamphetamine and racemic methamphetamine are **Schedule II controlled substances** in the **United** States, and similarly the production, distribution, sale, and possession of methamphetamine is restricted or **illegal** in many jurisdictions .

1208: what is 6 pin din connector — *Mini-DIN connector*
Mini-DIN is similar to the larger, older **DIN connector** .

1211: Where does the word baptism come from — *Baptism*
While **John** the Baptist 's **use** of a deep river for his **baptism** suggests **immersion**, pictorial and archaeological evidence of **Christian baptism** from the 3rd century onward indicates that a normal form was to have the candidate stand in **water** while **water** was poured over the upper body.

1212: what school did Zach Thomas play for before making it in to the NFL — *Zach Thomas*
Zachary Michael **Thomas** (born September 1, 1973) is a former American college and professional football **player** who was a linebacker in the National Football League (**NFL**) for thirteen seasons.

1213: what order is the moth — *Moth*
Most of this **order** are **moths**; there are thought to be approximately 160,000 species of **moth** (nearly ten times the number of species of butterfly), with thousands of species yet to be described.

1215: what are the uses for gui — *Graphical user interface*
The term **GUI** is restricted to the scope of two-dimensional display screens with display resolutions able to describe generic information, in the tradition of the computer **science** research at the PARC (Palo Alto Research Center).

1217: who won antm cycle 12 — *America's Next Top Model, Cycle 12*
America's Next Top **Model**, **Cycle 12** is the twelfth **cycle** of **America's** Next Top **Model** and the sixth **season** to air on The CW network.

1219: who built the globe — *Globe Theatre*
A modern reconstruction of the **Globe**, named " Shakespeare's **Globe** ", opened in 1997 approximately from the site of the original theatre.

1221: what are warehouse spreadsheets used for — *Data warehouse*
In computing , a data **warehouse** or enterprise data **warehouse** (**DW**, DWH, or EDW) is a database **used** for **reporting** and data **analysis** .

1224: Who Started the Mormon Church — *Mormons*
Today **Mormons** are understood to be **members** of The **Church** of Jesus **Christ** of **Latter-day** Saints (**LDS** Church).

1226: what part of the earth's structure is believed to consist of tectonic plates — *Plate tectonics*
Plate tectonics (from the Late Latin **tectonicus**, from the "pertaining to building") is a scientific theory that describes the large-scale **motions** of **Earth** 's **lithosphere** .

1232: who won the 2009 super bowl — *Super Bowl XLIII*
The club became an unexpected **winner** during the regular **season**, compiling a 9–7 **record**, and the playoffs with the aid of head **coach** Ken Whisenhunt , who was the Steelers' offensive coordinator in **Super Bowl** XL , and the re-emergence of **quarterback** Kurt **Warner** , who was the **Super Bowl** MVP in **Super Bowl** XXXIV with his former **team**, St. Louis Rams .

1233: what is a popular people meter — *People meter*
The **People Meter** is an electronic method of television measurement that moved from active and diary-based to passive and **meter-monitored**.
Finally in 1986, Nielsen developed an electronic **meter**, **People Meter**, to solve the problem.

1235: what are the characteristics of bryophytes — *Bryophyte*
Bryophytes produce enclosed **reproductive** structures (**gametangia** and sporangia), but they produce neither flowers nor seeds , **reproducing** via spores .

#1236: when to use semicolon — *Semicolon*
The **semicolon** (;) is a punctuation mark with several **uses**.
The modern **uses** of the **semicolon** relate either to the listing of items or to the linking of related clauses .
Ben Jonson was the first notable English writer to **use** the **semicolon** systematically.

#1238: when did xbox release — *Xbox*
Although these two are free while **Xbox** Live required a subscription, as well as broadband-only connection which was not completely adopted yet, **Xbox** Live was a success due to better servers , features such as a buddy list, and milestone titles like Halo 2 **released** in November 2004, which is the best-selling **Xbox** video game and was by far the most popular online game for years.

1242: what is a roll play games — *Role-playing game*
Role-playing **games** also **include single-player** offline role-playing video **games** in which players **control** a character or team who undertake quests, and may **include** capabilities that advance using statistical mechanics.

1254: what is a forward in soccer — *Forward (association football)*
Modern team formations usually **include** one to three **forwards**; **two** is the most common.

1255: WHO QUALIFIES AS A DOMESTIC PARTNER IN CA — *Domestic partnership in California*
Initially, **domestic partnerships** enjoyed very few privileges—principally just hospital-visitation **rights** and the **right** to be claimed as a next of kin of the estate of a deceased **partner**.

1256: where did mark jackson play pro basketball? — *Mark Jackson (basketball)*
Mark A. **Jackson** (born April 1, 1965) is an American **basketball coach**, and former player who is the current head **coach** of the Golden State Warriors of the NBA .

1267: what are stink bombs made of — *Stink bomb*
A **stink bomb** is a device designed to create an unpleasant smell .
A prank **stink bomb**.

1270: who said "A picture is worth a thousand words"? — *A picture is worth a thousand words*
The discussion of "One **Picture Worth Thousand Words**" versus "One **Picture Worth** Ten **Thousand Words**" Wan yen I hua and 10.000 miles **worth** 10.000 books is cited in Information graphics where the concept of many in different disciplines and cultures.

1287: what teams won super bowl — *List of Super Bowl champions*
Super Bowl III in 1969 was the first such game that carried the "**Super Bowl**" moniker, the names " **Super Bowl** I " and " **Super Bowl** II " were retroactively applied to the first two games.

#1288: what year was christianity introduced to sub-saharan africa — *Christianity in Africa*
However, the most recent December 18, 2012 **Pew Forum** research **estimates** that in 2010, 6.010 **million Christians**, 3.270 **million traditional African religion** followers, 610,000 Muslims and 50,000 unaffiliated (no known **religion**) peoples lived in South Sudan.

#1295: when did expos become nationals — *Washington Nationals*
In , the **Expos** won a division **championship**, won their first-ever playoff series by defeating the Philadelphia Phillies , 3–2, and advanced to the **National** League **Championship** Series , where they would go on to lose to the Los Angeles Dodgers , 3–2, in their only postseason appearance during the strike-shortened season.

1296: what is a contingent fee with an attorney — *Contingent fee*
A **contingent fee** (in the **United** States) or conditional **fee** (in England and Wales) is any **fee** for services provided where the **fee** is payable only if there is a favourable result.

1297: what is a D.O. stand for medical doctor — *Doctor of Osteopathic Medicine*
Many D.**O**. physicians attend the same graduate **medical** education programs as their M.**D**. counterparts, and then take M.**D**. specialty **board** exams, while other D.**O**. graduates enter osteopathic programs and take D.**O**. specialty **board** examinations.

1307: who reports the consumer price index — *Consumer Price Index*
A **consumer price index** (**CPI**) measures changes in the **price** level of a market basket of **consumer** goods and services purchased by households.

1309: what are rocker arms? — *Rocker arm*
The effective leverage of the **arm** (and thus the force it can exert on the valve stem) is determined by the **rocker arm** ratio, the ratio of the distance from the **rocker arm's** center of rotation to the tip divided by the distance from the center of rotation to the point acted on by the camshaft or pushrod.

1313: where does the return address go on mail — *Return address*
Should the **return address** be of a different state or country, the **mail** may be routed through that location for ease of **return**.

#1314: when did steven adler play for guns and roses — *Steven Adler*
During the 2000s, **Adler** was the drummer of the band **Adler's** Appetite , and since **2012**, he has held the same position in the band **Adler**.

#1327: what year was gulf war — *Gulf War*
The **war** is also known under other **names** , such as the Persian **Gulf War**, First **Gulf War**, **Gulf War** I, or the First Iraq **War**, before the **term** "Iraq **War**" became identified instead with the 2003 Iraq **War** (also referred to in the U.S. as "**Operation** Iraqi Freedom").

#1335: what year was elvis born — *Elvis Presley*
In **1968**, after seven **years** away from the stage, he returned to live performance in a celebrated comeback television **special** that led to an extended **Las Vegas** concert residency and a string of profitable tours.

1339: where did columbus really land in 1492? — *Christopher Columbus*
Though **Columbus** was not the first European explorer to reach the Americas (having been preceded by the Norse expedition led by Leif Ericson in the 11th century), **Columbus's voyages** led to the first **lasting** European contact with the Americas, inaugurating a period of European exploration, conquest, and colonization that **lasted** for several **centuries**.

1340: what is a brindle boxer — *Boxer (dog)*
Boxers were **first** exhibited in a dog show for St. Bernards in Munich in 1895, the **first Boxer** club being founded the next year.

1342: who invented the internet — *History of the Internet*
Since the mid-1990s, the **Internet** has had a revolutionary impact on culture and **commerce**, including the rise of near-instant communication by electronic mail , instant messaging , Voice over **Internet** Protocol (VoIP) "phone calls", two-way interactive video calls , and the World **Wide Web** with its discussion forums , blogs , social networking , and online shopping sites.

#1345: when does the royal standard fly? — *Royal Standard of the United Kingdom*
If the ancient **Royal Standard** of Scotland is flying above Holyrood **Palace** or Balmoral Castle , instead of the **Royal Standard** of the United Kingdom used in Scotland, it also indicates that the Queen is not in residence.

1354: what the atmosphere on mercury — *Atmosphere of Mercury*
The existence of a **atmosphere** had been contentious before 1974, although by that time a consensus had formed that **Mercury**, like the Moon , lacked any substantial **atmosphere**.

#1356: what year did the last monkeys go into space — *Monkeys in space*
Before humans **went** into **space** , several animals were **launched** into **space** , including numerous **monkeys** , so that scientists could investigate the biological effects of **space** travel .

1364: what religion is westminster abbey — *Westminster Abbey*
Westminster Abbey is a collegiate **church** governed by the **Dean** and Chapter of **Westminster** , as established by Royal charter of Queen Elizabeth I in 1560, which **created** it as the Collegiate **Church** of St Peter **Westminster** and a Royal Peculiar under the personal jurisdiction of the Sovereign.

#1365: when did coca cola first come out — *Coca-Cola*
The most common of these is Diet **Coke** , with others including **Caffeine-Free Coca-Cola** , Diet **Coke Caffeine-Free**, **Coca-Cola** Cherry , **Coca-Cola** Zero , **Coca-Cola** Vanilla , and special versions with lemon, **lime** or coffee.

1366: what role do ombudsman play in the swedish government? — *Ombudsman*
In many countries where the **ombudsman's** remit extends beyond dealing with alleged maladministration to promoting and protecting **human** rights, the **ombudsman** is recognised as the national **human** rights institution .

1372: who played batman in dark knight — *The Dark Knight (film)*
Considered one of the **best films** of the 2000s and one of the **best superhero films** ever made , the **film** received **highly** positive **reviews** and **set** numerous records during its **theatrical** run.

1384: who won the super in xl1 — *Super Bowl XLI*
Super Bowl XLI was an **American football** game between the **American Football Conference** (AFC) **champion** Indianapolis **Colts** and the **National Football Conference** (NFC) **champion** Chicago Bears to decide the **National Football** League (NFL) **champion** for the 2006 **season** .

1389: what are the parts of plant stems? — *Plant stem*
In most **plants stems** are located above the soil surface but some **plants** have underground **stems** .

#1394: what year did martin luther king die — *Martin Luther King, Jr.*
Martin Luther King, **Jr**. **Day** was established as a U.**S**. federal **holiday** in **1986**.

1400: what spanish speaking countries have the most world cup titles — *FIFA World Cup*
The **World Cup** is the **world's** most widely **viewed** sporting event; an **estimated** 715.1 **million** people **watched** the final match of the 2006 **FIFA World Cup held** in **Germany** .

1402: who composed the Singapore national anthem — *Majulah Singapura*
Singaporeans are especially encouraged to sing the **national anthem** on occasions of **national** celebration or **national** significance such as at the **National** Day Parade , at **National** Day observance ceremonies conducted by educational institutions and government departments, and at sporting events at which **Singapore** teams are participating.

1411: when does v start — *V (2009 TV series)*
V stars Morena Baccarin , Lourdes Benedicto , Morris Chestnut , Joel Gretsch , Logan Huffman , Charles Mesure , Elizabeth Mitchell , Laura Vandervoort and **Scott** Wolf , and was executive produced by **Scott** Rosenbaum , Yves Simoneau , **Scott** Peters , and Jace Hall .

1416: when did spongebob first air — *SpongeBob SquarePants*
He teamed up with several **Nickelodeon** veterans and Rocko crew members, including creative director Derek Drymon (Action League Now! , Hey Arnold! , and Rocko's Modern Life) writers and directors Sherm Cohen , and Dan Povenmire , writer Tim Hill , actor and writer Martin Olson , animation director Alan Smart (all from Rocko's Modern Life), and story **editor** Merriwether Williams (The Angry Beavers), who worked on the **series** for its **first** few seasons and switched to **SpongeBob** SquarePants in July 1999.

1418: what are d.o. of medicine — *Doctor of Osteopathic Medicine*
Many D.**O**. **physicians** attend the same graduate medical education programs as their M.**D**. counterparts, and then take M.**D**. specialty **board** exams, while other D.**O**. graduates enter osteopathic programs and take D.**O**. specialty **board** examinations.

1425: what is a medallion guarantee — *Medallion signature guarantee*
the United States and Canada , a **medallion** signature **guarantee** is a special signature **guarantee** for the transfer of securities .
A **medallion** signature **guarantee** is not the same as an acknowledgment by a notary public , in the sense that a "signature **guarantee**" is a certification by the institution that the signature is authentic, and an acknowledgment is a certification by a notary public attesting that the signer signed a document voluntarily.

1438: what is "thin film" technology — *Thin film*
Thin films are also used in dye-sensitized **solar cells** .

1443: what zones are tropical — *Tropics*
The **tropics** are also referred to as the **tropical zone** and the torrid **zone** (see geographical **zone**).

1444: what type of business is walmart — *Walmart*
Walmart remains a family-owned **business** , as the company is controlled by the Walton family , who own a 48 percent stake in **Walmart**.

1445: what part of the plant are avocados — *Avocado*
The **avocado** (Persea americana) is a **tree** native to Central Mexico, classified in the **flowering plant** family Lauraceae along with cinnamon , camphor and bay laurel .

1449: when does air bag deploy — *Airbag*
Modern vehicles may contain multiple airbag **modules** in various side and frontal locations of the **passenger** seating positions, and sensors may **deploy** one or more airbags in an impact zone at variable rates based on the type, angle and severity of impact; the airbag is designed to only inflate in moderate to severe frontal **crashes**.

1452: what are some chinese inventions — *List of Chinese inventions*
For the purposes of this list, **inventions** are regarded as technological firsts developed in **China**, and as such does not include foreign technologies which the **Chinese** acquired through contact, such as the windmill from the Middle East or the telescope from Early modern Europe .

1457: when did wwi begin — *World War I*
It was predominantly called the **World** War or the **Great** War from its occurrence until the start of **World** War II in 1939, and the First **World** War or **World** War I thereafter.

1461: what two empires fought to control afghanistan — *Afghanistan*
Afghanistan (; ;), officially the **Islamic** Republic of **Afghanistan**, is a landlocked sovereign state forming part of South **Asia** , Central **Asia** , and to some extent Western **Asia** .

1462: what triggered the civil war — *American Civil War*
The **American Civil War** (ACW), also known as the **War** between the States or **simply** the **Civil War** (see naming), was a **civil war** fought from 1861 to **1865** between the United States (the "Union" or the "**North**") and several Southern **slave** states that declared their secession and formed the Confederate States of **America** (the "Confederacy" or the "South").

1466: who owns youtube — *Youtube*
Most of the content on **YouTube** has been uploaded by individuals, although media corporations including CBS , the BBC , Vevo , Hulu , and other organizations offer some of their material via the site, as part of the **YouTube** partnership program.

1468: what are tires made of — *Tire*
Metal **tires** are still used on locomotives and railcars , and solid rubber (or other polymer) **tires** are still used in various non-automotive applications, such as some casters , carts , lawnmowers , and wheelbarrows .

1469: who killed general warren in bunker hill — *Joseph Warren*
Warren had been commissioned a Major **General** in the colony's militia shortly before the June 17, 1775 Battle of **Bunker Hill** .
His death, immortalized in John Trumbull 's painting, The Death of **General Warren** at the Battle of **Bunker's Hill**, June 17, 1775, galvanized the rebel forces, and he has been memorialized in many place names in the United States.

1472: where do women ejeculation exactly coming from — *Female ejaculation*
The **exact** source and nature of the fluid continue to be a topic of debate among medical professionals, which is also related to doubts over the existence of the G-Spot .

1473: who won the women's world cup — *FIFA Women's World Cup*
The first **Women's World Cup** tournament, named the **Women's World** Championship, was held in 1991, sixty-one years after the men's first **FIFA World Cup** tournament in 1930.

1478: when did texas become a state — *Texas*
Houston is the **largest** city in **Texas** and the fourth-largest in the United **States**, while San Antonio is the **second largest** in the **state** and seventh **largest** in the United **States**.

1484: who owns smirnoff — *Smirnoff*
Smirnoff products include vodka , **flavoured** vodka , and **malt beverages** .

1485: who made the matrix — *The Matrix*
The success of the film led to the **release** of two feature film sequels, both **written** and directed by the Wachowskis, The **Matrix** Reloaded and The **Matrix** Revolutions .

1486: what are some six sigma tools used — *Six sigma*
A **six sigma** process is one in which 99.99966% of the products manufactured are **statistically** expected to be free of defects (3.4 defects per million), although, as discussed below , this defect level corresponds to only a 4.5 **sigma** level.

1493: what is a bad beat in poker — *Bad beat*
There is no consensus among **poker** players as to what exactly constitutes a **bad beat** and often players will disagree about whether a particular hand was a **bad beat**.

1495: what part of beef are rouladen cut from? — *Rouladen*
Rouladen (or Rinderroulade, singular: **roulade**) is a German meat **roulade** usually consisting of bacon, onions, mustard and pickles wrapped in thinly sliced **beef** which is then cooked.

1497: where does cellular respiration occur — *Cellular respiration*
While the overall **reaction** is a combustion **reaction**, no single **reaction** that comprises it is a combustion **reaction**.

1500: what is a lapping machine — *Lapping*
The other form of **lapping** involves a softer **material** such as pitch or a ceramic for the **lap**, which is "charged" with the abrasive.
The first type of **lapping** (traditionally called grinding), typically involves rubbing a brittle **material** such as glass against a surface such as iron or glass itself (also known as the "**lap**" or grinding tool) with an abrasive such as aluminum oxide , jeweller's rouge , optician's rouge , emery , silicon carbide , diamond , etc., in between them.

1502: who plays ethan in my babysitter's a vampire — *My Babysitter's a Vampire (TV series)*
My **Babysitter's** a **Vampire** (French : Ma gardienne est un **vampire**) is a 2011 Canadian **television** series, based on the **television** film of the same name .

1513: what is a base SI unit — *SI base unit*
The International System of **Units** (**SI**) defines seven **units** of measure as a basic set from which all other **SI units** are derived .

1516: when did ww1 end? — *World War I*
It was predominantly called the **World War** or the Great **War** from its occurrence until the start of **World War** II in 1939, and the First **World War** or **World War** I thereafter.

1519: what state is area code 419 — *Area codes 419 and 567*
The main **area code**, **419**, was created as one of the original **area codes** in October 1947; the overlay **area code** 567 was created on January 1, 2002.

1521: what is 9/11 bombings — *September 11 attacks*
The **fourth plane**, United **Airlines Flight** 93 , was targeted at the United States Capitol in Washington, D.C., but **crashed** into a field near Shanksville, Pennsylvania , after its **passengers** tried to overcome the **hijackers**.

1528: when did charles dickens live — *Charles dickens*
Born in **Portsmouth**, England, **Dickens** left **school** to work in a factory after his father was thrown into **debtors' prison**.

1529: when did the civil rights movement begin — *Civil rights movement*
Civil rights movements ranging from the global LGBT **rights movement** to the global Women's **rights movement** to various racial minority **rights movements** around the **world** continue.

1532: who shot john lennon? — *Death of John Lennon*
John Lennon was an English musician who gained worldwide fame as one of the founders of The Beatles , for his subsequent solo career, and for his political activism and pacifism .

1534: what state is new orleans in — *New Orleans*
The **New Orleans metropolitan area** (**New** Orleans–Metairie–Kenner **Metropolitan** Statistical **Area**) had a population of 1,167,764 in 2010 and was the 46th largest in the United **States**.

1535: what where the most important factors that led to the defeat of the democrates in 1968? — *United States presidential election, 1968*
This was the last **election** in which **New** York had the most **votes** in the **electoral** college (43 votes).

1537: when did kurt cobain kill himself — *Kurt Cobain*
Kurt Donald **Cobain** (February 20, 1967 – **April** 5, 1994) was an American musician and **artist** , best known as the lead singer , guitarist and primary songwriter of the grunge band Nirvana .

1538: who starred in the original true grit — *True Grit (1969 film)*
True Grit is a **1969** American western **film** written by Marguerite Roberts and directed by Henry Hathaway .

1539: what is a millwright worker — *Millwright*
Modern **millwrights work** with steel and other materials in addition to wood and must often combine the skills of several skilled trades in order to successfully fabricate industrial machinery or to assemble machines from pre-fabricated parts.

1548: who created the tourbillon movement? — *Tourbillon*
Stührling **Tourbillon Movement** ().

1549: what is an .msi file — *Windows Installer*
The installation information, and often the **files** themselves, are packaged in installation packages , loosely relational databases structured as COM Structured Storages and **commonly** known as "**MSI files**", from their default **file** extension .

1550: what year did mexico gain independence from spain — *Mexican War of Independence*
The movement, which became known as the **Mexican** War of **Independence**, was led by Mexican-born **Spaniards** , Mestizos and Amerindians who sought **independence** from **Spain**.

1553: what is a Four Lokos — *Four Loko*
The name "**Four**" is derived from the original energy drink's **four** main ingredients: alcohol , caffeine , taurine , and guarana .

1555: what is a fret on a guitar — *Fret*
The neck of a **guitar** showing the nut (in the background, coloured white) and first four metal **frets**.

1559: what year was smokey the bear invented — *Smokey Bear*
Smokey Bear (often called **Smokey** the **Bear** or **Smokey**) is a mascot of the United States Forest Service created to educate the public about the dangers of forest fires .

1561: who said tv is a vast wasteland — *Newton N. Minow*
His speech referring to television as a "**vast wasteland**" is cited even as the speech has passed its 50th anniversary.

1575: who discovered the 2 moons of mars,Phobos and Deimos — *Moons of Mars*
It is possible that **Mars** may have **moons** smaller than 50 - 100 meters and a dust ring between **Phobos** and **Deimos** may be present but none have been **discovered**.

1580: where does angela davis currently work — *Angela Davis*
Angela Yvonne **Davis** (born January 26, 1944) is an American political activist , scholar, and author.

1587: who killed robert kennedy — *Assassination of Robert F. Kennedy*
The **assassination** of **Robert** Francis "Bobby" **Kennedy** , a United States Senator and brother of **assassinated** President John Fitzgerald "Jack" **Kennedy** , took place shortly after midnight on June 5, 1968, in Los Angeles , California , during the campaign season for the United States Presidential election, 1968 .

1589: when Harry met Sally case — *When Harry Met Sally...*
When **Harry Met** Sally... is a 1989 American romantic comedy film written by Nora Ephron and directed by Rob Reiner . When **Harry Met** Sally... grossed a total of US$92 .8 million in North America.

1590: who owns hamburger helper — *Hamburger Helper*
The **Hamburger Helper** mascot is the "**Helping** Hand", an anthropomorphic animated , four fingered left-hand glove, which appears in the product's television commercials and on the packages.

1591: what makes a dwarf planet — *Dwarf planet*
The exclusion of **dwarf planets** from the roster of **planets** by the IAU has been both praised and criticized; it was said to be the "right decision" by Mike Brown , who **discovered** and other new **dwarf planets**, but has been rejected by Alan Stern , who had coined the term **dwarf planet** in 1990.

1593: when did jack lalanne die — *Jack LaLanne*
On the occasion of **LaLanne's death**, Schwarzenegger **credited LaLanne** for being "an apostle for fitness" by inspiring "billions all over the world to **live** healthier lives," and, as governor of California , had earlier placed him on his Governor's Council on Physical Fitness.

1598: who made the original care bears — *Care Bears*
The **Care Bears appeared** in their own TV **specials called** The **Care Bears** in the Land Without Feelings (1983) and The **Care Bears** Battle the Freeze Machine (1984).

1601: when do solar eclipses happen? — *Solar eclipse*
Earth's orbit is called the **ecliptic plane** as the Moon's orbit must cross this **plane** in order for an **eclipse** (both **solar** as well as lunar) to occur.

1606: what latitude is tropic of cancer — *Tropic of cancer*
The **Tropic** of **Cancer**, also referred to as the Northern **tropic**, is the circle of **latitude** on the Earth that marks the most northerly position at which the Sun may appear directly overhead at its zenith .

1611: what state is the capital in — *List of capitals in the United States*
In addition, each of the 50 U.S. **states** and the five principal territories of the United **States** maintains its own **capital**.

1612: what are square diamonds called? — *Princess cut*
The **square** princess cut **diamond** is usually slightly cheaper than round brilliant cut **diamonds** of the same carat weight because it retains about 80% of the rough **diamond**, as opposed to the round brilliant which retains only about 50% of the rough.

1615: where scottsdale? — *Scottsdale, Arizona*
Scottsdale is bordered to the west by **Phoenix** and **Paradise** Valley , to the north by Carefree , to the south by **Tempe** , and to the east by Fountain Hills and the Salt River Pima-Maricopa Indian Community .

1616: who has brad pitt dated — *Brad Pitt*
In addition, **Pitt** owns a **production** company, Plan B Entertainment , whose **productions** include The Departed (2006), which won the Academy **Award** for Best Picture , and Moneyball, which garnered a Best Picture nomination.

1617: what is a dogs classification — *Dog*
MtDNA evidence shows an evolutionary split between the **modern dog's** lineage and the **modern** wolf's lineage around 100,000 years ago but, , the oldest fossil specimens genetically linked to the **modern dog's** lineage date to approximately 33,000-36,000 years ago.

1622: what nationality is wendy williams — *Wendy Williams*
She hosts a syndicated television **talk show**, The **Wendy Williams Show** .

1626: what president was theodore roosevelt — *Theodore Roosevelt*
Roosevelt was 42 years old when sworn in as **President** of the **United** States in 1901, making him the youngest **president** ever; he beat out the youngest elected **president**, John F. Kennedy , by only one year.

1628: who starred in webster — *Webster (TV series)*
Webster is an American situation comedy that aired on ABC from **September** 16, 1983 until May 8, 1987, and in first-run **syndication** from **September** 21, 1987 until **March** 10, 1989.

1633: what is a monarch to a monarchy — *Monarchy*
The **monarchs** of **Cambodia**, **Japan**, Jordan, Malaysia and **Morocco** "**reign**, but do not rule" although there is considerable variation in the amount of authority they wield.

1637: what is an "N.M? — *Newton metre*
The symbolic form is N **m** or N·m,.

1652: who made hubble telescope — *Hubble Space Telescope*
Although not the first space **telescope**, **Hubble** is one of the **largest** and most versatile, and is well known as both a vital research **tool** and a **public** relations boon for astronomy .

1659: when did andrea doria sink — *SS Andrea Doria*
Named after the 16th-century **Genoese** admiral **Andrea Doria** , the ship had a gross register tonnage of 29,100 and a capacity of about 1,200 passengers and 500 crew.

1673: who wrote second corinthians — *Second Epistle to the Corinthians*
The **Second** Epistle to the **Corinthians**, often referred to as **Second Corinthians** (and **written** as 2 Corinthians), is the eighth book of the New Testament of the Bible.

1675: who first synthesized heroin — *Heroin*

Mexican cartels are also known to **produce** a third type of illicit **heroin**, commonly called black tar , which results from a simplified, **quicker** synthesis procedure and contains a high percentage of **morphine** derivates other than **heroin**, such as 6-monoacetylmorphine (6-MAM).

1678: what is a wiki platform *Wiki*
Wikis are powered by **wiki** software .

1686: who plays dumbledore in harry potter 6 *Albus Dumbledore*
Dumbledore is **portrayed** by Richard **Harris** in the **film** adaptions of **Harry Potter** and the Philosopher's **Stone** and **Harry Potter** and the **Chamber** of Secrets .

1689: what type of land is savannah *Savanna*
Typical tropical savanna in Northern Australia demonstrating the high tree density and regular spacing characteristic of many **savannas**.

1695: who set the world record for women for high jump *High jump*
Javier Sotomayor (Cuba) is the current men's **record** holder with a **jump** of **set** in 1993, the longest standing **record** in the history of the men's **high jump**.

1701: where does ground pepper come from *Black pepper*
Peppercorns, and the **ground pepper derived** from them, may be described simply as **pepper**, or more precisely as black **pepper** (cooked and **dried** unripe fruit), green **pepper** (**dried** unripe fruit) and **white pepper** (**dried** ripe seeds).

1707: what type of ecosystem does stingrays live in *Stingray*
They are classified in the suborder Myliobatoidei of the order Myliobatiformes , and consist of eight families: Hexatrygonidae (sixgill stingray), Plesiobatidae (deep water stingray), Urolophidae (stingarees), Urotrygonidae (round rays), Dasyatidae (whiptail stingrays), Potamotrygonidae (river stingrays), Gymnuridae (butterfly rays), and Myliobatidae (eagle rays).

1712: what religion is primary in africa? *Religion in Africa*
Religious distribution in **Africa**
Religion in **Africa** is multifaceted and has been a major influence on art , culture and philosophy .

1713: who won season 2 of project runway *Project Runway (season 2)*
Project Runway Season 2 was the second **season** of Bravo 's successful **Project Runway** , a reality competition for fashion designers.

1715: what is a mule in coins *Mule (coin)*
The name derives from the **mule** , the hybrid offspring of a horse and a donkey, due to such a **coin** having two sides intended for different **coins**, much as a **mule** has parents of two different species.

1721: what is an arc in a story plot *Story arc*
Although **story arcs** have existed for decades, the term "**story arc**" was coined in 1988 in relation to the television series Wiseguy , and was quickly adapted for other uses.

1727: who sang what a wonderful world *What a Wonderful World*
"What a **Wonderful World**" is a song written by Bob Thiele (as "George Douglas") and George David Weiss .

1728: what month is the president inaugurated *United States presidential inauguration*
(Prior to the Twentieth Amendment , the **date** was **March 4**, the **day** of the year on which the Constitution of the United States first took effect in 1789; the last **inauguration** to take place on the older **date** was Franklin D. Roosevelt 's first one on **March 4**, 1933.)

1751: what are club seats *Club seating*
Club level **seating** is a special section of **seating** in modern sports stadiums .

1766: what is a redshirt freshman football player *Redshirt (college sports)*
For example, a coach may choose to **redshirt** a **player** who is then referred to as a **redshirt freshman** or simply a **redshirt**.

♯1776: what year did South Africa become a team in rugby *South Africa national rugby union team*
The **South Africa** national **rugby union team** (known as the Springboks) represents **South Africa** in **rugby union** .

1782: who did john f kennedy run against? *John F. Kennedy*
After military **service** as **commander** of the Motor Torpedo Boats PT-109 and PT-59 during World War **II** in the South **Pacific** , **Kennedy** represented Massachusetts' 11th congressional district in the U.S. House of Representatives from 1947 to **1953** as a Democrat .

1785: what season is dexter on *Dexter (TV series)*
Set in Miami , the show's first **season** was largely based on the **novel** Darkly Dreaming **Dexter** , the first of the **Dexter** series **novels** by Jeff **Lindsay** .

1786: What are context effects of memory? *Context-dependent memory*
However, the research **literature** on context-dependent **memory describes** a number of different types of contextual information that may affect recall such as environmental context-dependent **memory**, state-dependent learning , **cognitive** context-dependent **memory** and **mood-congruent memory**.

1789: who won the 1998 world cup *1998 FIFA World Cup*
The **1998 FIFA World Cup** was the 16th **FIFA World Cup** , the **world** championship for men's national association football teams.

1791: where did the vietnamese settle in america *Vietnamese American*
A **Vietnamese** American () is an American of **Vietnamese descent**.

♯1793: when did hitler kill himself *Death of Adolf Hitler*
Accounts differ as to the cause of **death**; one that he **died** by poison only and another that he **died** by a self-inflicted gunshot, while biting down on a cyanide capsule.

♯1796: when did world war 2 end *World War II*
With an invasion of the **Japanese** archipelago imminent, and the Soviet **Union** having declared **war** on **Japan** by invading **Manchuria** , **Japan** surrendered on 15 August **1945**, **ending** the **war** in Asia and cementing the total **victory** of the Allies over the Axis.

1805: who won the 1967 nba championship *1967 NBA Finals*
The **1967 NBA** World **Championship** Series was the **championship** series of the 1966-67 National Basketball Association season , and was the conclusion of the **1967 NBA** Playoffs .

1806: who make airbus *Airbus*
Airbus began as a consortium of **aerospace manufacturers**, **Airbus** Industrie.

1809: who invented the television *History of television*
As **electronic camera** and **display tubes** were perfected, electromechanical **television** gave way to all-electronic systems in nearly all **applications**.

1812: what are the side effects for lyme disease *Lyme disease*
Lyme disease, **Lyme** borreliosis is an infectious **disease** caused by at least three species of bacteria belonging to the genus Borrelia .

1814: who wrote the song for star wars *Star Wars music*
Additionally, music for **Star Wars**: The Clone **Wars** was **written** by Kevin Kiner, and further music has been composed for **Star Wars** video **games** and **works** in other media.

1822: where does ray lamontagne live *Ray lamontagne*
Raymond "**Ray**" Charles Jack **LaMontagne** (; born June 18, 1973) is an American singer-songwriter .

1827: who made facebook *Facebook*
Critics , such as **Facebook** Detox , state that **Facebook** has turned into a national obsession in the United States, resulting in vast amounts of time lost and encouraging narcissism.

1831: where do you find iodine *Iodine*
Because of this function, radioisotopes of **iodine** are concentrated in the thyroid gland along with nonradioactive **iodine**.

1833: what kind of people are on the show skins *Skins (UK TV series)*
Other ventures to expand the brand have **included** a failed North **American** adaptation , which **aired** on MTV in **2011** but it was **cancelled** after one season after **advertisers** abandoned the **series** in response to low **ratings** and the **significant** controversy which arose over its depiction of **teen** sexuality.

1834: who played the drums in the band cream back in 1968 *Cream (band)*
Cream were inducted into the Rock and Roll **Hall** of **Fame** in 1993.

1837: what president made decision to buy louisiana *Louisiana Purchase*
The **Louisiana Purchase** ("**Sale** of **Louisiana**") was the acquisition by the United States of America in 1803 of of France 's claim to the territory of **Louisiana** .

1838: who sang cool jerk *Cool Jerk*
In the feature film Home Alone 2: Lost in New York , Uncle Frank (Gerry Bamman) **sings** "**Cool Jerk**" in the shower.

1839: what is a league in the sea *League (unit)*
A **league** is a **unit** of length (or, rarely, area).

♯1845: when will ie9 be released *Internet Explorer 9*
The system requirements for **Internet** Explorer **9** are **Windows** 7 , **Windows** Server **2008** R2 , **Windows** Vista Service Pack 2 or **Windows** Server **2008** SP2 with the **Platform** Update .

♯1849: when did classification of races begin *Race (human classification)*
While **biologists** sometimes **use** the **concept** of **race** to make **distinctions** among fuzzy sets of traits, others in the **scientific** community suggest that the idea of **race** often is used in a naive or simplistic way, i.e. that among humans, **race** has no taxonomic significance: all living humans belong to the same **species**, **Homo sapiens** and **subspecies**, **Homo sapiens sapiens**.

1855: what states has the electric chair *Electric chair*
Execution by **electrocution**, usually performed using an **electric chair**, is an execution **method** originating in the United **States** in which the condemned person is strapped to a specially built wooden **chair** and **electrocuted** through **electrodes** placed on the body.

1857: what are private labels *Private label*
McBride plc is an example of a European based provider of **private label** household and **personal** care products.

♯1866: When did the New Deal start *New Deal*
The **New Deal** produced a **political** realignment, making the Democratic **Party** the majority (as well as the **party** that held the White House for seven out of nine Presidential terms from **1933** to 1969), with its base in **liberal ideas**, the white **South**, traditional **Democrats**, big city machines, and the **newly** empowered labor unions and ethnic minorities.

♯1873: when did daylight savings time start *Daylight saving time*
Daylight saving time (DST)—also summer **time** in British **English** — is the practice of advancing clocks during the lighter months so that evenings have more **daylight** and mornings have less.

1874: What region of France is Montargis in? *Montargis*
Montargis is a commune in the Loiret department in north-central **France** on the Loing river.

♯1877: what year lord of rings made? *The Lord of the Rings*
The enduring **popularity** of The **Lord** of the **Rings** has led to numerous references in **popular culture**, the founding of many societies by fans of **Tolkien's works** , and the **publication** of many **books** about **Tolkien** and his **works**.

1878: what time will the world end on may 21 *2011 end times prediction*
The **2011 end times** prediction made by American Christian radio **host** Harold Camping stated that the Rapture and Judgment Day would take **place** on May **21**, **2011**, and that the **end** of the **world** would take **place** five months later on **October 21**, **2011**.

1885: what is file based system *File system*
Some **file systems** are "virtual", in that the "**files**" supplied are **computed** on request (e.g. procfs) or are merely a mapping into a **different file system** used as a backing store.

1887: What political conflicts marked the presidency of William Howard Taft? *William Howard Taft*
William Howard Taft (September 15, 1857 – **March** 8, 1930) was the 27th **President** of the **United States** (1909–1913) and later the tenth **Chief Justice** of the **United States** (1921–1930).

1890: what temperature is a salt ice bath *Cooling bath*
Cooling **baths** are generally one of two types: (a) a cold fluid (particularly liquid nitrogen , water , or even air) — but most commonly the term refers to (b) a mixture of 3 components: (1) a cooling agent (such as dry **ice** or water **ice**); (b) a liquid 'carrier' (such as liquid water, ethylene glycol, acetone, etc.), which transfers heat between the **bath** and the vessel; ; and (c) an additive to depress the melting-point of the solid/liquid system.

♯1893: when slavery abolished *Slavery in the United States*
However, by 1804, all **states north** of the Mason and Dixon Line had either **abolished slavery** outright or **passed laws** for the gradual **abolition** of **slavery**.

1905: what are circumpolar constellations *Circumpolar constellation*
In the northern hemisphere, we will always be able to see stars and **constellations** in the northern **circumpolar** sky, while in the southern hemisphere, we will always be able to see stars and **constellations** in the southern **circumpolar** sky.

1906: where do sesame seeds come from *Sesame*
The world's **largest** exporter of **sesame seeds** was India, and Japan the **largest** importer.

1912: who owned kansas before it became a state *Kansas*
When officially opened to **settlement** by the U.S. **government** in 1854, abolitionist Free-Staters from New England and pro-slavery **settlers** from neighboring **Missouri** rushed to the territory to determine if **Kansas** would become a free **state** or a **slave state**.

1915: where did the olmecs come from *Olmec*
Among other "firsts", the **Olmec** appeared to practice ritual bloodletting and played the **Mesoamerican** ballgame , hallmarks of nearly all subsequent **Mesoamerican** societies.

1918: what states allow same sex marriage *Same-sex marriage in the United States*
The Defense of **Marriage** Act (DOMA), enacted in 1996, prevents the **federal** government from **recognizing same-sex marriages** and allows each **state** to refuse recognition of **same-sex marriages** performed in other **states**.

1923: who won the most nba championships *List of NBA players with most championships*
Saul **won** consecutive **championships** with the Rochester Royals and the Minneapolis Lakers in the 1950s, while Kerr **won** consecutive **championships** with the Bulls and the Spurs in the 1990s,
Horry **won** seven **championships** with the Houston Rockets , the Los Angeles Lakers and the San Antonio Spurs , while Salley **won** four **championships** with the Detroit Pistons , the Bulls and the Lakers.

♯1925: when did the word fuck begin *Fuck*
In modern usage, **fuck** and its derivatives (such as **fucker** and **fucking**) can be used in the position of a noun , a verb , an adjective or an adverb .

1926: what are the 7 continents *Continent*
Depending on the convention and model , some **continents** may be consolidated or subdivided: for example, Eurasia is most often subdivided into Europe and Asia (red shades), while North and South America are sometimes recognized as one American **continent** (green shades).

1930: what state is milwaukee in *Milwaukee*
Known for its brewing traditions, major new additions to the city include the **Milwaukee** Riverwalk , the Delta Center (formerly "Frontier Airlines Center"), Miller Park , an internationally renowned addition to the **Milwaukee** Art Museum , **Milwaukee** Repertory Theater , and Pier **Wisconsin** , as well as major **renovations** to the U.S. Cellular Arena .

♯1935: what years was the 18th century *18th century*
To **historians** who expand the **century** to include larger **historical** movements, the "long" **18th century** may run from the Glorious **Revolution** of 1688 to the battle of Waterloo in 1815 or even later.

♯1937: what year was the eiffel tower made *Eiffel Tower*
The **Eiffel Tower** (,) is an **iron lattice tower** located on the Champ de Mars in Paris , **named** after the **engineer** Gustave **Eiffel** , whose company designed and built the **tower**.

♯1942: when monopoly came out *History of the board game Monopoly*
Also in the 1970s, Professor Ralph Anspach , who had himself published a board game intended to illustrate the principles of both **monopolies** and trust busting , fought Parker Brothers and its then parent company, General Mills , over the trademarks of the **Monopoly** board game.

1954: who passed no child left behind *No Child Left Behind Act*
The No **Child Left** Behind **Act** of **2001** (NCLB) is a United States **Act** of Congress that is a **reauthorization** of the **Elementary** and Secondary Education **Act** , which included Title I , the government's flagship aid program for disadvantaged students.

1956: what are the charges against Casey Anthony *Death of Caylee Anthony*
Caylee Marie Anthony (**August** 9, 2005 – **2008**) was a two-year-old American girl who **lived** in Orlando, Florida with her **mother**, **Casey Marie Anthony**, and her **maternal** grandparents, **George** and **Cindy Anthony**.

1965: where did world war 1 start *World War I*
It was predominantly called the **World War** or the **Great War** from its occurrence until the **start** of **World War II** in 1939, and the **First World War** or **World War** I thereafter.

1966: what species is a spider *Spider*
It now appears that the spiral orb **web** may be one of the earliest forms, and **spiders** that produce **tangled** cobwebs are more abundant and diverse than **orb-web spiders** .

♯1968: when can you use a defibulator *Defibrillation*
Some external units, known as **automated** external defibrillators (AEDs), **automate** the diagnosis of treatable rhythms, meaning that lay responders or bystanders are able to **use** them successfully with little, or in some cases no training at all.

1969: who created the cato institute *Cato Institute*
According to the 2011 **Global** Go To Think **Tank** Index, **Cato** is the 6th most influential US based think **tank** , ranking 3rd in Economic Policy and 2nd in Social Policy.

♯1977: when does a demand curve shift? *Demand curve*
The **demand curve** for all consumers together follows from the **demand curve** of every individual consumer: the individual **demands** at each price are added together.

1992: Where Elephants Live *Elephant*
Traditionally, two species are recognised, the **African elephant** (Loxodonta africana) and the **Asian elephant** (**Elephas** maximus), although some evidence **suggests** that **African** bush **elephants** and **African** forest **elephants** are separate species (L. africana and L. cyclotis respectively).

1997: what are the quad muscles *Quadriceps femoris muscle*

The **quadriceps** femoris (Latin for "four-headed **muscle** of the femur "), also called simply the **quadriceps**, **quadriceps** extensor, **quads**, is a large **muscle** group that includes the four prevailing **muscles** on the front of the thigh .

#2005: when was james madison in the house of representatives *James Madison*

James Madison, **Jr**. (**March** 16, 1751 (O.S. **March** 5) – June 28, 1836) was an American statesman and political theorist, the fourth **President** of the United States (1809–1817).

2007: What is and where is hydraulic fluid found and used for *Hydraulic fluid*

Hydraulic fluids, also called **hydraulic** liquids, are the medium by which power is transferred in **hydraulic** machinery .

2008: Where Are Mahindra Tractors Made *Mahindra Tractors*

Mahindra Tractors, the farm equipment division of **Mahindra** & **Mahindra** , builds and sources **tractors** that are sold worldwide across six continents.

2012: what is middle class in the us *American middle class*

One of the first major studies of the **middle class** in **America** was White Collar: The **American Middle Classes** , published in 1951 by sociologist C. Wright Wright_Mills Mills Wright_Mills .

2013: what is the scientific name of the eastern tiger salamander? *Tiger Salamander*

The proper common **name** is the **eastern tiger salamander**, to differentiate it from other closely related species.

2020: what was the steelworkers strike *Steel strike of 1919*

The **strike** began on September 21, 1919, and **collapsed** on January 8, 1920.

#2022: When was the first Mary Poppins book written *Mary Poppins*

Mary Poppins is the title character of a **series** of children's **books written** by P. L. **Travers** .

2024: what is sims language *Simlish*

Initially, inspired by the Native American code talkers of World War II , **Sims** creator Will Wright and **language** expert Marc Gimbel suggested experimenting with the Navajo **language** to create Simlish.

2037: who is elizabeth from general hospital who are the boys fathers *Elizabeth Webber*

Elizabeth is part of two supercouple pairings, Lucky Spencer and **Elizabeth Webber** and **Jason** Morgan and **Elizabeth Webber** .

2038: what is reagan known for *Presidency of Ronald Reagan*

The United **States presidency** of Ronald **Reagan** , also **known** as the **Reagan** administration, was a Republican administration headed by Ronald **Reagan** from **January 20**, 1981, to **January 20**, 1989.

2040: what is evoked otoacoustic emissions *Otoacoustic emission*

Broadly speaking, there are two types of **otoacoustic emissions**: spontaneous **otoacoustic emissions** (SOAEs), which can occur without external stimulation, and **evoked otoacoustic emissions** (EOAEs), which require an **evoking** stimulus .

2041: what is chep pallet *CHEP*

CHEP offers wooden and plastic **pallets**, small display **pallets**, crates and IBC containers.

#2046: when was scooby doo created *Scooby-Doo*

This Saturday morning **cartoon** series featured four teenagers— Fred Jones , Daphne Blake , Velma Dinkley , and Norville "Shaggy" Rogers — and their talking brown Great Dane **dog** named Scooby-Doo , who solve **mysteries** involving supposedly **supernatural** creatures through a series of antics and missteps.

2054: who is the CEO OF FACEBOOK *Mark Zuckerberg*

In 2007, at the age of twenty-three years, Zuckerberg became a billionaire as a result of **Facebook** and the number of **Facebook** users worldwide reached a total of one billion in 2012.

2057: What Is the Largest Whale *Blue whale*

The blue **whale** (Balaenoptera musculus) is a marine mammal belonging to the suborder of baleen **whales** (called Mysticeti).

2058: What Is Range in Math *Range (mathematics)*

This is the current usage for **range** in computer science .
Sometimes "**range**" refers to the codomain and sometimes to the image.
Some books say that **range** of this function is its codomain, the set of all real numbers, reflecting that the function is real-valued.
Other books say that the **range** is the function's image, the set of non-negative real numbers, reflecting that a number can be the output of this function if and only if it is a non-negative real number.
In this case, the larger set containing the **range** is called the codomain.
In mathematics , the **range** of a function refers to either the codomain or the image of the function, depending upon usage.

2060: who is on the hundred dollar bill *United States one hundred-dollar bill*

The **United States** one **hundred-dollar bill** ($100) is a **denomination** of **United States** currency .

2061: what is the ingredient in mustard *Mustard (condiment)*

The other four **mustards** pictured are a simple table **mustard** with turmeric coloring (center left), a Bavarian sweet **mustard** (center-right), a Dijon **mustard** (lower-left), and a coarse French **mustard** made mainly from black **mustard** seeds (lower-right).

2062: who are the members of the climax blues band? *Climax Blues Band*

The **Climax Blues Band** (originally known as the **Climax** Chicago **Blues Band**) were formed in Stafford , England in 1968.

#2064: when was the web invented *World Wide Web*

With a **web browser** , one can view **web** pages that may contain text, images, videos, and other multimedia , and navigate between them via hyperlinks .

2066: what is kathmandu known for *Kathmandu*

Kathmandu's sister **cities** (Lalitpur Patan) and Bhaktapur are integral to **Kathmandu's** cultural heritage, **tourism** industry, and **economy**; therefore UNESCO's World Heritage Site lists all three cities' monuments and attractions together under one heading, "**Kathmandu** Valley- UNESCO World Heritage Site ".

2071: where is mark sanchez from *Mark Sanchez*

Despite a subpar performance, **Sanchez** led the Jets to the AFC **Championship Game** , a losing effort to the Indianapolis Colts , becoming the fourth rookie quarterback in NFL history to win his first playoff **game** and the second to win two playoff **games**.

2080: where is cougar town filmed *Cougar Town*

The **show** was created by Bill **Lawrence** and Kevin **Biegel** and is **produced** by Doozer (Lawrence's company) and Coquette **Productions** in association with **ABC Studios** .

2089: where is green bay packers from *Green Bay Packers*

The **Green Bay Packers** have won **13** league **championships** (more than any other team in the NFL), including nine NFL **championships** prior to the **Super** Bowl era and four **Super** Bowl victories —in **1967** (**Super** Bowl I), 1968 (**Super** Bowl II), 1997 (**Super** Bowl XXXI) and 2011 (**Super** Bowl XLV).

#2091: when was the world of coca cola built *World of Coca-Cola*

Its well-known advertising as well as a host of entertainment areas and attractions, and is **located** in Atlanta , Georgia (where the company's headquarters are **located**) at Pemberton Place (named in honor of John Pemberton , the inventor of Coca-Cola).

2096: what is the highest mountain in america and where is is located? *Mount McKinley*

Mount McKinley (also known as Denali taken from the Inuit Koyukon Athabaskan language meaning "The Great One") is the **highest mountain** peak in the United States and in North **America** with a summit **elevation** of above sea level .

2097: what was the first honda car *Honda S600*

Available as a roadster – bearing strong resemblance to the **Honda** S500 – and as a fastback coupé – introduced in March 1965 – the S600 was the **first Honda** available in two trim levels.

2100: what is the population of san francisco *San Francisco*

San Francisco (), **officially** the City and **County** of **San Francisco**, is the leading financial and cultural center of Northern California and the **San Francisco** Bay Area .

2102: who is the writer of the beowulf poem? *Beowulf*

Beowulf (; in **Old English** or) is the conventional title of an **Old English** heroic epic **poem** consisting of 3182 alliterative long lines , set in **Scandinavia** , commonly cited as one of the most important **works** of **Anglo-Saxon literature** .

2114: what is sherlock holmes job *Sherlock Holmes*

The **character** grew tremendously in popularity with the first **series** of **short** stories in The Strand Magazine , beginning with A Scandal in Bohemia in 1891; further **series** of **short** stories and two novels **published** in **serial** form appeared between then and 1927.

2119: what is social security card used for *Social Security number*

A **Social Security** number may be obtained by applying on Form SS-5, "Application for A **Social Security** Number **Card**".

2120: what is the minimalist trend *Minimalism*

Minimalism is any design or **style** in which the simplest and fewest elements are used to create the maximum effect.

2126: who is the current Chief Justice of the U.S. supreme court? *Chief Justice of the United States*

The **Chief Justice** of the United **States** is the **head** of the United **States** federal **court** system (the judicial branch of the federal government of the United **States**) and the **chief** judge of the **Supreme Court** of the United **States** .

#2129: when was the american labor union formed *Labor unions in the United States*

The economist Joseph Stiglitz has asserted that, "Strong **unions** have **helped** to reduce inequality, whereas weaker **unions** have made it easier for CEOs , sometimes **working** with market forces that they have **helped** shape, to increase it."

2132: what is the largest credit union *Navy Federal Credit Union*

Navy Federal **Credit Union** (or Navy Federal) is a **credit union** headquartered in Vienna, Virginia , chartered and regulated under the authority of the National **Credit Union** Administration (NCUA) of the U.S. federal government.

2135: what is definition of psychotic *Psychosis*

Despite this, "**psychosis**" is generally given to noticeable deficits in normal behavior (negative signs) and more commonly to diverse **types** of **hallucinations** or **delusional beliefs** (e.**g**. grandiosity, **delusions** of persecution).

2136: Who is General Grievous of Star Wars *General Grievous*

General Grievous is a fictional **character** in the **Star Wars** universe .

2139: where is scotty mccreery from *Scotty McCreery*

He also released a **Christmas album**, **Christmas** with **Scotty McCreery** , which has been certified gold.

2151: what is my resting heart rate at age 24 *Heart rate*

Tachycardia is **defined** as a **resting heart rate** above 100 **bpm**, though persistent **rest rates** between 80-100 **bpm**, mainly if they are present during sleep, may be signs of hyperthyroidism or anemia (see below).

2152: what is the name for an old horse-drawn vehicle *Carriage*

Working **vehicles** such as the (four-wheeled) wagon and (**two-wheeled**) **cart** share important parts of the **history** of the carriage, as is the fast (**two-wheeled**) chariot .

2154: where was the tsunami in 2005 *2004 Indian Ocean earthquake and tsunami*

The resulting **tsunami** was given various **names**, including the **2004 Indian** Ocean **tsunami**, **South** Asian **tsunami**, **Indonesian tsunami**, and the Boxing Day **tsunami**.

2155: what is the source of geothermal energy *Geothermal energy*

The **geothermal** gradient , which is the difference in temperature between the core of the **planet** and its **surface**, drives a continuous conduction of thermal **energy** in the form of heat from the core to the **surface**.

2158: who was ho chi minh in vietnam war *Ho chi minh*

Hồ Chí Minh (**Northern Vietnamese** pronunciation : , Southern **Vietnamese** pronunciation :); 19 May 1890 – 2 **September** 1969), born **Nguyễn Sinh** Cung and also known as **Nguyễn** Tất Thành and **Nguyễn** Ái Quốc, was a **Vietnamese** communist **revolutionary** leader who was prime minister (1945–1955) and president (1945–**1969**) of the **Democratic Republic** of **Vietnam** (**North** Vietnam).

2167: who was charged with murder after the massacre at My lai *My Lai Massacre*

The My **Lai Massacre** (, ; , , or) was the Vietnam **War mass murder** of between 347 and 504 unarmed civilians in South Vietnam on **March** 16, 1968, by United States Army soldiers of "Charlie" Company of 1st **Battalion** , 20th **Infantry** Regiment , 11th Brigade of the Americal Division .

2173: what is metal music about *Heavy metal music*

Underground scenes produced an array of more extreme, aggressive styles: thrash **metal** broke into the mainstream with bands such as Metallica , Megadeth , **Slayer** , and Anthrax , while other styles of the most extreme **subgenres** of **metal** like death **metal** and black **metal** remain subcultural phenomena.

2175: what is the title of Hobbes main work *Thomas Hobbes*

Thomas **Hobbes** of Malmesbury (5 April 1588 – 4 December 1679), in some older texts Thomas **Hobbs** of Malmsbury, was an English philosopher, best known today for his **work** on political philosophy .

2188: what is the cabin pressure of us airlines *Cabin pressurization*

The **cabin pressure** is regulated by the outflow **valve**.

2191: what is buttoned-down shirt? *Dress shirt*

A **shirt**, dress **shirt**, button-front, button-front **shirt**, or button-up **shirt** is a garment with a collar , a full-length opening at the front from the collar to the hem, and sleeves with cuffs .

#2192: when is administrative assistant day *Administrative Professionals' Day*

Administrative Professionals **Day** (also known as Secretaries **Day** or Admin **Day**) is an unofficial secular holiday observed in several countries to recognize the work of secretaries , **administrative assistants**, receptionists , and other **administrative** support professionals.

2204: what is steam by valve corporation *Steam (software)*

The **Steam** logo is a stylised left-side fly-crank and rod from the Walschaerts **valve** gear of a **steam** locomotive .

2209: what is civil engineering aBOUT *Civil engineering*

It is traditionally broken into several sub-disciplines including environmental **engineering** , geotechnical **engineering** , geophysics , geodesy , control **engineering** , structural **engineering** , biomechanics , nanotechnology , transportation **engineering** , earth science , atmospheric sciences , forensic **engineering** , municipal or urban **engineering** , water resources **engineering** , materials **engineering** , coastal **engineering** , surveying , and construction **engineering** .

2217: who was john f kennedy up against *United States presidential election, 1960*

This election is notable as being the first time in U.S. history that two sitting U.S. Senators (**Kennedy** and **Johnson**) were elected as **president** and **vice-president**, a phenomenon that has been repeated once, by Barack **Obama** and Joe Biden in 2008 (in both cases, the **president** was the younger, more junior senator).

2221: what is the controlled substance act known as *Controlled Substances Act*

Two **federal** agencies, the Drug Enforcement Administration and the Food and Drug Administration , determine which **substances** are added to or removed from the various **schedules**, though the **statute passed** by Congress created the initial listing, and Congress has sometimes **scheduled** other **substances** through legislation such as the Hillory J. Farias and Samantha Reid Date-Rape **Prevention Act** of 2000, which placed gamma hydroxybutyrate in **Schedule** I. Classification decisions are **required** to be made on criteria including potential for **abuse** (an undefined term), currently accepted medical use in treatment in the United States, and international treaties.

2226: what is the si unit of pressure *Pressure*

While **pressure** may be **measured** in any **unit** of force divided by any **unit** of area, the **SI unit** of **pressure** (the newton per square **metre**) is called the pascal (Pa) after the seventeenth-century philosopher and scientist Blaise Pascal .

2227: what is the great basin area *Great Basin*

It is noted for both its **arid** conditions and its **Basin** and **range** topography that varies from the North American low point at Badwater **Basin** to the highest point of the contiguous United States , less than away at the summit of Mount Whitney .

#2230: when was fdr elected as president *Franklin D. Roosevelt*

A dominant leader of the **Democratic Party** and the only **American president elected** to more than two **terms**, he built a **New** Deal Coalition that realigned **American politics** after 1932, as his domestic policies defined **American** liberalism for the middle third of the 20th century.

2231: where was ms-13 originally from *MS-13*

In the U.S., the **MS-13** has an especially heavy presence in Los Angeles County and the San Francisco Bay Area in Northern California; the Washington, D.C. metropolitan areas of Fairfax County, Virginia, Montgomery County, Maryland, and Prince George's County, Maryland; Long Island, **New** York; the Boston, Massachusetts area; Charlotte, North **Carolina**; and Houston, Texas.

2234: where was the super bowl in 1991 *Super Bowl XXV*

The Bills and their explosive no-huddle offense were making their first **Super Bowl** appearance after finishing the regular **season** with a 13-3 record, and leading the league in total points scored with 428.
Super Bowl XXV was an American football game between the American Football Conference (AFC) champion Buffalo Bills and the National Football Conference (NFC) champion New York Giants to decide the National Football League (NFL) champion for the 1990 **season** .
In advancing to their second **Super Bowl**, the Giants also posted a 13-3 regular **season** record, but with a ball-control offense and a defense that allowed a league low 211 points.

2237: who are the characters in 90210 in season 3 *90210 (season 3)*

Christmas-themed episode "Holiday Madness", hit **season** highs in all key **demos** with 2.1 in The CW's target **demo** of women 18–34, a 1.4 in **adults** 18–34 and a 1.1 in **adults** 18–49.

#2238: when was queen elizabeth ii married *Elizabeth II*

There have been times of **personal** sorrow for her which include the **death** of her **father** at 56, the assassination of **Prince Philip's** uncle, Lord Mountbatten , the breakdown of her **children's marriages** in 1992 (a year deemed her annus horribilis), the **death** in 1997 of her former daughter-in-law, Diana, **Princess** of Wales , and the **deaths** of her **mother** and sister in **2002**.

2249: who is is the singer westlife *Westlife*

Westlife sold over 50 **million** records **worldwide**, a **total** that **included** studio albums, singles, video releases, and compilation albums.

2250: what is vitamin a for *Vitamin A*

The associated acid (retinoic acid), a metabolite that can be irreversibly synthesized from **vitamin** A, has only partial **vitamin** A activity, and does not function in the retina for the visual cycle.

#2254: when is it memorial day *Memorial Day*

Memorial Day is not to be confused with Veterans **Day** ; **Memorial Day** is a **day** of remembering the men and women who died while serving, while Veterans **Day** celebrates the service of all U.S. military veterans, living or dead.

2259: what is the function of the liver *Liver*

The **liver** is necessary for survival; there is currently no way to compensate for the absence of **liver function** in the long **term**, although new **liver** dialysis techniques can be used in the short **term**.

2261: what is the formula for calcium nitrate *Calcium nitrate*

A variety of related salts are known including **calcium** ammonium **nitrate** decahydrate and **calcium** potassium **nitrate** decahydrate.

2262: where is dia de los muertos celebrated — *Day of the Dead*
The holiday has spread throughout the world: In Brazil , **Dia de** Finados is a public holiday that many Brazilians **celebrate** by visiting cemeteries and **churches**.

2263: where is the 2011 mlb all star game location — *2011 Major League Baseball All-Star Game*
The **2011 Major League** Baseball All-Star **Game** was the 82nd in-season exhibition **game** between the All-Stars of the National **League** (NL) and the American **League** (AL) ; the **leagues** composing **Major League** Baseball .

2267: where is the island New Guinea? — *New Guinea*
The Germans **annexed** the northern coast of the **eastern** half of the **island** as German **New Guinea** in their pre–World War I effort to establish themselves as a **colonial** power, whilst the south **eastern** portion was reluctantly **claimed** by Britain.

2272: what is in milk — *Milk*
New Zealand , the European Union 's 27 member states, Australia , and the **United** States are the world's largest exporters of **milk** and **milk** products.

2273: What is motorcycle speedway racing — *Motorcycle speedway*
There are now both **domestic** and international **competitions** in a number of **countries** including the **Speedway World Cup** whilst the highest overall scoring individual in the **Speedway Grand Prix** events is pronounced the **world champion**.

2289: where are the most concentration of jews living — *Jews*
Converts to Judaism , whose status as **Jews** within the **Jewish** ethnos is equal to those born into it, have been absorbed into the **Jewish** people throughout the millennia.

2293: who was the first european in the americas — *European colonization of the Americas*
In 1497, sailing from the **north** on behalf of England , John Cabot landed on the **North American coast**, and a year later, **Columbus's** third voyage **reached** the South **American coast**.

2295: where is testosterone produced — *Testosterone*
On average, in adult human males, the plasma concentration of **testosterone** is about 7–8 times as great as the concentration in adult human females' plasma, but as the metabolic consumption of **testosterone** in males is greater, the daily **production** is about 20 times greater in men.

2299: what is cu the element — *Copper*
Its **compounds** are commonly encountered as copper(**II**) **salts**, which often impart **blue** or green colors to minerals such as azurite and turquoise and have been widely used historically as pigments.

2301: what was the parthenon used for — *Parthenon*
In the 5th century AD, the **Parthenon** was converted into a **Christian church** dedicated to the **Virgin** Mary .

♯2307: when was jamestown colonized — *Jamestown, Virginia*
Historic Jamestowne , the archaeological **site** on **Jamestown** Island, is a cooperative effort by **Jamestown** National **Historic Site** (part of Colonial National **Historical** Park), and **Preservation** Virginia .

2308: what was the first year of kentucky derby — *Kentucky Derby*
The **Kentucky Derby** () is a Grade I **stakes** race for three-year-old Thoroughbreds , held annually in Louisville, **Kentucky**, United States , on the **first** Saturday in May, capping the two-week-long **Kentucky Derby** Festival .

2311: who is on the $10. bill — *United States ten-dollar bill*
The source of the face on the $**10 bill** is John Trumbull 's 1805 portrait of **Hamilton** that belongs to the portrait collection of New York City Hall .

2318: who is norah jones parents — *Norah Jones*
Norah Jones (born Geetali **Norah Jones Shankar**; **March 30**, 1979) is an **American** singer-songwriter, pianist, and actress.

2325: what is Carbon 14 dating is a type of ? — *Carbon-14*
There are three naturally occurring isotopes of **carbon** on Earth: 99% of the **carbon** is **carbon-12** , 1% is **carbon-13** , and **carbon-14** occurs in trace amounts, i.e., making up about 1 part per trillion (0.0000000001%) of the **carbon** in the atmosphere.

♯2331: when was the battle at tombstone fought — *Gunfight at the O.K. Corral*
The gunfight, believed to have lasted only about thirty seconds, was **fought** between the **outlaw** Cowboys **Billy Claiborne** , **Ike** and **Billy Clanton** , and Tom and Frank McLaury , and the opposing town **Marshal** Virgil Earp and his brothers Assistant Town **Marshal** Morgan and temporary lawman Wyatt , aided by Doc **Holliday** designated as a temporary **marshal** by Virgil.

2335: what is the population of algoma wi for 2010 — *Algoma, Wisconsin*
See also the **Town** of **Algoma** in Winnebago County, **Wisconsin** .

2349: who is Dr. JB Danquah — *J. B. Danquah*
During his political career, he was one of the primary opposition **leaders** to Ghanaian president and independence **leader** Kwame Nkrumah .

2356: where is kj 52 from — *KJ-52*
The "**KJ**" part of his **name** refers to his old rap alias, "King J. Mac," a **name** which he later described in one of his podcasts as "horribly cheesy."

2360: what is the prognosis of stomach cancer — *Stomach cancer*
Stomach cancer, or gastric **cancer**, refers to **cancer** arising from any part of the **stomach** .

2367: what is el mate — *Mate (beverage)*
"Tea-bag" type infusions of **mate** (**mate** cocido) have been on the market in Argentina, Paraguay and Uruguay for many years under such trade names as " Taragüi Vitality" in Argentina, "Pajarito" and "Kurupí" in Paraguay, and in Brazil under the name "**Mate** Leão".

2369: what is dave batista doing now — *Dave Batista*
David Michael "**Dave**" Bautista, Jr. (born January 18, 1969), is an American mixed martial artist , bodybuilder , actor, and former professional wrestler best known for his time in World Wrestling Entertainment competing under the ring name **Batista**.

2371: where is the arctic circle located on the earth — *Arctic circle*
In fact, because of atmospheric refraction and mirages , and because the sun appears as a disk and not a point, part of the midnight sun may be seen on the night of the northern summer solstice up to about 50 ′ () south of the **Arctic Circle**; similarly, on the day of the northern winter solstice , part of the sun may be seen up to about 50′ north of the **Arctic Circle**.

2375: what is the erb/heart — *Erb's point (cardiology)*
Heart valves are labeled with "B", "T", "A", and "P".First **heart** sound: caused by atrioventricular valves - Bicuspid/Mitral (B) and Tricuspid (T).
Second **heart** sound caused by semilunar valves – Aortic (A) and Pulmonary/Pulmonic (P).
Front of thorax , showing surface relations of bones , lungs (purple), pleura (blue), and **heart** (red outline).

♯2377: when was pokemon first started — *Pokémon*
The term **Pokémon**, in addition to referring to the **Pokémon franchise** itself, also collectively refers to the 649 fictional species that have made appearances in **Pokémon** media as of the **release** of the fifth generation titles **Pokémon** Black **2** and White **2** ; with the upcoming **releases** of **Pokémon** X and Y , 6 **new Pokémon** have been featured in promotions for the **games** .

2379: where are colors on stoplight — *Traffic light*
Traffic lights alternate the right of way accorded to road users by displaying lights of a standard **color** (**red**, yellow/**amber**, and **green**) following a universal **color** code .

2392: what is mincemeat made of — *Mincemeat*
Variants of **mincemeat** are found in Australia , Brittany , Canada , northern Europe , Ireland , South Africa , the **United** Kingdom and the **United** States .

2399: what was the GE building in rockefeller plaza called before. — *GE Building*
The **GE Building** is an Art Deco skyscraper that forms the centerpiece of **Rockefeller** Center in Midtown Manhattan , New York City , USA .

2400: who are the two senators of louisiana — *List of United States Senators from Louisiana*
Louisiana was admitted to the Union on April 30, 1812, and elects **senators** to Classes 2 and Class 3 .

2408: who is st patty? — *Saint Patrick's Day*
Saint Patrick's Day or the Feast of **Saint Patrick** (, "the Day of the Festival of **Patrick**") is a cultural and religious holiday celebrated on 17 March.

2427: WHAT IS THE LENGTH OF A NAUTICAL MILE — *Nautical mile*
The **nautical mile** is nearly equal to a minute of latitude on a chart, so a distance measured with a chart divider can be roughly converted to **nautical miles** using the chart's latitude scale.

2431: what is firewire used for — *IEEE 1394*
Apple first included **FireWire** in some of its 1999 models, and most Apple computers since the year 2000 have included **FireWire** ports, though, as of 2013, nothing beyond the 800 version (IEEE-1394b).

♯2432: when was srv born — *Stevie Ray Vaughan*
As the younger **brother** of **Jimmie** Vaughan , Vaughan started playing the guitar at age seven and formed several bands that **occasionally** performed in local nightclubs.

2433: what is stent surgery — *Stent*
The term may also refer to a tube used to temporarily hold such a natural conduit open to allow access for **surgery**.

2439: what is tofu made of — *Tofu*
There are many different varieties of **tofu**, including fresh **tofu** and **tofu** that has been processed in some way.

2440: Where is Bubbles the Chimp now — *Bubbles (chimpanzee)*
In 2003 the public learned that, like all captive chimpanzees, **Bubbles** had matured into a large and aggressive adult **chimp** unsuitable as a companion animal.

2441: what is muse's lead singer's name — *Muse (band)*
Muse are known for their energetic and extravagant live performances and their fusion of many music genres , including space **rock** , progressive **rock** , **alternative rock** , heavy metal , classical music and electronica .

♯2451: when was the internet started — *History of the Internet*
The **Internet** was commercialized in 1995 when **NSFNET** was decommissioned, removing the last restrictions on the use of the **Internet** to carry **commercial** traffic.

2455: where is rolling rock brewed — *Rolling Rock*
Rolling Rock is a 4.6% abv pale lager launched in 1939 by the Latrobe **Brewing** Company .

♯2456: when was Pope Benedict XVI elected? — *Pope Benedict XVI*
Prior to becoming **pope**, he was "a major figure on the **Vatican** stage for a quarter of a century" as "one of the most respected, influential and controversial members of the College of **Cardinals**"; he had an influence "second to none when it came to **setting** church priorities and directions" as one of **Pope John Paul II** 's closest confidants.

2458: Who was John Adam's children — *John Adams*
John Adams (**October** 30, 1735 (O.S. **October** 19, 1735) – July 4, 1826) was the second president of the **United** States (1797–1801), having earlier served as the **first** vice president of the **United** States .

2459: who was on the 10 dollar bill — *United States ten-dollar bill*
The $**10 bill** is the only U.S. paper currency in circulation in which the portrait faces to the left (the $100,000 **bill** featured a portrait of Woodrow Wilson facing to the left, but was used only for intra-government transactions).

2460: what is water jet propulsion — *Jetboat*
Unlike these previous **waterjet** developments, such as Campini's and the Hanley Hydrojet, Hamilton had a specific need for a **propulsion** system to operate in very-shallow **water**, and the **waterjet** proved to be the ideal solution.

2463: who is E from entourage — *Eric Murphy*
Murphy is a fictional **character** on the comedy-drama television series **Entourage** .

2472: what is puerto rico currency — *Currencies of Puerto Rico*
The Banco Español de **Puerto Rico** was renamed **Bank** of **Porto Rico** and issued bills equivalent to the **United States dollar** , creating the **Puerto** Rican **dollar**.

2486: what is endodontic dentistry — *Endodontics*
Endodontists perform a variety of procedures including **endodontic** therapy (commonly known as "root canal therapy"), **endodontic** retreatment , surgery, treating cracked teeth , and treating dental trauma .

2491: who is the junior senator of nc — *Kay Hagan*
When Hagan defeated Republican **incumbent** Elizabeth Dole in the 2008 United States **Senate** election , she became the first woman to defeat an **incumbent** woman in a **Senate** election.

2492: who is heisman trophy named after — *Heisman Trophy*
The **Heisman** Memorial **Trophy** Award (usually known colloquially as the **Heisman Trophy** or the Heisman), is awarded annually to the player deemed the most outstanding player in collegiate **football**.

2498: what is sado masochism — *Sadomasochism*
Similarly, **sexual** sadism within the context of mutual consent should not be mistaken for acts of **sexual** violence or aggression.

2501: where are Giant Panda Bears found? — *Giant Panda*
"black and white cat-foot"), also known as the **giant panda** to distinguish it from the unrelated **red panda** , is a **bear** native to central-western and south western China .

2505: what is the format of the canadian citizenship test — *Canadian Citizenship Test*
The **Canadian Citizenship Test** is a **test**, administered by **Citizenship** and Immigration **Canada** (CIC), that is required for all applicants for **Canadian citizenship** who are aged between 18 and 54 and who meet the basic requirements for **citizenship**.

2510: What is up with Kent Hovind — *Kent Hovind*
Since January 2007, **Hovind** has been serving a ten-year **prison** sentence after being convicted of 58 **federal** counts, including 12 tax offenses, one count of obstructing **federal** agents, and 45 counts of structuring **cash** transactions .

2518: where is good morning america studio — *Times Square Studios*
The **studio** is best known as the production home of ABC News ' **Good Morning America** (GMA), a **morning** news and talk program and segments for GMA on ABC News Now.

2523: what is primary medicine — *Primary care*
Such a professional can be a **primary** care physician , such as a general **practitioner** or family physician , or depending on the locality, health system organization, and patient's discretion, they may see a pharmacist , a physician assistant , a **nurse practitioner** , a **nurse** (such as in the United Kingdom), a clinical officer (such as in **parts** of Africa), or an Ayurvedic or other traditional **medicine** professional (such as in **parts** of Asia).

♯2530: when was the first super bowl — *Super Bowl*
Super Bowl XLV, played in 2011, became the most-watched American television program in history, drawing an average audience of 111 million viewers and taking over the spot held by the previous **year's Super Bowl**, which itself had taken over the #1 spot held for twenty-eight **years** by the final episode of M*A*S*H .

2532: what is the official language of america? — *Languages of the United States*
The situation is quite varied at the state and territorial levels, with some states mirroring the federal policy of **adopting** no **official language** in a **de** jure capacity, others **adopting English** alone, others **officially adopting English** as well as local **languages**, and still others **adopting** a policy of **de** facto bilingualism.

2534: where is shropshire, uk — *Shropshire*
The two **Shropshire** unitary areas (covering all of the ceremonial county), together with the authorities covering the ceremonial county of Staffordshire, comprise the "**Shropshire** and Staffordshire" NUTS 2 region .
Shropshire is one of England's most rural and sparsely populated counties, with the population density of the **Shropshire** Council area being just 91/km2 (337/sq mi).

2539: what is the social norm approach? — *Social Norms Approach*
Despite the fact that college drinking is at elevated levels, the perceived amount almost always exceeds actual behavior The **social norms approach** has shown signs of countering misperceptions, however research on **resulting changes** in behavior **resulting** from **changed** perceptions varies between mixed to conclusively nonexistent.

2545: what is an sd memory card reader — *Memory card reader*
There are three categories of **card readers** sorted by the type and quantity of the **card** slots: single **card reader** (e.g. 1x SD-only), multi **card reader** (e.g. 9-in-1) and series **card reader** (e.g. 4x **SD** only).

2547: What was Captain Ahab's Ship in the novel "Moby Dick" — *Moby-Dick*
Ishmael soon learns that **Ahab** has one purpose on this **voyage**: to seek out **Moby Dick**, a ferocious, enigmatic white sperm **whale** .

♯2563: when is halley's comet next — *Halley's Comet*
Halley's Comet or **Comet Halley** (or), officially designated 1P/**Halley**, is the best-known of the short-period **comets** and is visible from Earth every 75–76 years.

2566: what is soy made from — *Soybean*
Fat-free (**defatted**) soybean meal is a significant and cheap source of **protein** for animal feeds and many prepackaged meals ; **soy vegetable oil** is another **product** of processing the soybean crop.

2569: what is linkedin used for — *LinkedIn*
, **LinkedIn** reports more than 200 million acquired **users** in more than 200 countries and territories.

2596: what is captcha code — *CAPTCHA*
Although most **CAPTCHAs** are letter pictures randomly generated, many of them have become difficult even for a human to read , so picture **CAPTCHAs** were created in which a human is shown a simple test to show a picture of a certain animal (given few animal pictures), which is simple for a human being to process, and therefore easy to pick, while a bot cannot process and solve the question because although it can analyze the picture, it cannot easily guess the animal.

2611: who is in the group trinity 5 7 — *Trin-i-tee 5:7*
In 1998, the **group** released their first album entitled **Trin-i-tee** 5:7 which debut at #1 on the Top Gospel charts .

♯2619: when is international men's day — *International Men's Day*
International Men's Day is celebrated in over 60 countries, including Trinidad and Tobago, Jamaica, Australia, India, China, United States, Romania, Singapore, Malta, United Kingdom, South Africa, Tanzania, **Zimbabwe**, Botswana, Seychelles, **Burundi**, Hungary, Ireland, **Isle** of **Man**, Ghana, **Canada**, Denmark, Norway, Austria, **Bosnia** and Herzegovina, Ukraine, France, **Italy**, Pakistan, **Cuba**, Antigua and Barbuda, **St**. **Kitts** and **Nevis** , **St**. **Lucia** , **Grenada** and Cayman Islands , on 19 November, and global support for the celebration is broad.

♯2635: when was the patriot act enacted — *Patriot Act*

On May 26, **2011**, President **Barack Obama** used an Autopen to **sign** the **PATRIOT Sunsets Extension Act** of **2011** , a four-year **extension** of three key provisions in the USA **PATRIOT Act** while he was in France: **roving** wiretaps , **searches** of business **records** (the " **library records** provision "), and conducting **surveillance** of "**lone wolves**" — individuals suspected of terrorist-related activities not linked to terrorist groups.

#2637: when was the tacoma bridge collapse? *Tacoma Narrows Bridge*

Historically, the name "**Tacoma** Narrows **Bridge**" has applied to the original **bridge** nicknamed "Galloping **Gertie**", which opened in July 1940 but **collapsed** because of aeroelastic flutter four months later, as well as the replacement of the original **bridge** which opened in 1950 and still stands today as the **westbound** lanes of the present-day twin **bridge** complex.

2640: what is brown flax *Flax*

Flax (also **known** as common **flax** or linseed) (binomial name: Linum usitatissimum) is a member of the genus Linum in the family Linaceae .

#2669: when was the lady gaga judas song released *Judas (song)*

Gaga has performed "**Judas**" on a **number** of television **shows**, including The **Graham** Norton **Show** , Saturday Night Live , Good Morning Americas "Summer Concert Series", the French X Factor as well as on The Ellen DeGeneres **Show** .

2670: what is mpeg4 avc or sp/asp *H.264/MPEG-4 AVC*

The **ITU-T** H.264 **standard** and the ISO/IEC MPEG-4 **AVC standard** (**formally**, ISO/IEC 14496-10 – MPEG-4 **Part** 10, **Advanced** Video Coding) are jointly maintained so that they have identical technical content.

2685: what is the definition of a map *Map*

Although most commonly used to depict geography, **maps** may represent any space , real or imagined, without regard to context or scale ; e.g. brain **mapping** , DNA **mapping** and extraterrestrial **mapping**.

2688: who is on blink 182 album cover *Enema of the State*

Produced by Jerry **Finn** , the **album** was released on June 1, 1999 in the **United** States on MCA **Records** .

2691: What is the name of the late Major League old professor *Casey Stengel*

After his **major league** career began, he acquired the **nickname** "**Casey**", which originally came from the initials of his hometown ("K. C."), which evolved into "**Casey**", influenced by the wide popularity of the poem **Casey** at the **Bat** .

#2693: when was the constitution written *Constitution*

The **Constitution** of India is the longest **written constitution** of any sovereign **country** in the world, **containing** 448 articles, 12 schedules and 100 amendments, with 117,369 words in its English language version, while the **United States Constitution** is the shortest **written constitution**, at 7 articles and 27 amendments.

2697: what is the use of a sales invoice? *Invoice*

In English, the context of the term **invoice** is usually **used** to clarify its meaning, such as "We sent them an **invoice**" (they owe us money) or "We received an **invoice** from them" (we owe them money).

2698: what was the actress who played the pink power ranger *Kimberly Hart*

Kimberly is best remembered as the first **Pink Ranger** and first **Pink Ninja Ranger** from the first entry of the **franchise** Mighty Morphin **Power Rangers** .

2700: what is squash the sport *Squash (sport)*

The **game** was formerly called **squash racquets**, a reference to the "**squashable**" soft ball used in the **game** (compared with the fatter ball used in its parent **game racquets** or rackets).

2702: where is the world cup in 2010 *2010 FIFA World Cup*

The **2010 FIFA World Cup** was the 19th **FIFA World Cup** , the **world** championship for men's national association football teams.

2707: what is hosting a website *Web hosting service*

The most basic is web page and small-scale **file hosting**, where **files** can be uploaded via **File** Transfer Protocol (FTP) or a Web interface.

#2708: when was the first nfl madden game released *Madden NFL*

Madden NFL (known as John **Madden Football** before 1993) is an American **football** video **game series developed** by Electronic Arts Tiburon for **EA** Sports .

2738: what is the purpose of child support? *Child support*

The 1992 United Nations Convention on the Rights of the **Child** , a binding convention signed by every member nation of the United Nations and formally ratified by all but Somalia and the United States, declares that the upbringing and development of **children** and a standard of **living** adequate for the **children's** development is a common responsibility of both parents and a fundamental human right for **children**, and asserts that the primary responsibility to provide such for the **children** rests with their parents.

2741: where was martin luther king shot? *Martin Luther King, Jr.*

Martin Luther King, **Jr**. **Day** was established as a U.S. federal **holiday** in **1986**.

2747: who is the mayor of chicago 2011 *Chicago mayoral election, 2011*

The city of **Chicago**, Illinois held a nonpartisan **mayoral** election on Tuesday, February 22, **2011**.

#2749: when is the wv state fair *State Fair of West Virginia*

The **State Fair** of **West Virginia** is an **annual state fair** for **West Virginia** .

#2755: when was the state of utah established *Utah*

Approximately 80% of **Utah's** 2,817,222 people live along the Wasatch Front , centering on Salt **Lake** City , leaving vast expanses of the **state** nearly uninhabited and making the **population** the sixth most urbanized in the U.S. **Utah** is bordered by **Colorado** on the east, Wyoming on the northeast, Idaho on the north, Arizona on the south, and Nevada on the west.

#2762: when is world war hulk movie come *World War Hulk*

The **series** consists of five main issues titled **World War Hulk**, with Greg Pak as writer and John Romita, Jr. as penciller , and three other limited **series** : **World War Hulk**: Front Line, **World War Hulk**: Gamma Corps , and **World War Hulk**: X-Men .

2763: WHERE WAS JOHN WAYNE BORN *John Wayne*

Marion Mitchell Morrison (**born** Marion **Robert** Morrison; May 26, 1907 – June 11, 1979), better known by his stage name **John Wayne**, was an American film **actor**, director and producer.

2765: what is the difference between alpha lipoic acid and lipoic acid *Lipoic acid*

The carbon atom at C6 is chiral and the molecule exists as two enantiomers (R)-(+)-lipoic **acid** (RLA) and (S)-(-)-lipoic **acid** (SLA) and as a racemic mixture (R/S)-lipoic **acid** (R/S-LA).

2767: what is the organic layer in an amoled screen *OLED*

An OLED (**organic light-emitting diode**) is a **light-emitting diode** (LED) in which the emissive **electroluminescent layer** is a film of **organic** compound which **emits light** in response to an **electric** current.

2783: what is bourbon made of *Bourbon whiskey*

The name of the spirit derives from its historical association with an area known as Old **Bourbon**, around what is now **Bourbon** County, Kentucky (which, in turn, was named after the French House of **Bourbon** royal family).

2802: where is La Palma africa *La Palma*

It was as a result of his visit to **La Palma** and Tenerife where he visited the **Las** Cañadas and Taburiente calderas, that the Spanish word for cauldron - " Caldera " - was introduced into the English language geological vocabulary.

2810: where was hillary clinton born *Hillary Rodham Clinton*

That **election** marked the **first** time an American **First Lady** had **run** for public office; **Clinton** was also the **first** female senator to represent the **state**.

#2819: When was 27th amendment proposed *Twenty-seventh Amendment to the United States Constitution*

The Twenty-seventh **Amendment** (**Amendment** XXVII) prohibits any law that increases or decreases the salary of members of the Congress from taking effect until the start of the next set of terms of office for Representatives.

2821: what is singapore's currency *Singapore dollar*

The Monetary Authority of **Singapore** and the **Brunei Currency** and Monetary Board still maintain the historic exchangeability of their two **currencies**, the **Singaporean** dollar and the **Brunei** dollar , respectively.

2828: where was paul revere born *Paul Revere*

Following the **war**, **Revere** returned to his **silversmith** trade and used the profits from his expanding business to finance his work in iron **casting**, bronze bell and cannon **casting**, and the forging of copper bolts and spikes.

2841: what is the type of democracy in which all citizens have the right to make government decisions *Democracy*

One **form** of **democracy** is **direct democracy** , in which all eligible **citizens** have **direct** and **active** participation in the **decision making** of the **government**.

2847: what is ratchet from ratchet and clank *Ratchet & Clank*

The franchise was created and developed by Insomniac Games and published by Sony Computer Entertainment for many different **PlayStation** consoles, such as **PlayStation** 2 and **PlayStation** 3 with the exclusion of Size Matters and Secret Agent **Clank** , which were developed by High Impact Games for the **PlayStation** Portable .

2848: where was the first real roller coaster *roller coaster*

Some **roller coasters**, notably wild mouse **roller coasters** , run with single cars.

2849: what is in a will *Will and testament*

In the strictest sense, a "will" has historically been limited to **real property** while "**testament**" applies only to dispositions of **personal property** (thus giving rise to the popular title of the document as "Last Will and **Testament**"), though this distinction is seldom observed today.

2862: what is bracketing in photography *Bracketing*

Autobracketing is automatic **bracketing** by using a setting on the camera to take several **bracketed** shots (in contrast to the **photographer** altering the settings by hand between each shot).

2869: What is the disk defragmenter tool *Defragmentation*

Defragmentation is advantageous and relevant to file systems on electromechanical **disk** drives .

2870: what is level of agreement mean *Service-level agreement*

As an example, **internet** service providers will commonly include service **level agreements** within the **terms** of their **contracts** with customers to define the level(s) of service being sold in plain language **terms**.

2883: who is the actor that plays harry potter *List of Harry Potter cast members*

She **announced** that she could never have dreamt of the talented **actors** and **actresses** who have appeared in all the **Harry Potter** movies.

2884: what is the concept of "wellness" ? *Wellness (alternative medicine)*

Wellness grew as a popular **concept** starting in the 19th century, just as the middle class began emerging in the industrialized world, and a time when a newly prosperous public had the time and the resources to pursue **wellness** and other forms of self-improvement.

2899: what is the defensive line in football called *American football positions*

This has resulted in the development of three "platoons" of players, the offense (the team with the ball, who is **trying** to score), the **defense** (the team **trying** to prevent the other team from scoring, and to take the ball from them), and the special teams (who play in kicking situations).

2903: what is the measurements of saturn's moons *Moons of Saturn*

They include the seven major satellites, four small **moons** which exist in a **trojan** orbit with larger **moons**, two mutually co-orbital **moons** and two **moons** which act as shepherds of **Saturn's F Ring** .

2907: who are the girls from the bad girls club? *Bad Girls Club*

If a "**bad girl**" breaks a rule, she is evicted from the show and, if it is early in the **season**, replaced by a new "**bad girl**".

2916: what was bernie madoff selling *Bernard Madoff*

Madoff founded the Wall Street firm **Bernard** L. **Madoff** Investment Securities **LLC** in 1960, and was its chairman until his arrest on December 11, 2008.

2923: who is inventor of the radio *Invention of radio*

Others, notably Guglielmo **Marconi** , were concerned with practical improvements and the commercial application of **radio** to wireless telegraphy .

2925: what is scurvy disease *Scurvy*

Other eponyms for **scurvy** include Moeller's **disease** and Cheadle's **disease**.

2931: what is google openid ***OpenID***

An extension to the standard (the **OpenID** Attribute **Exchange**) facilitates the transfer of user attributes, such as name and gender, from the **OpenID** identity provider to the relying party (each relying party may request a different set of attributes, depending on its requirements).

2933: what is atherosclerotic heart disease *Atherosclerosis*

Arteriosclerosis is a general **term** describing any hardening (and loss of elasticity) of medium or large **arteries** (from the Greek arteria , meaning **artery**, and , meaning hardening); arteriolosclerosis is any hardening (and loss of elasticity) of **arterioles** (small arteries); **atherosclerosis** is a hardening of an **artery** specifically due to an atheromatous plaque.

2935: what was the name of the uss john kennedy before *USS John F. Kennedy (CV-67)*

The ship is **named** after the 35th President of the United **States**, **John** F. **Kennedy** , and is nicknamed "Big John."

2939: what were 3 important things that douglas MacArthur did? *Douglas MacArthur*

General of the **Army Douglas MacArthur** (26 January 18805 April **1964**) was an American **general** and field marshal of the Philippine **Army** who was Chief of Staff of the **United States Army** during the 1930s and played a prominent **role** in the Pacific theater during World War II.

2946: what is santa's wife's name *Mrs. Claus*

Mrs. Claus is the **wife** of **Santa** Claus , the Christmas gift-bringer in North American and European Christmas tradition .

#2949: when was Saddle Creek founded *Saddle Creek Records*

Saddle Creek first appeared in print on a show flyer, offering to "Spend an evening with **Saddle Creek**" (later to be the title of the label's DVD.)

2958: Who is the highest scoring NBA player *List of National Basketball Association season scoring leaders*

The National **Basketball** Association 's (**NBA**) **scoring** title is awarded to the **player** with the **highest** points per game average in a given season.

2961: what is direct marketing channel *Direct Marketing*

Direct marketing relies on being able to **address** the members of a target **market** .

2974: what is the name of the family who own the biltmore estates in nc *Biltmore Estate*

Biltmore Estate is a large private **estate** and tourist attraction in Asheville , North Carolina .

2979: where are the kidneys in your body *Kidney*

Common clinical conditions involving the **kidney** include the nephritic and nephrotic syndromes , renal cysts , acute **kidney** injury , **chronic kidney** disease , urinary tract infection , nephrolithiasis , and urinary tract obstruction .

2983: what is white chocolate made of *White chocolate*

White chocolate is a **chocolate** derivative.

2986: what is the latest season of psych *Psych*

USA renewed the series for a seventh **season**, to include 16 episodes, on January 10, **2012**, and again for an **eighth season**, to include eight episodes, on December 19, **2012**.

2990: what is the main component of vaccines *vaccine*

Vaccines may be prophylactic (example: to **prevent** or ameliorate the effects of a future infection by any natural or "wild" pathogen), or therapeutic (e.g. **vaccines** against cancer are also being investigated; see cancer **vaccine**).

3005: what is in the alpine mountain systems *Alps*

The Alps are one of the great **mountain** range **systems** of Europe stretching approximately across eight **Alpine** countries from Austria and Slovenia in the **east**, Switzerland , Liechtenstein , Germany , **France** to the **west** and Italy and Monaco to the **south**.

3008: who is mary matalin married to *Mary Matalin*

She appears in the award-winning documentary film Boogie Man: The Lee Atwater Story and also played herself, opposite her husband, James **Carville**, John Slattery , and **Mary** McCormack in the short **lived** HBO series K Street .

3010: where is basque spoken *Basque language*

These **provinces** and many **areas** of **Navarre** are heavily populated by ethnic Basques, but **Basque** was, at least until the 1990s, not **spoken** as a native language in most of Álava, western parts of Biscay and central and southern **areas** of **Navarre** , either because it had been replaced by **Spanish** along the centuries, in some **areas**, or because it had never been **spoken** there, in other **areas**.

3020: where is dear john filmed at *Dear John (2010 film)*

Dear John is a **2010** American romantic drama - war **film** starring Amanda Seyfried and Channing Tatum .

3034: what is melissa and joey about *Melissa & Joey*

Melissa & **Joey** is an **ABC Family** original television series starring **Melissa** Joan Hart and **Joey** Lawrence .

3039: what is the average american income *Household income in the United States*

Income includes wages and salaries, unemployment insurance, disability payments, child support payments received, regular rental receipts, as well as any personal business, investment, or other kinds of **income** received routinely.
As households tend to share a similar economic context, the use of household **income** remains among the most widely accepted measures of **income**.

#3042: When was Apple Computer founded *Apple Inc.*

Apple Inc., formerly **Apple Computer**, Inc., is an American **multinational corporation headquartered** in Cupertino , California that designs, **develops**, and **sells** consumer **electronics** , **computer software** and personal **computers**.

5 Approximate word clustering in selected Germanic languages

In this section, we present the approximate clustering algorithms for three representative Germanic languages: Danish, Dutch and German. All these three languages underwent orthographic reforms at certain points in the 20th century. We devise algorithms that accommodate to both the new and old orthographies.

It is worth mentioning that the three aforementioned Germanic languages contain some morphological features not found in English (an atypical Germanic language):

- Both Danish and German have highly inflected nouns and adjectives, and the same is true for the written standard of pre-1946 Dutch.
- Both Dutch and German verbs conjugate in the persons, and both these languages have verbs with separable prefixes.
- All these three languages allow/require a compound to be written as a single word without spaces or hyphens.

Up to minor changes, these descriptions for Dutch extend to Afrikaans (a variety of Dutch spoken in South Africa), while the features of Danish are also found in Norwegian and Swedish (two Scandinavian languages that share the same ancestor with Danish). Therefore, the algorithms presented here can be generalized to these languages with minimal modifications. We note, however, that Old Norse (the ancestral language of modern Scandinavian languages) and its conservative modern descendants (Faroese and Icelandic) have much higher morphological complexity, not yet fully captured by our current algorithms.

Like English, the Germanic strong verbs conjugate in a wide variety of "irregular" patterns. Our algorithms address the majority of these strong verbs, but do not cover all the irregularities in verb conjugation.

We will not impose the "heuristic capitalisation" procedure (Definition 4.1) on any of the non-English languages treated in this document. Accordingly, all the words in a vocabulary list must be reduced to lowercase form before they are submitted to any one of our modified Porter stemming algorithm for languages other than English.

5.1 Modified Porter stemming algorithm for Danish

Some adverbs, articles, prepositions and pronouns occur frequently in Danish texts, while conveying very little information in their own right. They are regarded as stop words.

Definition 5.1 (Danish stop words)**.** If a word belongs to the following list[49]:

> *ad*, *af*, *al*, *aldrig*, *alle*, *allerede*, *alligevel*, *alt*, *altfor*, *altid*, *anden*, *andet*, *andre*, *andres*, *at*, *atter*, *bag*, *bagved*, *bare*, *blandt*, *blev*, *blevet*, *bliv*, *blive*, *bliver*, *blot*, *bort*, *borte*, *burde*, *burdet*, *bør*, *både*, *baade*, *da*, *de*, *dem*, *den*, *denne*, *dennelunde*, *dens*, *der*, *deres*, *derfor*, *derfra*, *derhenne*, *derpå*, *derpaa*, *dersom*, *dertil*, *des*, *desto*, *det*, *dets*, *dette*, *dig*, *din*, *dine*, *disse*, *dit*, *dog*, *du*, *efter*, *efterhånden*, *efterhaanden*, *eftersom*, *egen*, *eget*, *egne*, *eller*, *en*, *én*, *end*, *endnu*, *endvidere*, *ene*, *eneste*, *enhver*, *enhvert*, *ens*, *enten*, *er*, *et*, *ethvert*, *flere*, *flest*, *fleste*, *for*, *foran*, *forbi*, *fordi*, *forudsat*, *fra*, *før*, *førend*, *ganske*, *gennem*, *gjorde*, *gjort*, *gør*, *gøre*, *ham*, *han*, *hans*, *har*, *havde*, *have*, *heller*, *hellere*, *hen*, *hende*, *hendes*, *henne*, *her*, *herfra*, *herhen*, *herhenne*, *hermed*, *hertil*, *hinanden*, *hinsides*, *hos*, *hun*, *hvad*, *hvadenten*, *hvem*, *hver*, *hverken*, *hvert*, *hvilke*, *hvilken*, *hvilket*, *hvis*, *hvor*, *hvordan*, *hvorfor*, *hvorhen*, *hvorhenne*, *hvormed*, *hvorvidt*, *i*, *iblandt*, *idet*, *iflad*, *ifølge*, *igen*, *igennem*, *ihvorvel*, *ikke*, *imellem*, *imens*, *imod*, *ind*, *inde*, *inden*, *indtil*, *ingen*, *intet*, *ja*, *jeg*, *jer*, *jeres*, *jo*, *kan*, *kun*, *kunde*, *kunne*, *kunnet*, *langs*, *lidt*, *ligesom*, *ligeså*, *ligesaa*, *man*, *mange*, *mangen*, *med*, *medens*, *megen*, *meget*, *mellem*, *men*, *mens*, *mere*, *mest*, *mig*, *min*, *mine*, *mit*, *mod*, *må*, *maa*, *måske*, *maaske*, *måtte*, *maatte*, *måttet*, *maattet*, *ned*, *nede*, *nej*, *nemlig*, *nogen*, *noget*, *nogle*, *nok*, *nu*, *nærmlig*, *nær*, *nærved*, *næsten*, *når*, *naar*, *nårsomhelst*, *naarsomhelst*, *ofte*, *oftere*, *oftest*, *ofteste*, *og*, *også*, *ogsaa*, *om*, *omend*, *omkring*, *omme*, *op*, *oppe*, *os*, *osv*, *over*, *overfor*, *ovre*, *på*, *paa*, *selv*, *selvom*, *siden*, *sidst*, *sidste*, *sig*, *sikken*, *sin*, *sine*, *sit*, *skal*, *skulde*, *skulle*, *snart*, *som*, *sommetider*, *straks*, *så*, *saa*, *sådan*, *saadan*, *sådanne*, *saadanne*, *sådant*, *saadant*, *såfremt*, *saafremt*, *således*, *saaledes*, *såvel*, *saavel*, *thi*, *til*, *tilbage*, *tit*, *trods*, *uagtet*, *ud*, *ude*, *uden*, *under*, *undertiden*, *undtagen*, *var*, *vi*, *vil*, *vilde*, *ville*, *villet*, *vor*, *vore*, *vores*, *vort*, *vær*, *være*, *været*,

then we consider it a Danish stop word. All the Danish stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

[49] Our list of Danish stop words is based on `snowball.tartarus.org/algorithms/danish/stop.txt`, together with forms derived from substitutions *å* → *aa* (accommodating to pre-1948 Danish orthography).

5.1.1 Effective spelling and essential root

Algorithm 5.2 (Danish effective spelling)**.** *Set* **V** = (*a*|*e*|*i*|*o*|*u*). *For a Danish word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Convert to lowercase, and replace*[50]

dug	*dvikl*	*eliza***X**~	*fattig*	*formue*	*forster*	*fætter*	*hoved*	*jone*	*klæd*
δuγ	*dβikl*	*ελiζ*	*πpuρrg*	*foρτṽ*	*φoρστρ*	*ṽefφe*	*ηεaδ*	*γoṽe*	*κλaδd*

kone	*langsom*	*læk*~	*mund*	*mål*	*ordn*~	*peg*	*rolig*	*sidde*~	*sider*
koṽε	*σλoωm*	*δλak*	*μuṽδ*	*σμaλ*	*ωordn*	*ποιṽτ*	*ρωλig*	*σiτe*	*πaγer*

sl(*og*|*å*)	*soldat*	*stolt*	*sukk*	*sum*~	*sund*	*synd*	*sød*	*tur*	*uværdig*~	*miss*	*s*(*a*|*i*)*d*
ζληa	*σoλδaτ*	*στolτζ*	*σukk*	*σuμm*	*γσuṽδ*	*τσyṽ*	*σωiτ*	*τuρσ*	*ωoθλσ*	*zμiσ*	*σiτe*

(2) *Replace*

’	$_{\hat{\chi}\notin}$(*u*)*fuld***X**	**X**$^{\in}$(*g*|*l*|*v*)*ent*	*ee*	*gods*~	*guv*	*hær*	*hår*	*janu*~	*kon*~	*kær*~	*land*	*lind*	*mis*$_{\hat{\chi}\notin}$(*t*)~
∅	$\hat{\chi}$	**X***enτ*	*eê*	*γoδσ*	*γuv*	*ηoστ*	*χaρ*	*jnu*	*con*	*kær*	*lanδ*	*λind*	*μsχ̂*

mæssig	*part*	*præ*~	*ros*~	*sol*~	*sonet*~	*ss*	*uge*	*varm*~	$\hat{\chi}$*værdig***X**	*yv*	*ø*~
∅	*parτ*	*πρæ*	*ρos*	*σoλ*	*σonet*	*β*	*ωeκ*	*varμ*	$\hat{\chi}$	*oj*	*zxo*

(*liden*|*mindre*|*mindst*|*mindste*|*smaa*|*små*)	*ro*(∅|*en*)	*seen*(∅|*de*)	~$_{\hat{\chi}\notin}$(*a*|*e*|*i*|*o*|*u*)*tes*	~*isme*(∅|*en*)	~*ist*(∅|*en*|*er*|*erne*)(∅|*s*)
lille	*ρωλig*	*σσaw*	$\hat{\chi}$*es*	*e*	*e*

(3) *Replace*

f(*a*|*æ*)*d*	(*here*|*herre*)~	(*ond*|*værre*|*værst*)~	**X**$^{\in}$(*tagg*|*tigd*|*tigg*)~
faδ	*zherre*	*onxved*	*b***X**

X$^{\in}$*lege*(∅|*d*|*r*|*t*)~	*ald*~	*b*(*o*|*ø*)*g*~	*bild*	*bill*~	*bland*~	*bord*~	*bragt*~	*bu*	*bøde*~	*dam*~	*dans*~	*dig*~	*dum*~
*zz***X**	*gamle*	*kbog*	*βild*	*tbill*	*mbland*	*ebord*	*bring*	*βu*	*fbode*	*ldam*	*danzts*	*δich*	*sdum*

dyd~	*d*(*å*|*aa*)*r*~	*fald*~	*far*~	*for*~	*frugt*~	*først*~	*før*~	*glim*~	*god*~	*grad*~	*grød*~	*gud*~	*gæst*~	*hils*~
vdyd	*δcar*	*efald*	*rfar*	*vfor*	*fruct*	*1st*	*zxfor*	*γlim*	*γod*	*kgrad*	*rgrød*	*hgud*	*hgast*	*ghils*

hind~	*hum*	*hund*~	*hus*	*hør*~	*iv*~	*kald*~	*kend*~	*kold*~	*komp*~	*kone*~	*kun*~	*kør*~	*lav*~	*luk*~	*lys*~	*løb*~
χind	*huμ*	*dχund*	*huσs*	*xχor*	*xiv*	*ckald*	*κend*	*κold*	*comp*	*wkone*	*gkun*	*dkor*	*λav*	*λok*	*xlys*	*rlob*

mind~	*mini*~	*mørk*~	*m*(*å*|*aa*)*d*~	*ny*(∅|*e*|*ere*|*est*|*et*|*t*)~	*nå*~	*rede*~	*r*(*y*|*ø*)*g*~	*sag*$_{\hat{\chi}\notin}$(*d*|*t*)	*se*~	*sir*
μind	*μini*	*dmork*	*wmad*	*wnu*	*rna*	*ρede*	*ρyg*	*zsag*$\hat{\chi}$	*zse*	*σiρ*

skænd~	*smal*~	*smul*~	*som*~	*svær*	*syg*~	*søg*	*søn*	*til*~	*tro*(∅|*ede*|*en*|*ens*|*et*|*r*|*s*)~	*tung*~
zskand	*zmal*	*xmul*	*zom*	*dfsvar*	*σyg*	*soeek*	*ζon*	*zil*	*gθtro*	*wtung*

ty~	*undre*~	*v*(*å*|*aa*)*gn*~	*yngre*~	(*å*|*aa*)*b*~	(*å*|*aa*)*r*~	(*gik*|*g*(*å*|*aa*)*r*)	*bed*(*re*|*st*|*ste*)	*f*(*å*|*aa*)(∅|*et*|*r*)	*fik*
θy	*wundre*	*wvagn*	*ung*	*oabp*	*jår*	*ga*	*γod*	*zfa*	*zfa*

mr	*mrs*	*mød*(∅|*t*|*te*|*tes*)	*møde*(∅|*r*|*rne*|*rnes*|*rs*|*s*|*t*|*ts*)	*sag*	*sagtens*	*se*(∅|*r*|*s*|*t*)
zηeρρe	*zφuρ*	*μod*	*μod*	*zsag*	*zzsagtens*	*σσaw*

s(*å*|*aa*)(∅|*s*)	*ti*	*ved*	*æld*(*re*|*st*|*ste*)	~*st*(*od*|(*å*|*aa*)(∅|*elig*(∅|*e*|*t*)|*et*|*r*))
σσaw	*10ti*	*vid*	*gammel*	*stand*

~**X**$^{\in}$(*br*|*m*)(*oder*|*or*)(∅|*en*|*ens*|*s*)	~**X**$^{\in}$(*br*|*m*)*ødre*(∅|*ne*|*nes*|*s*)	~*far*(∅|*en*|*ens*|*s*)	~*l*(*å*|*aa*)
X*oδ*	**X***oδ*	*faδ*	*ligger*

(4) *Do* $st_{\hat{\chi}\notin}(o|\emptyset)\sim\to szt\hat{\chi}$, $vi\to vzi$, $y\to u$, $æ\to a$, $ø\to o$, $(å|aa)\to a$ *and undouble characters (i.e. do* $mm\to m$ *etc.);*

(5) *If the result so far is an empty string, stop here. Otherwise, break down the result into* $\hat{\sigma}_1\hat{\sigma}_2$*, where* $\ell(\hat{\sigma}_1)=1$*, and work on* $\hat{\sigma}_2$ *in four sequential steps:*

(5.1) *Do* $el\to l$, $\hat{e}\to e$, $riv\to irv$;

(5.2) *Do* $_{\hat{\chi}\notin}\mathbf{V}re\to\hat{\chi}er$;

[50] To avoid confusion of Greek nu with Latin vee, we write the former as $\tilde{\nu}$ in substitution rules.

(5.3) *Do inter→inτer;*

(5.4) *Call the result so far as* $\hat{\sigma}'_2$. *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

Definition 5.3 (Danish protected range)**.** Set $\mathbf{V} = (a|e|i|o|u)$, $\mathbf{V}^*_m = (a|i|o|u)_m$ and $\mathbf{C}_{m_0} = \overline{(a|e|i|o|u)}_{m_0}$. Let $\hat{\sigma}$ be the effective spelling of a Danish word, its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \max\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is determined by two non-negative integers $\lambda_1(\hat{\sigma})$ and $\lambda_2(\hat{\sigma})$ specified through the following procedures:

- Look for the string pattern $(\varnothing|ab|af|be|in|\pi\rho a)\mathbf{C}_{m_0}(e|\mathbf{V}^*_m)\overline{(a|e|i|o|r|u)}_{[0,1]}\sim$ in the string $\hat{\sigma}$;[51]
- If the string pattern above is found, the last position occupied by such a string defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Do $\sim(\varnothing|t)(\varnothing|s) \to \varnothing$ on $\hat{\sigma}$, and call the result $\hat{\sigma}'$;
- Look for the pattern $(a|o|s|\beta|t|u)$ in the string $\hat{\sigma}'$;
- If a letter in the pattern above is found, the last position occupied by such a letter defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = 0$.

□

Algorithm 5.4 (Danish essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Danish word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$.

(2) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*[52]

(2.1) *Do* $\sim ne(\varnothing|s) \to \varnothing$*;*

(2.2) *Do* $\sim e(d|e|n|r|s|t)_{m_0} \to \varnothing$.

The result after these two steps of operations is called $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$*, before doing* $\sim s(\varnothing|t) \to \varnothing$*;*

(4) *Do* $\sim^{\hat{\chi}\in}(b|d|g|k)t \to \hat{\chi}$, $\sim er \to e$, $\sim gd \to g$.

Example 5.4.1. In many cases, the essential root of a Danish word is already a very close approximation to its etymological stem. For instance, we may consider the inflected forms of two Danish nouns *datter* "daughter" and *mand* "man", as tabulated below:

$\hat{\sigma}$	$\mathrm{EffSpell}(\hat{\sigma})$	$\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\sigma}))$
datter	*dater*	*dat*
datteren	*dateren*	*dat*
datterens	*daterens*	*dat*
datters	*daters*	*dat*
døtre	*doter*	*dot*
døtrene	*doterne*	*dot*
døtrenes	*doternes*	*dot*
døtres	*doters*	*dot*

$\hat{\sigma}$	$\mathrm{EffSpell}(\hat{\sigma})$	$\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\sigma}))$
mand	*mand*	*mand*
manden	*manden*	*mand*
mandens	*mandens*	*mand*
mands	*mands*	*mand*
mænd	*mand*	*mand*
mændene	*mandene*	*mand*
mændenes	*mandenes*	*mand*
mænds	*mands*	*mand*

The declensions of *datter* "daughter" exhibit vowel alternations in their essential roots. This phenomenon, together with the vowel alternations in Danish irregular verbs, will be handled in §5.1.2.

5.1.2 Admissible mutation and approximate clustering

The Danish vowel blotting mechanism is slightly different from the English counterpart (Algorithm 4.9).

Algorithm 5.5 (Danish vowel blotting)**.** *Set* $\mathbf{V} = (a|e|i|o|u)$ *and* $\mathbf{C}_{m_0} = \overline{(a|e|i|o|u)}_{m_0}$. *For a token string* $\hat{\sigma}$*, its blotted form* $\mathrm{BlotV}_1(\hat{\sigma})$ *is constructed as follows:*

- *If the string pattern* $(\varnothing|ab|af|be)\mathbf{C}_{m_0}\mathbf{V}\sim$ *can be found in the string* $\hat{\sigma}$*, then the last position occupied by such a pattern is replaced by the letter "a".*
- *Otherwise, leave the string* $\hat{\sigma}$ *intact.*

[51] See Definition 3.3 for the multiplicity notation $(\hat{\sigma}_1|\cdots|\hat{\sigma}_n)_{[0,1]}$.

[52] In other words, the core algorithm for essential root extraction runs as follows: keep the last "strong" vowel *a*, *ï*, *o* or *u* in non-final position, plus one subsequent letter; delete final *a*; erase the final appearance of *e* and all the letters thereafter.

Example 5.5.1. As a continuation of Example 5.4.1, we point out that $\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\sigma})))$ evaluates to *dat* (resp. *mand*) for all the declensions of *datter* (resp. *mand*). If we consider $\text{BlotV}_1(\text{EffSpell}(\hat{\sigma}))$ instead, the output is identical to $\text{EffSpell}(\hat{\sigma})$ for every declined form of *mand*, but involves a substitution $o \to a$ for the case of *datter*.

Example 5.5.2. A few more examples involving inflected forms of the Danish irregular verb *finde* "find":

$\hat{\sigma}$	*fandt*	*find*	*finder*	*funden*	*fundet*	*fundne*
$\text{EffSpell}(\hat{\sigma})$	*fandt*	*find*	*finder*	*funden*	*fundet*	*fundne*
$\text{BlotV}_1(\text{EffSpell}(\hat{\sigma}))$	*fandt*	*fand*	*fander*	*fanden*	*fandet*	*fandne*
$\text{EssRoot}(\text{EffSpell}(\hat{\sigma}))$	*fand*	*find*	*find*	*fund*	*fund*	*fund*
$\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\sigma})))$	*fand*	*fand*	*fand*	*fand*	*fand*	*fand*

Similar to what we did in §4.2.2 for the case of English, we will construct a "heredity test function" $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ from a "simple heredity test function" in Algorithm 5.6 and a set of "admissible suffix mismatch and vowel alternation" rules in Algorithm 5.8.

Algorithm 5.6 (Simple heredity test)**.** *Suppose that $\hat{\alpha}$ and $\hat{\beta}$ are both lowercase strings. Set* $\mathbf{V} = (a|e|i|o|u)$*. The Boolean-valued function* $\text{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if $\hat{\alpha}$ contains at least one instance of* **V***,* **AND** *at least one of the following three conditions holds:*[53]

(i) $\hat{\alpha} = \hat{\beta}$*;*

(ii) $\hat{\beta} = \hat{\alpha}t$*;*

(iii) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) > \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha} = \hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha})+1\}} = (e|s)$*. (See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

In what follows, we shall define $\text{SuffixNW}(\hat{\alpha},\hat{\beta})$ and $\text{SuffixSW}(\hat{\alpha},\hat{\beta})$ as what is done in the English case (Algorithm 4.11), but the definitions of $\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta})$ and $\text{SuffixSW}(\hat{\alpha},\hat{\beta}), \text{SW}^*(\hat{\alpha},\hat{\beta})$ will be subtly different.

Algorithm 5.7 (Roots and Suffixes by NW and SW)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the function values* $\text{RootNW}(\hat{\alpha},\hat{\beta})$*,* $\text{SuffixNW}(\hat{\alpha},\hat{\beta})$ *and* $\text{NW}^*(\hat{\alpha},\hat{\beta})$ *are determined through the following procedure:*

- *Use the Needleman–Wunsch algorithm to align the sequences as* $\text{NW}(\hat{\alpha},\hat{\beta})$*.*
- *If* $\text{NW}(\hat{\alpha},\hat{\beta})$ *ends with a mismatch (shown in brackets in Examples 3.7.1 and 3.7.2), use this mismatch to define* $\text{SuffixNW}(\hat{\alpha},\hat{\beta})$*; otherwise, define* $\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [\varnothing,\varnothing]$*.*
- *Construct* $\text{RootNW}(\hat{\alpha},\hat{\beta})$ *from all the (not necessarily contiguous) matching string portions in* $\text{NW}(\hat{\alpha},\hat{\beta})$*, joined together by* # = U+0023*;*
- *If the first mutation bracket in* $\text{NW}(\hat{\alpha},\hat{\beta})$ *does not appear at the end* [*which in turn, makes a non-void contribution to* $\text{SuffixNW}(\hat{\alpha},\hat{\beta})$]*, then define* $\text{NW}^*(\hat{\alpha},\hat{\beta})$ *as this first mismatch; otherwise, define* $\text{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$*;*

The function values $\text{RootSW}(\hat{\alpha},\hat{\beta})$*,* $\text{SuffixSW}(\hat{\alpha},\hat{\beta})$ *and* $\text{SW}^*(\hat{\alpha},\hat{\beta})$ *are determined similarly, through the* SW *function.*

Example 5.7.1. Let $\hat{\alpha}$ =*fandt*, $\hat{\beta}$ =*find*, then we have $\text{NW}(\hat{\alpha},\hat{\beta}) = \text{SW}(\hat{\alpha},\hat{\beta})$ =*f*[*a*, *i*]*nd*[*t*, ∅], leading to $\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = \text{SuffixSW}(\hat{\alpha},\hat{\beta})$ =[*t*, ∅], $\text{RootNW}(\hat{\alpha},\hat{\beta}) = \text{RootSW}(\hat{\alpha},\hat{\beta})$ =*f*#*nd*, $\text{NW}^*(\hat{\alpha},\hat{\beta}) = \text{SW}^*(\hat{\alpha},\hat{\beta})$ =[*a*, *i*].

Algorithm 5.8 (Admissible suffix mismatch and vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{SuffixNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if*

$$\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [(\varnothing|(e|nd|t)_m|((bar|dom|hed|ig|lig|skab)\mathbf{X})), (\varnothing|(e|nd|t)_m|((bar|dom|hed|ig|lig|skab)\mathbf{X}))]$$

AND *at least one of the following two conditions holds:*

(i) $\text{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$ **AND** $\text{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* $\mathbf{V} = (a|e|i|o|u)$*;*

(ii) $\text{NW}^*(\hat{\alpha},\hat{\beta}) = ([a, i]|[a, o]|[a, u]|[e, i]|[i, u]|[o, u])$.

[53] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmMut}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 5.9 (Heredity test function)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function* $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\text{SimpHrdTest}(\hat{\alpha},\hat{\beta}) = \textbf{TRUE}$*;*

(ii) $\ell(\text{RootNW}(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^{*}(\hat{\alpha},\hat{\beta})) = \textbf{TRUE}$*;*

(iii) $\ell(\text{RootSW}(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\text{AdmMut}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^{*}(\hat{\alpha},\hat{\beta})) = \textbf{TRUE}$.

Our approximate clustering algorithm for Danish differs from the English counterpart (Algorithm 4.15) in one specific detail: in Algorithm 5.10 below, vowel blotting is applied to both the effective spelling and the essential root in two stages, while Algorithm 4.15 has restricted the use of vowel blotting to the essential root.

Algorithm 5.10 (Approximate clustering of Danish words)**.** *The approximate clustering of a list of Danish words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1,\text{EffSpell}(\hat{\alpha}_1),\text{BlotV}_1(\text{EffSpell}(\hat{\alpha}_1))), \ldots, (\hat{\alpha}_N,\text{EffSpell}(\hat{\alpha}_N),\text{BlotV}_1(\text{EffSpell}(\hat{\alpha}_N)))\}$ *alphabetically according to the third component (with higher priority) and the second component (with lower priority). If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)},\text{EffSpell}(\hat{\alpha}_{(1)}),\text{BlotV}_1(\text{EffSpell}(\hat{\alpha}_{(1)}))), \ldots, (\hat{\alpha}_{(N)},\text{EffSpell}(\hat{\alpha}_{(N)}),\text{BlotV}_1(\text{EffSpell}(\hat{\alpha}_{(N)})))\}$ *satisfy* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n)}),\text{EffSpell}(\hat{\alpha}_{(n+1)})) = \textbf{FALSE}$*, where $n \in \mathbb{Z} \cap [1, N)$, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)},\text{EffSpell}(\hat{\alpha}_{(1,1)}),\text{BlotV}_1(\text{EffSpell}(\hat{\alpha}_{(1,1)}))), \ldots, (\hat{\alpha}_{(1,n_1)},*,*)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)},*,*), \ldots, (\hat{\alpha}_{(M,n_M)},*,*)\}\}$,[54] *where each sublist also preserves alphabetic order. In particular, we have $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ and $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.*

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)},*,*), \ldots, (\hat{\alpha}_{(m,n_m)},*,*)\}$ *where $m \in \mathbb{Z} \cap [1, M]$, we augment it into a tagged entry* $G_m = (g_m,\text{EffSpell}(\hat{\alpha}_{(m,1)}),\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})),\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\text{BlotV}_1(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})))$ *(with highest priority),* $\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with medium priority), and* $\text{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)},\hat{\gamma}'_{(1)},\hat{\gamma}''_{(1)},\hat{\gamma}'''_{(1)}), \ldots, G_{(M)} = (g_{(M)},\hat{\gamma}'_{(M)},\hat{\gamma}''_{(M)},\hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\text{HrdTest}(\hat{\gamma}''_{(m)},\hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)},\hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, then add a demarcation line between these two entries. In this manner, the alphabetized list $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots, vowel blotted forms) from each Γ_k, where $k \in \mathbb{Z} \cap [1, K]$.*

Example 5.10.1. As an illustration of our clustering algorithm, we pick the following families of Danish words, where an "approximate translation" in English, enclosed in quotation marks, is appended to the end of each family:[55]

and, *anden*, *andens*, *ands*, *ænder*, *ænderne*, *ændernes*, *ænders* — "duck";
Anders, *Andersen*, *Andersens* — "Andrew";
barn, *barnet*, *barnets*, *barns*, *børn*, *børnene*, *børnenes*, *børns* — "child";
bonde, *bonden*, *bondens*, *bondes*, *bønder*, *bønderne*, *bøndernes*, *bønders* — "farmer";
datter, *datteren*, *datterens*, *datters*, *døtre*, *døtrene*, *døtrenes*, *døtres* — "daughter";
ferie, *ferien*, *feriens*, *ferier*, *ferierne*, *feriernes*, *feriers*, *feries* — "holiday";
fængsel, *fængselerne*, *fængselet*, *fængselets*, *fængsels*, *fængsler*, *fængslernes*, *fængslers*, *fængslet*, *fængslets* — "jail";
himle, *himlen*, *himlene*, *himlenes*, *himlens*, *himles*, *himmel*, *himmelen*, *himmelens*, *himmels* — "heaven";
hus, *hus'*, *huse*, *husende*, *husene*, *husenes*, *huser*, *huses*, *huset*, *husets*, *hust*, *huste* — "house";

[54] Hereafter, we use an asterisk (∗) to abbreviate components that are clear from context.

[55] We note that within each family, words may be related to each other through both inflections (adjective comparisons, noun declensions, verb conjugations) and derivations (such as compound nouns formed from several components, verbs derived from nouns, etc.). Therefore, the "approximate translation" only matches certain members in the specified word family.

husven, *husvennen*, *husvennens*, *husvenner*, *husvennerne*, *husvennernes*, *husvenners*, *husvens* — "family friend";
hænder, *hænderne*, *hændernes*, *hænders*, *Haand*, *hånd*, *hånden*, *hånds* — "hand";
kø, *køen*, *køens*, *køer*, *køerne*, *køernes*, *køers*, *køs* — "queue";
køb, *købene*, *købenes*, *køber*, *købet*, *købets*, *købs*, *købt*, *købte*, *købtes* — "purchase";
København, *Københavns* — "Copenhagen";
mand, *manden*, *mandens*, *mands*, *mænd*, *mændene*, *mændenes*, *mænds* — "man";
prins, *prinse*, *prinselig*, *prinsen*, *prinsens*, *prinser*, *prinserne*, *prinsernes*, *prinsers*, *prinses* — "prince";
prinsesse, *prinsessekrone*, *prinsessen*, *prinsessens*, *prinsesser*, *prinsesserne*, *prinsessernes*, *prinsessers*, *prinsesses* — "princess";
ske, *skeen*, *skeens*, *skeer*, *skeerne*, *skeernes*, *skeers*, *skes* — "spoon";
ske, *sker*, *sket*, *skete* — "happen";
stor, *store*, *stort*, *større*, *størst* — "big";
ung, *ungdom*, *unge*, *ungt*, *yngre*, *yngst* — "young";
ven, *vennen*, *vennens*, *venner*, *vennerne*, *vennernes*, *venners*, *vens* — "friend";
venlig, *venligere*, *venligst*, *venligste*, *venligt* — "friendly";
vent, *vente*, *ventede*, *ventedes*, *venter*, *ventes*, *ventet* — "wait";
vin, *vine*, *vinen*, *vinene*, *vinenes*, *vinens*, *vines*, *vins* — "wine";
vinter, *vinteren*, *vinterens*, *vinters*, *vintre*, *vintrene*, *vintrenes*, *vintres* — "winter";
vinterferie, *vinterferien*, *vinterferiens*, *vinterferier*, *vinterferierne*, *vinterferiernes*, *vinterferiers*, *vinterferies* — "winter holiday";
ø, *øen*, *øens*, *øer*, *øerne*, *øernes*, *øers*, *øs* — "island";
å, *åen*, *åens*, *åer*, *åerne*, *åernes*, *åers*, *ås* — "creek";
år, *årene*, *årenes*, *året*, *årets*, *års* — "year".

Applying Algorithm 5.10 to this list of words, we obtain the following clustering results:

{*and*, *anden*, *andens*, *Anders*, *Andersen*, *Andersens*, *ands*, *ænder*, *ænderne*, *ændernes*, *ænders*}, {*barn*, *barnet*, *barnets*, *barns*, *børn*, *børnene*, *børnenes*, *børns*}, {*bonde*, *bonden*, *bondens*, *bondes*, *bønder*, *bønderne*, *bøndernes*, *bønders*}, {*datter*, *datteren*, *datterens*, *datters*, *døtre*, *døtrene*, *døtrenes*, *døtres*}, {*ferie*, *ferien*, *feriens*, *ferier*, *ferierne*, *feriernes*, *feriers*, *feries*}, {*fængsel*, *fængselerne*, *fængselet*, *fængselets*, *fængsels*, *fængsler*, *fængslernes*, *fængslers*, *fængslet*, *fængslets*}, {*himle*, *himlen*, *himlene*, *himlenes*, *himlens*, *himles*, *himmel*, *himmelen*, *himmelens*, *himmels*}, {*hus*, *hus'*, *huse*, *husende*, *husene*, *husenes*, *huser*, *huses*, *huset*, *husets*, *hust*, *huste*, *husven*, *husvennen*, *husvennens*, *husvenner*, *husvennerne*, *husvennernes*, *husvenners*, *husvens*}, {*hænder*, *hænderne*, *hændernes*, *hænders*, *Haand*, *hånd*, *hånden*, *hånds*}, {*kø*, *køen*, *køens*, *køer*, *køerne*, *køernes*, *køers*, *køs*}, {*køb*, *købene*, *købenes*, *køber*, *købet*, *købets*, *købs*, *købt*, *købte*, *købtes*}, {*København*, *Københavns*}, {*mand*, *manden*, *mandens*, *mands*, *mænd*, *mændene*, *mændenes*, *mænds*}, {*prins*, *prinse*, *prinsen*, *prinsens*, *prinser*, *prinserne*, *prinsernes*, *prinsers*, *prinses*, *prinsesse*, *prinsessekrone*, *prinsessen*, *prinsessens*, *prinsesser*, *prinsesserne*, *prinsessernes*, *prinsessers*, *prinsesses*}, {*prinselig*}, {*ske*, *skeen*, *skeens*, *skeer*, *skeerne*, *skeernes*, *skeers*, *sker*, *skes*, *sket*, *skete*}, {*stor*, *store*, *stort*, *større*, *størst*}, {*ung*, *ungdom*, *unge*, *ungt*, *yngre*, *yngst*}, {*ven*, *veninde*, *veninden*, *venindens*, *veninder*, *veninderne*, *venindernes*, *veninders*, *venindes*, *vennen*, *vennens*, *venner*, *vennerne*, *vennernes*, *venners*, *vens*}, {*venlig*, *venligere*, *venligst*, *venligste*, *venligt*}, {*vent*, *vente*, *ventede*, *ventedes*, *venter*, *ventes*, *ventet*}, {*vin*, *vine*, *vinen*, *vinene*, *vinenes*, *vinens*, *vines*, *vins*}, {*vinter*, *vinteren*, *vinterens*, *vinters*, *vintre*, *vintrene*, *vintrenes*, *vintres*}, {*vinterferie*, *vinterferien*, *vinterferiens*, *vinterferier*, *vinterferierne*, *vinterferiernes*, *vinterferiers*, *vinterferies*}, {*ø*, *øen*, *øens*, *øer*, *øerne*, *øernes*, *øers*, *øs*}, {*å*, *åen*, *åens*, *åer*, *åerne*, *åernes*, *åers*, *ås*}, {*år*, *årene*, *årenes*, *året*, *årets*, *års*}.

Example 5.10.2. To further test our algorithm against Danish irregular verbs, we throw the following list:

afbrudt, *afbryd*, *afbryde*, *afbryder*, *afbrød* — "interrupt";
bar, *bær*, *bære*, *bærer*, *båret* — "carry";
besat, *besatte*, *besæt*, *besætte*, *besætter* — "occupy";
beskrev, *beskreven*, *beskrevet*, *beskrevne*, *beskriv*, *beskriver* — "describe";
blotlagde, *blotlagt*, *blotlæg*, *blotlægge*, *blotlægger* — "expose";
bræk, *brække*, *brækkede*, *brækker*, *brækket* — "break";
drak, *drik*, *drikker*, *drukken*, *drukket*, *drukne* — "drink";
fandt, *find*, *finder*, *funden*, *fundet*, *fundne* — "find";
fik, *få*, *fået*, *får* — "get";
flyv, *flyve*, *flyver*, *fløj*, *fløjet* — "fly";
gav, *gavs*, *giv*, *give*, *givende*, *givendes*, *giver*, *gives*, *givet*, *givets* — "give";
gentag, *gentagen*, *gentager*, *gentaget*, *gentagne*, *gentog* — "repeat";

græd, *græde*, *græder*, *grædt* — “weep”;
hjalp, *hjulpen*, *hjulpet*, *hjulpne*, *hjælp*, *hjælpe*, *hjælper* — “help”;
kom, *komme*, *kommen*, *kommende*, *kommer*, *kommet*, *komne* — “come”;
lad, *lader*, *ladet*, *ladt*, *lod* — “allow”;
lig, *ligge*, *ligger*, *ligget*, *lå* — “lie (be in horizontal position)”;
løb, *løbe*, *løber*, *løbet* — “run”;
optræd, *optræde*, *optræder*, *optrådt*, *optrådte* — “appear”;
ryg, *ryge*, *ryger*, *røg*, *røget* — “smoke”;
sang, *sunget*, *syng*, *syngende*, *synger* — “sing”;
trak, *trukken*, *trukket*, *trukne*, *træk*, *trækkende*, *trækker* — “pull”;
æd, *æde*, *æder*, *ædt*, *åd* — “eat (like an animal)”

into our algorithm, and obtain

{*afbrudt*, *afbryd*, *afbryde*, *afbryder*, *afbrød*}, {*bar*, *bær*, *bære*, *bærer*, *båret*}, {*besat*, *besatte*, *besæt*, *besætte*, *besætter*}, {*beskrev*, *beskreven*, *beskrevet*, *beskrevne*, *beskriv*, *beskriver*}, {*blotlagde*, *blotlagt*, *blotlæg*, *blotlægge*, *blotlægger*}, {*bræk*, *brække*, *brækkede*, *brækker*, *brækket*}, {*drak*, *drik*, *drikker*, *drukken*, *drukket*, *drukne*}, {*fandt*, *find*, *finder*, *funden*, *fundet*, *fundne*}, {*fik*, *få*, *fået*, *får*}, {*flyv*, *flyve*, *flyver*, *fløj*, *fløjet*}, {*gav*, *gavs*, *giv*, *give*, *givende*, *givendes*, *giver*, *gives*, *givet*, *givets*}, {*gentag*, *gentagen*, *gentager*, *gentaget*, *gentagne*, *gentog*}, {*græd*, *græde*, *græder*, *grædt*}, {*hjalp*, *hjulpen*, *hjulpet*, *hjulpne*, *hjælp*, *hjælpe*, *hjælper*}, {*kom*, *komme*, *kommen*, *kommende*, *kommer*, *kommet*, *komne*}, {*lad*, *lader*, *ladet*, *ladt*, *lod*}, {*lig*, *ligge*, *ligger*, *ligget*, *lå*}, {*løb*, *løbe*, *løber*, *løbet*}, {*optræd*, *optræde*, *optræder*, *optrådt*, *optrådte*}, {*ryg*, *ryge*, *ryger*, *røg*, *røget*}, {*sang*, *sunget*, *syng*, *syngende*, *synger*}, {*trak*, *trukken*, *trukket*, *trukne*, *træk*, *trækkende*, *trækker*}, {*æd*, *æde*, *æder*, *ædt*, *åd*}.

Example 5.10.3. Combining the inputs from the last two examples, we obtain the following output:

{*afbrudt*, *afbryd*, *afbryde*, *afbryder*, *afbrød*}, {*and*, *anden*, *andens*, *Anders*, *Andersen*, *Andersens*, *ands*, *ænder*, *ænderne*, *ændernes*, *ænders*}, {*bar*, *bær*, *bære*, *bærer*, *båret*}, {*barn*, *barnet*, *barnets*, *barns*, *børn*, *børnene*, *børnenes*, *børns*}, {*besat*, *besatte*, *besæt*, *besætte*, *besætter*}, {*beskrev*, *beskreven*, *beskrevet*, *beskrevne*, *beskriv*, *beskriver*}, {*blotlagde*, *blotlagt*, *blotlæg*, *blotlægge*, *blotlægger*}, {*bonde*, *bonden*, *bondens*, *bondes*, *bønder*, *bønderne*, *bøndernes*, *bønders*}, {*bræk*, *brække*, *brækkede*, *brækker*, *brækket*}, {*datter*, *datteren*, *datterens*, *datters*, *døtre*, *døtrene*, *døtrenes*, *døtres*}, {*drak*, *drik*, *drikker*, *drukken*, *drukket*, *drukne*}, {*fandt*, *find*, *finder*, *funden*, *fundet*, *fundne*}, {*ferie*, *ferien*, *feriens*, *ferier*, *ferierne*, *feriernes*, *feriers*, *feries*}, {*fik*, *få*, *fået*, *får*}, {*flyv*, *flyve*, *flyver*, *fløj*, *fløjet*}, {*fængsel*, *fængselerne*, *fængselet*, *fængselets*, *fængsels*, *fængsler*, *fængslernes*, *fængslers*, *fængslet*, *fængslets*}, {*gav*, *gavs*, *giv*, *give*, *givende*, *givendes*, *giver*, *gives*, *givet*, *givets*}, {*gentag*, *gentagen*, *gentager*, *gentaget*, *gentagne*, *gentog*}, {*græd*, *græde*, *græder*, *grædt*}, {*himle*, *himlen*, *himlene*, *himlenes*, *himlens*, *himles*, *himmel*, *himmelen*, *himmelens*, *himmels*}, {*hjalp*, *hjulpen*, *hjulpet*, *hjulpne*, *hjælp*, *hjælpe*, *hjælper*}, {*hus*, *hus'*, *huse*, *husende*, *husene*, *husenes*, *huser*, *huses*, *huset*, *husets*, *hust*, *huste*, *husven*, *husvennen*, *husvennens*, *husvenner*, *husvennerne*, *husvennernes*, *husvenners*, *husvens*}, {*hænder*, *hænderne*, *hændernes*, *hænders*, *Haand*, *hånd*, *hånden*, *hånds*}, {*kom*, *komme*, *kommen*, *kommende*, *kommer*, *kommet*, *komne*}, {*kø*, *køen*, *køens*, *køer*, *køerne*, *køernes*, *køers*, *køs*}, {*køb*, *købene*, *købenes*, *køber*, *købet*, *købets*, *købs*, *købt*, *købte*, *købtes*}, {*København*, *Københavns*}, {*lad*, *lader*, *ladet*, *ladt*, *lod*}, {*lig*, *ligge*, *ligger*, *ligget*, *lå*}, {*løb*, *løbe*, *løber*, *løbet*}, {*mand*, *manden*, *mandens*, *mands*, *mænd*, *mændene*, *mændenes*, *mænds*}, {*optræd*, *optræde*, *optræder*, *optrådt*, *optrådte*}, {*prins*, *prinse*, *prinsen*, *prinsens*, *prinser*, *prinserne*, *prinsernes*, *prinsers*, *prinses*, *prinsesse*, *prinsessekrone*, *prinsessen*, *prinsessens*, *prinsesser*, *prinsesserne*, *prinsessernes*, *prinsessers*, *prinsesses*}, {*prinselig*}, {*ryg*, *ryge*, *ryger*, *røg*, *røget*}, {*sang*, *sunget*, *syng*, *syngende*, *synger*}, {*ske*, *skeen*, *skeens*, *skeer*, *skeerne*, *skeernes*, *skeers*, *sker*, *skes*, *sket*, *skete*}, {*stor*, *store*, *stort*, *større*, *størst*}, {*trak*, *trukken*, *trukket*, *trukne*, *træk*, *trækkende*, *trækker*}, {*ung*, *ungdom*, *unge*, *ungt*, *yngre*, *yngst*}, {*ven*, *veninde*, *veninden*, *venindens*, *veninder*, *veninderne*, *venindernes*, *veninders*, *venindes*, *vennen*, *vennens*, *venner*, *vennerne*, *vennernes*, *venners*, *vens*}, {*venlig*, *venligere*, *venligst*, *venligste*, *venligt*}, {*vent*, *vente*, *ventede*, *ventedes*, *venter*, *ventes*, *ventet*}, {*vin*, *vine*, *vinen*, *vinene*, *vinenes*, *vinens*, *vines*, *vins*}, {*vinter*, *vinteren*, *vinterens*, *vinters*, *vintre*, *vintrene*, *vintrenes*, *vintres*}, {*vinterferie*, *vinterferien*, *vinterferiens*, *vinterferier*, *vinterferierne*, *vinterferiernes*, *vinterferiers*, *vinterferies*}, {*æd*, *æde*, *æder*, *ædt*, *åd*}, {*ø*, *øen*, *øens*, *øer*, *øerne*, *øernes*, *øers*, *øs*}, {*å*, *åen*, *åens*, *åer*, *åerne*, *åernes*, *åers*, *ås*}, {*år*, *årene*, *årenes*, *året*, *årets*, *års*}.

5.1.3 Heuristic detection of compounds

In Danish (as well as German and Dutch), compounds are usually spelt as a single word, without spaces or hyphens between the constituting components. For example, the Danish compound *vinterferie* “winter holiday” is formed by joining *vinter*

"winter" and *ferie* "holiday". Aside from direct concatenation, Danish compounds may also involve an intervening letter *e* or *s* [46, §12.2.2.2].

Automatic detection of the "hidden word boundaries" within Germanic compounds is a non-trivial task. However, if we limit our scope to compounds with only two constituting components that both appear in isolations in a vocabulary list, there are heuristic methods to break down these binary compound words. We will integrate such a heuristic compound detection method with Danish word clustering, later in Algorithm 5.13 (which is a variation on Algorithm 5.10), where any detected binary compound is "dissolved" into its two constituting components ("head" and "tail") and filed as two copies, one clustered with the "head" word, the other with the "tail" word.

Before stating Algorithm 5.13 in full, we need some preparations (Definition 5.11 and 5.12) to sort out potential binary compounds and their constituting components.

Definition 5.11 (String minus operation)**.** Suppose that $\ell(\hat{\beta}) \geq \ell(\hat{\alpha})$ and $\hat{\beta} = \hat{\beta}^{[\ell(\hat{\alpha})]}\hat{\tau}$, then $\hat{\beta} \ominus \hat{\alpha} = \hat{\tau}$ defines the the "string minus operation". In other words, $\hat{\beta} \ominus \hat{\alpha}$ is a string that leaves out the first $\ell(\hat{\alpha})$ characters in $\hat{\beta}$. □

Algorithm 5.12 (Heuristic identification of Danish binary compounds)**.** *Let $\Lambda^{\hat{\rho}} = \{\hat{\rho}_1, \ldots, \hat{\rho}_Q\}$ be a list of distinct Danish essential roots (without vowel blotting) that contain at least one instance of* $\mathbf{V} = (a|e|i|o|u)$ *and DO NOT match the following string patterns:*

$$(\pi\rho a|a|af|ag|ak|al|an\mathbf{X}|ar|at|be|ben|dag|du|\,frem|lo|mi|mod|na|nar|\\ ned|om|op|ot|ov|rak|re|rig|rod|ru|sam|tag|u|ud|und|vfor\mathbf{X}|\mathbf{X}ant|zil\mathbf{X}).$$

The output of the function $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is obtained through the following procedures:*

(1) *Construct a list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$ *where* $\lambda_q^{\hat{\rho}} = \{\hat{\rho}_{(q,1)}, \ldots, \hat{\rho}_{(q,n_q)}\}$ *is a subset of* $\Lambda^{\hat{\rho}}$ *whose members all match the string pattern* $\hat{\rho}_q\sim$*, for* $q \in \mathbb{Z} \cap [1, Q]$.

(2) *Expand the aforementioned entry* $(\hat{\rho}_q, \lambda_q^{\hat{\rho}})$ *into a list of triplets* $\{(\hat{\rho}_{(q,1)}, \hat{\rho}_q, \hat{\rho}_{(q,1)} \ominus \hat{\rho}_q), \ldots, (\hat{\rho}_{(q,n_q)}, \hat{\rho}_q, \hat{\rho}_{(q,n_q)} \ominus \hat{\rho}_q)\}$ *for every* $q \in \mathbb{Z} \cap [1, Q]$ *such that* $\lambda_q^{\hat{\rho}} \neq \varnothing$*. Collect all these triplets as one runs through the list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$*. The list of these triplets* $\{(\hat{\rho}_{\langle 1\rangle}, \hat{\eta}_{\langle 1\rangle}, \hat{\rho}_{\langle 1\rangle} \ominus \hat{\eta}_{\langle 1\rangle}), \ldots, (\hat{\rho}_{\langle Q'\rangle}, \hat{\eta}_{\langle Q'\rangle}, \hat{\rho}_{\langle Q'\rangle} \ominus \hat{\eta}_{\langle Q'\rangle})\}$ *contains potentially valid decompositions of compounds.*

(3) *Screen the aforementioned list of triplets as follows: for every* $q' \in \mathbb{Z} \cap [1, Q']$*, if* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \hat{\tau}_{\langle q'\rangle} = \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle})$ *satisfies*

$$\ell(\hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle}) \geq 2 \quad \textbf{AND} \quad \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle} = \mathbf{X}_1(a|e|i|o|u)\mathbf{X}_2,$$

then construct $\hat{\tau}^*_{\langle q'\rangle}$ *by performing* $(e|s|t)\sim\!\!\rightarrow \varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$ *and* $\hat{\tau}^{**}_{\langle q'\rangle}$ *by doing* $(er|es)\sim\!\!\rightarrow \varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$*, before generating a list* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *by members of* $\Lambda^{\hat{\rho}}$ *that match the pattern* $\underline{(\hat{\tau}_{\langle q'\rangle}|\hat{\tau}^*_{\langle q'\rangle}|\hat{\tau}^{**}_{\langle q'\rangle})}$*; otherwise, set* $\lambda^{\hat{\tau}}_{\langle q'\rangle} = \varnothing$.

(4) *Collect all the triplets* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \lambda^{\hat{\tau}}_{\langle q'\rangle})$ *where* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *is non-void and* $\hat{\tau}_{\langle q'\rangle}$ *DOES NOT match the following string patterns:*

$$(dom\mathbf{X}|hed\mathbf{X}|lig\mathbf{X}|om).$$

This list of triplets $\mathrm{CpdDet}(\Lambda_{\hat{\rho}})$ *contains the heuristic decompositions of all the identified binary compounds.*

Algorithm 5.13 (Approximate clustering of Danish words with heuristic detection of compounds)**.** *The approximate clustering of a list of Danish words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *respecting compounding is completed in four stages:*

(1) *Do as in Algorithm 5.10(1).*

(2) *Do as in Algorithm 5.10(2). Save both the tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$ *and the list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *for further use.*

(3) *Construct a tagged list of word clusters* $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$ *where* $\Lambda_k^{\hat{\rho}}$ *is the union of all Danish essential roots (without vowel blotting) available to* Γ_k*, for* $k \in \mathbb{Z} \cap [1, K]$*. Set* $\Lambda^{\hat{\rho}} = \Lambda_1^{\hat{\rho}} \cup \cdots \cup \Lambda_K^{\hat{\rho}}$*, and evaluate* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$.

(4) *The first component* $\hat{\rho}_{\langle q''\rangle}$ *of each triplet* $(\hat{\rho}_{\langle q''\rangle}, \hat{\eta}_{\langle q''\rangle}, \lambda^{\hat{\tau}}_{\langle q''\rangle})$ *in* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is called a "dissolvable compound", the second component* $\hat{\eta}_{\langle q''\rangle}$ *a "heuristic head", and the first member in the third component* $\lambda^{\hat{\tau}}_{\langle q''\rangle}$ *a "heuristic tail". In the tagged list* $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$*, every entry containing a "dissolvable compound" is removed, and regrouped with the entries matching its "heuristic head" and "heuristic tail". Finally, remove all tags.*

Example 5.13.1. We may revisit Example 5.10.3 with the algorithm above. The results are given below:

{*afbrudt*, *afbryd*, *afbryde*, *afbryder*, *afbrød*}, {*and*, *anden*, *andens*, *Anders*, *Andersen*, *Andersens*, *ands*, *ænder*, *ænderne*, *ændernes*, *ænders*}, {*bar*, *bær*, *bære*, *bærer*, *båret*}, {*barn*, *barnet*, *barnets*, *barns*, *børn*, *børnene*, *børnenes*, *børns*}, {*besat*, *besatte*, *besæt*, *besætte*, *besætter*}, {*beskrev*, *beskreven*, *beskrevet*, *beskrevne*, *beskriv*, *beskriver*}, {*blotlagde*, *blotlagt*, *blotlæg*, *blotlægge*, *blotlægger*}, {*bonde*, *bonden*, *bondens*, *bondes*, *bønder*, *bønderne*, *bøndernes*, *bønders*}, {*bræk*, *brække*, *brækkede*, *brækker*, *brækket*}, {*datter*, *datteren*, *datterens*, *datters*, *døtre*, *døtrene*, *døtrenes*, *døtres*}, {*drak*, *drik*, *drikker*, *drukken*, *drukket*, *drukne*}, {*fandt*, *find*, *finder*, *funden*, *fundet*, *fundne*}, {*ferie*, *ferien*, *feriens*, *ferier*, *ferierne*, *feriernes*, *feriers*, *feries*, *vinterferie*, *vinterferien*, *vinterferiens*, *vinterferier*, *vinterferierne*, *vinterferiernes*, *vinterferiers*, *vinterferies*}, {*fik*, *få*, *fået*, *får*}, {*flyv*, *flyve*, *flyver*, *fløj*, *fløjet*}, {*fængsel*, *fængselerne*, *fængselet*, *fængselets*, *fængsels*, *fængsler*, *fængslernes*, *fængslers*, *fængslet*, *fængslets*}, {*gav*, *gavs*, *giv*, *give*, *givende*, *givendes*, *giver*, *gives*, *givet*, *givets*}, {*gentag*, *gentagen*, *gentager*, *gentaget*, *gentagne*, *gentog*}, {*græd*, *græde*, *græder*, *grædt*}, {*himle*, *himlen*, *himlene*, *himlenes*, *himlens*, *himles*, *himmel*, *himmelen*, *himmelens*, *himmels*}, {*hjalp*, *hjulpen*, *hjulpet*, *hjulpne*, *hjælp*, *hjælpe*, *hjælper*}, {*hus*, *hus'*, *huse*, *husende*, *husene*, *husenes*, *huser*, *huses*, *huset*, *husets*, *hust*, *huste*, *husven*, *husvennen*, *husvennens*, *husvenner*, *husvennerne*, *husvennernes*, *husvenners*, *husvens*}, {*hænder*, *hænderne*, *hændernes*, *hænders*, *Haand*, *hånd*, *hånden*, *hånds*}, {*kom*, *komme*, *kommen*, *kommende*, *kommer*, *kommet*, *komne*}, {*kø*, *køen*, *køens*, *køer*, *køerne*, *køernes*, *køers*, *køs*}, {*køb*, *købene*, *købenes*, *køber*, *købet*, *købets*, *købs*, *købt*, *købte*, *købtes*}, {*København*, *Københavns*}, {*lad*, *lader*, *ladet*, *ladt*, *lod*}, {*lig*, *ligge*, *ligger*, *ligget*, *lå*}, {*løb*, *løbe*, *løber*, *løbet*}, {*mand*, *manden*, *mandens*, *mands*, *mænd*, *mændene*, *mændenes*, *mænds*}, {*optræd*, *optræde*, *optræder*, *optrådt*, *optrådte*}, {*prins*, *prinse*, *prinselig*, *prinsen*, *prinsens*, *prinser*, *prinserne*, *prinsernes*, *prinsers*, *prinses*, *prinsesse*, *prinsessekrone*, *prinsessen*, *prinsessens*, *prinsesser*, *prinsesserne*, *prinsessernes*, *prinsessers*, *prinsesses*}, {*ryg*, *ryge*, *ryger*, *røg*, *røget*}, {*sang*, *sunget*, *syng*, *syngende*, *synger*}, {*ske*, *skeen*, *skeens*, *skeer*, *skeerne*, *skeernes*, *skeers*, *sker*, *skes*, *sket*, *skete*}, {*stor*, *store*, *stort*, *større*, *størst*}, {*trak*, *trukken*, *trukket*, *trukne*, *træk*, *trækkende*, *trækker*}, {*ung*, *ungdom*, *unge*, *ungt*, *yngre*, *yngst*}, {*ven*, *veninde*, *veninden*, *venindens*, *veninder*, *veninderne*, *venindernes*, *veninders*, *venindes*, *venlig*, *venligere*, *venligst*, *venligste*, *venligt*, *vennen*, *vennens*, *venner*, *vennerne*, *vennernes*, *venners*, *vens*}, {*vent*, *vente*, *ventede*, *ventedes*, *venter*, *ventes*, *ventet*}, {*vin*, *vine*, *vinen*, *vinene*, *vinenes*, *vinens*, *vines*, *vins*}, {*vinter*, *vinteren*, *vinterens*, *vinterferie*, *vinterferien*, *vinterferiens*, *vinterferier*, *vinterferierne*, *vinterferiernes*, *vinterferiers*, *vinterferies*, *vinters*, *vintre*, *vintrene*, *vintrenes*, *vintres*}, {*æd*, *æde*, *æder*, *ædt*, *åd*}, {*ø*, *øen*, *øens*, *øer*, *øerne*, *øernes*, *øers*, *øs*}, {*å*, *åen*, *åens*, *åer*, *åerne*, *åernes*, *åers*, *ås*}, {*år*, *årene*, *årenes*, *året*, *årets*, *års*}.

Here, the word family generated by *vinterferie* "winter holiday" is correctly redistributed to the constituting components *vinter* "winter" and *ferie* "holiday". In the meantime, there are no inappropriate compound detections, such as grouping *bære* "carry" with *barn* "child".

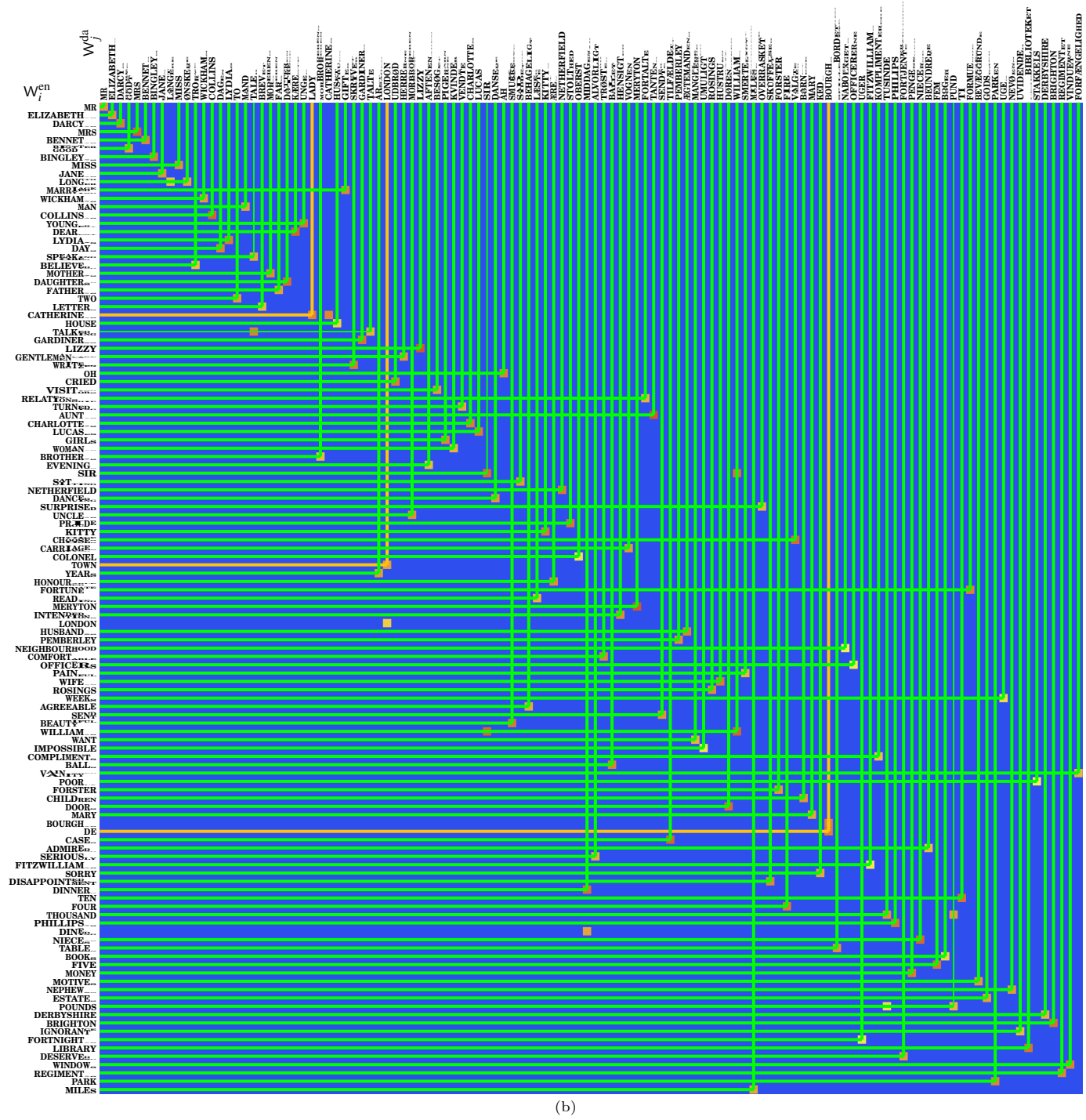
Example 5.13.2. In Fig. S4, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).

The Danish word *ved* can mean "at" or "know" (present tense), which can be either a function word or a highly frequent content word (from the English perspective). We have chosen not to include *ved* in our list of Danish stop words (Definition 5.1), but have made no efforts to resolve its polysemy. Thus, in the second cluster of Fig. S4a, we have conflated several concepts that are spelt as *ved* ("at" or "know") and *vide* ("know" or "wide").

Moreover, the English topic *de* (part of the proper name "de Bourgh") cannot be correctly translated by our algorithm, because the list of Danish stop words (Definition 5.1) includes *de* "the (plural), they, those". Nevertheless, we consider a link between English *de* and Danish *Bourgh* an excusable error (marked in amber color in Fig. S4b, since these two words always stay together in the entire text—with nearly identical kinetic behavior.

It should be noted that Jane Austen used the word *related* mostly in the sense of *told* in her novel, which is correctly translated into *fortalte* in Danish.

(a)



(b)

Fig. S4. Text mining in Danish. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Danish version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{da})$ between selected topics in English and Danish versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *amber*) cross-hair indicates an exact (resp. a close but non-exact) match.

5.2 Modified Porter stemming algorithm for German

Definition 5.14 (German stop words)**.** If the lowercase[56] form of a word belongs to the following list[57]:

> *ab*, *aber*, *all*, *alle*, *allein*, *allem*, *allen*, *aller*, *alles*, *als*, *also*, *am*, *an*, *ander*, *andere*, *anderem*, *anderen*, *anderer*, *anderes*, *anderm*, *andern*, *anders*, *aneinander*, *auch*, *auf*, *aufeinander*, *aufs*, *aus*, *auseinander*, *ausser*, *außer*, *ausserdem*, *außerdem*, *bei*, *beide*, *beidem*, *beiden*, *beider*, *beides*, *beieinander*, *beim*, *bevor*, *bin*, *binnen*, *bis*, *bisher*, *bislang*, *bist*, *da*, *dabei*, *dafür*, *dagegen*, *dahin*, *damit*, *danach*, *dann*, *daran*, *darauf*, *darf*, *darfst*, *darin*, *darüber*, *darum*, *das*, *dass*, *daß*, *dasselbe*, *davon*, *dazu*, *dein*, *deine*, *deinem*, *deinen*, *deiner*, *deines*, *deins*, *dem*, *deme*, *demselben*, *den*, *denen*, *denn*, *dennoch*, *denselben*, *der*, *deren*, *derer*, *dern*, *dero*, *derselbe*, *derselben*, *des*, *deshalb*, *deß*, *desselben*, *dessen*, *desto*, *dich*, *die*, *dies*, *diese*, *dieselbe*, *dieselben*, *diesem*, *diesen*, *dieser*, *dieses*, *dir*, *doch*, *dort*, *dorther*, *dorthin*, *du*, *durch*, *durcheinander*, *dürfe*, *dürfen*, *dürfend*, *dürfest*, *dürfet*, *dürft*, *durfte*, *dürfte*, *durften*, *dürften*, *durftest*, *dürftest*, *durftet*, *dürftet*, *ebenso*, *ein*, *einander*, *eine*, *einem*, *einen*, *einer*, *eines*, *einig*, *einige*, *einigem*, *einigen*, *einiger*, *einiges*, *einmal*, *empor*, *er*, *es*, *etwas*, *euch*, *euer*, *eure*, *eurem*, *euren*, *eurer*, *eures*, *falls*, *für*, *füreinander*, *ganz*, *ganze*, *ganzem*, *ganzen*, *ganzer*, *ganzes*, *gar*, *gedurft*, *gegen*, *gegeneinander*, *gegenüber*, *gehabt*, *gekonnt*, *gemocht*, *gemusst*, *gemußt*, *genug*, *gerade*, *geradezu*, *gesollt*, *getan*, *gewesen*, *gewollt*, *geworden*, *hab*, *habe*, *haben*, *habend*, *habest*, *habet*, *habt*, *hast*, *hat*, *hatte*, *hätte*, *hatten*, *hätten*, *hattest*, *hättest*, *hattet*, *hättet*, *her*, *herab*, *heran*, *herauf*, *heraus*, *herbei*, *herein*, *herüber*, *hervor*, *hier*, *hierher*, *hierhin*, *hin*, *hinab*, *hinauf*, *hinaus*, *hindurch*, *hinein*, *hingegen*, *hinten*, *hinter*, *hintereinander*, *hinüber*, *hinunter*, *hinweg*, *hinzu*, *ich*, *ihm*, *ihn*, *ihnen*, *ihr*, *ihre*, *ihrem*, *ihren*, *ihrer*, *ihres*, *im*, *immer*, *immerhin*, *in*, *indem*, *indes*, *indessen*, *ineinander*, *ins*, *iregendetwas*, *iregendwo*, *irgend*, *irgendein*, *irgendeine*, *irgendeinem*, *irgendeinen*, *irgendeiner*, *irgendeines*, *irgendjemand*, *irgendjemandem*, *irgendjemanden*, *irgendjemandes*, *irgendwann*, *irgendwas*, *irgendwelche*, *irgendwelchem*, *irgendwelchen*, *irgendwelcher*, *irgendwelches*, *irgendwer*, *irgendwie*, *irgendwohin*, *ist*, *ja*, *je*, *jede*, *jedem*, *jeden*, *jeder*, *jedes*, *jedoch*, *jemand*, *jemandem*, *jemanden*, *jemandes*, *jemands*, *jene*, *jenem*, *jenen*, *jener*, *jenes*, *jetzt*, *kann*, *kannst*, *kaum*, *kein*, *keine*, *keinem*, *keinen*, *keiner*, *keines*, *könne*, *können*, *könnend*, *könnest*, *könnet*, *könnt*, *konnte*, *könnte*, *konnten*, *könnten*, *konntest*, *könntest*, *konntet*, *könntet*, *los*, *machen*, *mag*, *magst*, *man*, *manche*, *manchem*, *manchen*, *mancher*, *manches*, *manchmal*, *mehr*, *mehre*, *mehrere*, *mehreres*, *mehres*, *mein*, *meine*, *meinem*, *meinen*, *meiner*, *meines*, *meins*, *mich*, *mir*, *mit*, *miteinander*, *mochte*, *möchte*, *mochten*, *möchten*, *mochtest*, *möchtest*, *mochtet*, *möchtet*, *möge*, *mögen*, *mögend*, *mögest*, *möget*, *mögt*, *muss*, *muß*, *müsse*, *müssen*, *müssend*, *müssest*, *müsset*, *musst*, *mußt*, *müsst*, *müßt*, *musste*, *mußte*, *müsste*, *müßte*, *mussten*, *mußten*, *müssten*, *müßten*, *musstest*, *mußtest*, *müsstest*, *müßtest*, *musstet*, *mußtet*, *müsstet*, *müßtet*, *na*, *nach*, *nachdem*, *nacheinander*, *nebeneinander*, *nein*, *nicht*, *nichts*, *nie*, *niemals*, *niemand*, *niemandem*, *niemanden*, *niemandes*, *niemands*, *nimmer*, *noch*, *nun*, *nur*, *ob*, *oben*, *obgleich*, *obwohl*, *obzwar*, *oder*, *oft*, *öfter*, *öftesten*, *ohne*, *pro*, *schon*, *sehr*, *sei*, *seien*, *seiet*, *sein*, *seine*, *seinem*, *seinen*, *seiner*, *seines*, *seins*, *seist*, *seit*, *seitdem*, *selbst*, *sich*, *sie*, *sind*, *so*, *sobald*, *sogar*, *solange*, *solch*, *solche*, *solchem*, *solchen*, *solcher*, *solches*, *soll*, *solle*, *sollen*, *sollend*, *sollest*, *sollet*, *sollst*, *sollt*, *sollte*, *sollten*, *solltest*, *solltet*, *sondern*, *sonst*, *soviel*, *soweit*, *sowie*, *sowohl*, *stets*, *tat*, *täte*, *taten*, *täten*, *tatest*, *tätest*, *tatet*, *tätet*, *trotz*, *trotzdem*, *tu*, *tue*, *tuen*, *tuend*, *tuest*, *tuet*, *tun*, *ture*, *tust*, *tut*, *über*, *überdies*, *übereinander*, *um*, *umsonst*, *und*, *uns*, *unse*, *unsem*, *unsen*, *unser*, *unsere*, *unserem*, *unseren*, *unserer*, *unseres*, *unserm*, *unsern*, *unses*, *unsre*, *unsrem*, *unsren*, *unsrer*, *unsres*, *unter*, *untereinander*, *usw*, *viel*, *viele*, *vielen*, *vieles*, *vielleicht*, *vielmehr*, *vom*, *von*, *voneinander*, *vor*, *vorbei*, *vordem*, *voreinander*, *vorüber*, *während*, *wann*, *war*, *wär*, *ward*, *warden*, *wardest*, *wardet*, *wardst*, *wäre*, *waren*, *wären*, *wärest*, *wäret*, *warst*, *wärst*, *wart*, *warum*, *was*, *weg*, *weil*, *weiter*, *welche*, *welchem*, *welchen*, *welcher*, *welches*, *wem*, *wen*, *wenig*, *wenige*, *wenigen*, *weniger*, *wenigsten*, *wenigstens*, *wenn*, *wenngleich*, *wer*, *werd*, *werde*, *werden*, *werdend*, *werdest*, *werdet*, *weshalb*, *wessen*, *weswegen*, *wie*, *wieder*, *wieso*, *wieviel*, *wiewohl*, *will*, *willst*, *wir*, *wird*, *wirst*, *wo*, *woher*, *wohin*, *wohl*, *wolle*, *wollen*, *wollend*, *wollest*, *wollet*, *wollt*, *wollte*, *wollten*, *wolltest*, *wolltet*, *worden*, *wurde*, *würde*, *wurden*, *würden*, *wurdest*, *würdest*, *wurdet*, *würdet*, *ziemlich*, *zu*, *zudem*, *zueinander*, *zum*, *zur*, *zusammen*, *zurück*, *zuvor*, *zwar*, *zwischen*,

then we consider it a German stop word. All the German stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

The post-1996 German orthography requires writing *ss* in place of *ß* in certain words. This new rule affects the stop word *daß*, whose new spelling *dass* has been added to our stop word list. We need to add *schon* "already" to the list of stop words, to avoid conflation with *schön* "beautiful" during approximate clustering (which ignores all umlauts). Although the words *schon* and *schön* were indeed etymologically related (reminiscent of the adverbial and adjectival uses of the word *pretty* in English), their current meanings are well separated.

[56] For the matching of German stop words, lowercase forms are enforced before comparison. Therefore, both *sie* "she, they" and *Sie* "you (polite form)" appearing in a German text are considered stop words, even though the capitalized form *Sie* does not appear in our list.

[57] Our list of German stop words is based on `snowball.tartarus.org/algorithms/german/stop.txt`, with extensive additions that roughly match their counterparts in English.

For clustering German words, it is necessary that we ignore umlauts, so that *Apfel* "apple" is clustered with *Äpfel* "apples" etc. We also need to convert all the occurrences of *ß* to *ss* (which matches the orthographic practice for German spelling in Switzerland), in order to accommodate to texts written before the German spelling reform in 1996: *Kuß* "kiss" (noun, pre-1996), *Kuss* "kiss" (noun, post-1996), *Küsse* "kisses" (noun, pre- and post-1996). This necessarily carries certain risk, as *Buße* "penance" (pre- and post-1996) will be conflated with *Busse* "buses" (pre- and post-1996).

All German nouns are capitalized, while adjectives and verbs derived from nouns are not: cf. *Deutsch* "German" (noun, "the German language") and *deutsch* "German" (adjective). It is thus advisable to ignore capitalization during clustering, contrary to the practice in English.

5.2.1 Effective spelling and essential root

Algorithm 5.15 (German effective spelling)**.** *Set* **V** = (*a*|*e*|*ê*|*i*|*î*|*o*|*u*) *and* **V*** = (*a*|*ê*|*î*|*i*|*o*|*u*)*. For a German word* σ̂*, its effective spelling* EffSpell(σ̂) *is constructed in the following steps:*[58]

(1) *Convert to lowercase, and replace*

(∅|ver)st(a|e|i|o|ü)rb~	χ̂kunft	χ̂sicht	χ̂tr(ä|a)cht	abwesen(d|h)**X**
στερb	χ̂χ̂⁺kunft	χ̂×2siχt	χ̂⁺χ̂tracht	αβσεντ

acht~	ähn	anwesen(d|h)**X**	b(ü|u)ch	bald~	benn~	bessie~	blut~	born~	brot	dame	dämmer
8e	σμähn	πρεsentτ	βuκ	βaλδ	βenn	βεσθie	βλωoδ	brunn	βροτ	δaμe	δaωmερ

dent~	diana~	dicht	ehe~	erinn~	f(ü|u)(ss|ß)	fessel	fessl	fest~	feuer	gardi~
δeν̃tτ	διαν̃a	θiκt	μρe	μν̃eμnn	φuτσζ	φeσel	φeσl	φeστ	φeueρ	γarδi

gef(ä|a)hr	gegenwart~	geheim	gestalt	gewalt~	gott~	h(ä|a|ie)lt	h(ö|o)he	hall
δaν̃γρr	πρεsentτ	σεκρετ	σηaιπ	φoρct	γoττ	χoλδ	hoche	hσαall

hannah~	helen~	hell	hüll	inne~	irr	jone~	leah~	ledig	liebst	lied	lond~
χaν̃ν̃aχ	ελeν̃η	hχeλl	ωρaπ	ιν̃ne	iιρr	γιone	λeααη	σiν̃γ	lieb	λieδ	λonδ

m(ama|utti)~	m(ü|u)nd~	meile	minut	müd	oa~	oberst	ort~	pferd	plötz	porträt	quell
mutter	μuν̃δ	μειλe	μiν̃ωτ	μτiρδ	ωa	oβerστ	ωort	χoρσ	σuδδeν̃	ροrτρät	qveλl

r(ö|o)t~	recht	ree~	reich	reis	relig	ri(ss|ß)~	robert	roches~	s(a|ä)(ss|ß)
roθ	ρiγχτ	ρee	ρρeich	ρεeiσs	ρeliγ	ρiφτ	ροβeρτ	rρωξeσ	sitz

schulter~	sommer~	spiel	still	talent	them	tot~	versuch~	weg	welk	well~	wohl	ach
schuλδer	σoμmer	πλal	σtiλl	ταλeν̃τ	θeμ	στερb	veρσuχ	ωeγ	ωiθρk	oν̃δal	whol	αaχ

besessen	de	gatte(∅|n)	gattin(∅|nen)	gesessen	gibt	mrs	tod(∅|e|en|es|s)	~leute(∅|n)
besitzen	δεe	ehemann	ehefrau	sitzen	geben	zφρau	στερb	mann

(2) *Replace*

'	(∅|ge)gönn~	(∅|ge)hind~	(bös|bosheit|boshaft**X**)	χ̂∈(a|e|i|k|n|o|p|s|t|u|y)tisch**X**
∅	goenn	χind	bboes	χ̂ik

X∈(a|i)list	**X**∈(ge|k|sch|ss|ß)ling	ä	absent	bek~	best$_{\hat{\chi}\notin}$(e)~	bett	brannt~	ch	eb~	ei
Xl	**X**lang	a	αβσεnτ	beκ	bstχ̂	zbett	brenn	ç	xeb	ê

ernst~	finger	fünf	gegess	gesund~	hübe	janu~	langsam~	leb	leib~	leist	lös	mäd	manier
erν̃στ	φinγer	5e	ess	γesund	hebe	jnu	γλangsam	leβ	lεib	λειστ	λoζ	makd	μanier

monat	ö	ob~	passag	ruh	schlecht	schock	setz	sitz
μonat	o	ωb	paσσag	ρuh	schλecht	schoκ	ζetz	σitz

sonn	ß	trän	tz	ü	über	überred~	verkünd	vier~	vulg
σoν̃ν̃	ss	τeaρ	ztz	u	xüber	πeσβ	πρoν̃σ	4e	xvxulg

wahr	wasser	welt	wert	wünsch**X**	zoll	gebrochen	labe	last	miss(∅|es)	~is(mus|tisch**X**)
vwahr	waσer	xwelt	fwert	wunsç	τoλλ	breçen	λaβ	lesen	μiσ	∅

(3) *Replace*

[58] To avoid confusion of Greek nu with Latin vee, we write the former as ν̃ in substitution rules.

(∅|be)str~	(∅|ge)last~	(∅|ge)lieb~	(braçte|gebraçt)	(lass|liess)	abend~	ahn~
στr	xlast	lxieb	bringe	λass	azbend	xzahn

alt~	anz$_{\hat{\chi}\notin}$(u)	arb	beσitz	bes$_{\hat{\chi}\notin}$(s|t)~	bet~	bild	bitter~	bus~	çs$_{\hat{\chi}\notin}$(t)	daçt	dank	dien
aλτ	ant$\hat{\chi}$	αrb	φoρτuṽ	ges$\hat{\chi}$	ββet	βild	βitter	zbus	x$\hat{\chi}$	denk	thank	δien

dumm	erb	fass	fê~	folg	froh	fund$_{\hat{\chi}\notin}$(e)	gast	gel(a|i|u)ng~	gew(a|i|o)nn	grund	gruss
δumm	eρβe	φass	fvê	vfolg	vroh	φund$\hat{\chi}$	γast	geλing	wiṽ	grunδ	kgruss

gut	hand	hart	hu~	ie	in$^{\mathbf{X}\in}$(d|g|k)~	kind	klag	komi	komp~	kumm	kund	kuns	lêd	lês~	list	lob
besser	hanδ	xhart	shu	î	în**X**	gkind	κlag	κomi	comp	xkumm	xkund	gkuns	λêd	σoφτ	ylist	λob

lock	lub	lust	mutter	nahr	name	nummer	ohr~	ort~	platz	raf~	reçn	rest~	saç	satz	sçon
λock	λub	zlust	μuther	ṽahr	ṽame	xnunmmer	ohρ	orτ	pplatz	ρaf	ρeçn	ρest	xsaç	xσatz	zsçon

sçl	segn	setz	siçt	sieg	sohn	stamm	stand	stemm	suç	tanz	tisç	tugend	uhr~	verg~	vett
sçλ	szegn	zsetz	σiçt	sxieg	szohn	στamm	steh	σztemm	zsuç	taṽz	τisχ	xtugend	uηρ	verzug	vwett

voll	w(ê|ie)s$_{\hat{\chi}\notin}$(s)	wên	wind	winter	wort	wusst	zêt
ffoll	zxwês$\hat{\chi}$	xvên	winδ	wvinter	wvort	wiss	zzêt

geh(∅|e|en|est|st|t)	bat(∅|en|est|et)	gebeten	herr(∅|en|n)	kam(∅|e|en|est|et|st|t)	mr
ging	bitten	bitten	hherr	komm	hherr

oh	sass(∅|en|est|et|t)	segen(∅|s)	tante(∅|n)	~çs	~ement(∅|s)	~enz(∅|en)	~w(ê|ie)s
zxoh	sitz	szegn	τante	x	îren	∅	zxwês

(4) *If the result so far is an empty string, stop here. Otherwise, break down the result into* $\hat{\sigma}_1\hat{\sigma}_2$*, where* $\ell(\hat{\sigma}_1) = \max\{3, \ell(\hat{\sigma})\}$*, obtain* $\hat{\sigma}'_1$ *by doing kam~*$\rightarrow$*zkam on* $\hat{\sigma}_1$ *before deducing* $\hat{\sigma}'_2$ *from* $\hat{\sigma}_2$ *in two sequential steps:*

(3.1) *Do el*$\rightarrow$*l;*

(3.2) *Do kation***X**$\rightarrow$*zîren, ~ln*$\rightarrow$*l,* (*ffol*|*haft*|*ist*|*massig*)**X**$\rightarrow\varnothing$.

Concatenate $\hat{\sigma}'_1$ *and* $\hat{\sigma}'_2$.

(5) *Do sçλoζ~*$\rightarrow$*sçλos,* $\sim_{\hat{\chi}\notin}$(a|e|ê|i|î|o|s|u)s$\rightarrow\hat{\chi}$, $\sim_{\hat{\chi}\notin}$**V**(çen|ling)(∅|e|en|s)$\rightarrow\hat{\chi}$, ~(h|k)êt(∅|en)$\rightarrow\varnothing$, ~ik(∅|en|er(∅|n|s))$\rightarrow\varnothing$, ~lên(∅|s)$\rightarrow\varnothing$, ~**VX**$_{\hat{\chi}\notin}$**V***st$\rightarrow$**VX**$\hat{\chi}$ *on* $\hat{\sigma}'_1\hat{\sigma}'_2$.

Definition 5.16 (German protected range)**.** Set $\mathbf{V}^* = (a|ê|î|i|o|u)$, $\mathbf{C}_{m_0} = \overline{(a|e|ê|i|î|o|u)}_{m_0}$ and $\mathbf{C}_m = \overline{(a|e|i|o|u|ê|î)}_m$. Let $\hat{\sigma}$ be the effective spelling of a German word, its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \max\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is determined by two non-negative integers $\lambda_1(\hat{\sigma})$ and $\lambda_2(\hat{\sigma})$ specified through the following procedures:

- Look for the string pattern $((\varnothing|a|be|e|ê|er|ge|i|mit|naç|u|vor|xüber|zer|zu)\mathbf{C}_{m_0}\mathbf{V}^*|\mathbf{C}_m e)\overline{(a|e|ê|i|î|r|u)}_{m_0}\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Look for the pattern $(a|o|u)$ in the string $\hat{\sigma}$;
- If a letter in the pattern above is found, the last position occupied by such a letter defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = 0$. □

Unlike Danish, but similar to Dutch, our German essential root is free from common separable verb prefixes (see step (3) in the algorithm below).

Algorithm 5.17 (German essential root)**.** *Set* $\mathbf{V} = (a|e|ê|i|î|o|u)$ *and* $\mathbf{C} = \overline{(a|e|ê|i|î|o|u)}$*. Let* $\hat{\sigma}$ *be the effective spelling of a German word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$.

(2) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*[59]

(2.1) *Do ~s*$\rightarrow\varnothing$*;*

(2.2) *Do* $\sim e(d|e|m|n|r|st|t)_{m_0}\rightarrow\varnothing$.

[59] In other words, the core algorithm for essential root extraction runs as follows: keep the last "strong" vowel *a*, *ï*, *o* or *u* in non-final position, plus one subsequent letter; delete final *a*; erase the final appearance of *e* and all the letters thereafter.

The result after these two steps of operations is called $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$*, before doing*

$$(\varnothing|ab|an|auf|aus|bê|durç|ên|h(er|in)(\varnothing|ab|an|auf|aus|ên|xüber)|mit|\\ naç|um|vor|wêter|w(i|î)der|xüber|zu(\varnothing|ruck))(\varnothing|ge|zu)\mathbf{XV}\sim\rightarrow\mathbf{XV};$$

(4) *Do* $(\hat{\chi}_{\times 2}t|\sim\hat{\chi}_{\times 2})\rightarrow\hat{\chi}$;[60] *(Check Example 3.4.2 to see the notation* $\hat{\chi}_{\times 2}$ *for double letters.)*

(5) *Do* $\sim\mathbf{C}t\rightarrow\mathbf{C}$, $nd\rightarrow\tilde{\nu}d$.

5.2.2 Admissible mutation and approximate clustering

Like English, German has hundreds of irregular verbs in daily use. Unlike English, these German verbs may not only have irregular past tense and past participles, but also have irregular present tense conjugations.

The German vowel blotting mechanism is similar to the Danish version (Algorithm 5.5).

Algorithm 5.18 (German vowel blotting)**.** *Set* $\mathbf{V}_m=(a|e|ê|i|î|o|u)_m$ *and* $\mathbf{C}_{m_0}=\overline{(a|e|ê|i|î|o|u)}_{m_0}$. *For a token string* $\hat{\sigma}$*, its blotted form* $\mathrm{BlotV}_1(\hat{\sigma})$ *is constructed as follows:*

- *If the string pattern* $(\varnothing|a|be|e|ê|i|o|u)\mathbf{C}_{m_0}\mathbf{V}_m\sim$ *can be found in the string* $\hat{\sigma}$*, then the last position occupied by such a pattern is replaced by the letter "a".*
- *Otherwise, leave the string* $\hat{\sigma}$ *intact.*

Similar to what we did in §5.1.2 for the case of Danish, we will construct a bivariate Boolean-valued function $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ on a "simple heredity test function" in Algorithm 5.19, and a set of "admissible suffix mismatch and vowel alternation" rules in Algorithm 5.20.

Algorithm 5.19 (Simple heredity test)**.** *Suppose that* $\hat{\alpha}$ *and* $\hat{\beta}$ *are both lowercase strings. Set* $\mathbf{V}=(a|e|ê|i|î|o|u)$. *The Boolean-valued function* $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if* $\hat{\alpha}$ *contains at least one instance of* **V***,* **AND** *at least one of the following four conditions holds:*[61]

(i) $\hat{\alpha}=\hat{\beta}$;

(ii) $\hat{\beta}=\hat{\alpha}d$;

(iii) $\hat{\beta}=\hat{\alpha}t$;

(iv) *Appending the last character of* $\hat{\alpha}$ *to itself, one obtains* $\hat{\beta}$*, i.e.* $\hat{\alpha}\Omega(\hat{\alpha})=\hat{\beta}$;

(v) $\ell(\hat{\beta})>\ell(\hat{\alpha})\geq\frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha}=\hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha})+1\}}=(e|n|s)$. *(See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

In what follows, we define $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}),\mathrm{RootNW}(\hat{\alpha},\hat{\beta}),\mathrm{NW}^*(\hat{\alpha},\hat{\beta})$ and $\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\mathrm{SW}^*(\hat{\alpha},\hat{\beta})$ as what is done in the Danish case (Algorithm 5.7).

Algorithm 5.20 (Admissible suffix mismatch and vowel alternation)**.** *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function*

$$\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}),\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}),\mathrm{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if*[62]

$$\begin{aligned}&(\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})=[(\varnothing|(d|e|n|t)_m|(\Omega(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}))|bar|er|ig|in|îr|isç|liç|sam|st|tum|ung)\mathbf{X})),\\ &\qquad(\varnothing|(d|e|n|t)_m|(\Omega(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}))|bar|er|ig|in|îr|isç|liç|sam|st|tum|ung)\mathbf{X}))]\\ &\textbf{OR}\ \mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})=[îh,og]\ \textbf{OR}\ \mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})=[og,îh])\end{aligned}$$

AND *at least one of the following three conditions holds:*[63]

(i) $\mathrm{NW}^*(\hat{\alpha},\hat{\beta})=\varnothing$ **AND** $\mathrm{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* $(a|e|ê|i|î|o|u|z)$;

[60] In *Mathematica* codes, the substitution rule reads `{x_~~x_~~"t"|WordBoundary :> x}`.

[61] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

[62] Depending on the programming language chosen, the string *îh* may be sorted before or after *og*.

[63] We note that $\mathrm{NW}^*(\hat{\alpha},\hat{\beta})$ and $\mathrm{NW}^*(\hat{\beta},\hat{\alpha})$ differ only in the order of the two components in the bracket.

(ii) $\mathrm{NW}^*(\hat{\alpha},\hat{\beta}) = ([a, e]||[a, i]||[a, î]||[a, î]||[a, o]||[a, u]||[ah, i]||[ah, o]||[au, î]||[au, o]||[e, i]||[e, î]||[e, o]||[ê, i]||[ê, î]||[eh, i]||[eh, o]||[i, o]||[î, i]||[î, o]||[î, u]||[o, u])$;

(iii) $\mathrm{NW}^*(\hat{\beta},\hat{\alpha}) = ([a, e]||[a, i]||[a, î]||[a, î]||[a, o]||[a, u]||[ah, i]||[ah, o]||[au, î]||[au, o]||[e, i]||[e, î]||[e, o]||[ê, i]||[ê, î]||[eh, i]||[eh, o]||[i, o]||[î, i]||[î, o]||[î, u]||[o, u])$.

Similarly, one can evaluate another Boolean-valued function

$$\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 5.21 (Heredity test function)**.** *In what follows,* $\ell([\hat{\sigma},\hat{\tau}]) = \min\{\ell(\hat{\sigma}),\ell(\hat{\tau})\}$. *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function* $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta}) =$ **TRUE**;

(ii) $\ell(\mathrm{RootNW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{NW}^*(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}),\ \mathrm{SuffixNW}(\hat{\alpha},\ \hat{\beta}),\ \mathrm{NW}^*(\hat{\alpha},\hat{\beta})) =$ **TRUE**;

(iii) $\ell(\mathrm{RootSW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SW}^*(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}),\ \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\ \mathrm{SW}^*(\hat{\alpha},\hat{\beta})) =$ **TRUE**.

Algorithm 5.22 (Approximate clustering of German words)**.** *The algorithm is essentially the same as Algorithm 5.10, except that German rules (instead of Danish rules) apply to all the tags (effective spelling, essential root, vowel blotting etc.).*

Example 5.22.1. As an illustration of our clustering algorithm, we pick the following families of German words, where an "approximate translation" in English, enclosed in quotation marks, is appended to the end of each family:

Apfel, *Äpfel*, *Äpfelchen*, *Äpfeln*, *Apfels* — "apple";
Arbeit, *arbeite*, *arbeiten*, *Arbeiten*, *arbeitest*, *arbeitet*, *arbeitete*, *arbeiteten*, *gearbeitet* — "work";
Arbeiter, *Arbeitern*, *Arbeiters* — "worker";
Arbeitsplatz, *Arbeitsplatze*, *Arbeitsplätze*, *Arbeitsplätzen*, *Arbeitsplatzes* — "workplace";
Arbeitstag, *Arbeitstage*, *Arbeitstagen*, *Arbeitstages*, *Arbeitstags* — "workday";
Bus, *Busse*, *Bussen*, *Busses* — "bus";
Buße, *Bußen* — "penance";
dunkel, *dunkelste*, *dunkelstem*, *dunkelsten*, *dunkelster*, *dunkelstes*, *dunkle*, *dunklem*, *dunklen*, *dunkler*, *dunklere*, *dunklerem*, *dunkleren*, *dunklerer*, *dunkleres*, *dunkles* — "dark";
Ei, *Eie*, *Eier*, *Eiern*, *Eies*, *Eis* — "egg";
Eigelb, *Eigelbe*, *Eigelben*, *Eigelbs* — "egg yolk";
Eis, *Eise*, *Eises* — "ice";
frei, *freie*, *freiem*, *freien*, *freier*, *freiere*, *freierem*, *freieren*, *freierer*, *freieres*, *freies*, *freiste*, *freistem*, *freisten*, *freisten*, *freister*, *freistes* — "free";
Freiheit, *Freiheiten* — "freedom";
gehaust, *haus*, *Haus*, *Häuschen*, *hause*, *Hause*, *hausen*, *hausend*, *Häuser*, *Hauses*, *haust*, *hauste*, *hauste*, *haustest* — "house";
geküsst, *Kuss*, *Kuß*, *Küsschen*, *Kusse*, *küsse*, *Küsse*, *küssen*, *Küssen*, *Kusses*, *küssest*, *küsst*, *küsste*, *küsstest*, *küsstet* — "kiss";
gelb, *gelbe*, *gelbem*, *gelben*, *gelber*, *gelbere*, *gelberem*, *gelberen*, *gelberer*, *gelberes*, *gelbes*, *gelbste*, *gelbstem*, *gelbsten*, *gelbster*, *gelbstes* — "yellow";
gross, *groß*, *große*, *großem*, *großen*, *großer*, *größere*, *größerem*, *größeren*, *größerer*, *größeres*, *großes*, *größte*, *größtem*, *größten*, *größter*, *größtes* — "big";
Öl, *Öle*, *Ölen*, *Öles*, *Öls* — "oil";
Platz, *Platze*, *Plätze*, *Plätzen*, *Platzes* — "place";
studier, *studiere*, *studiere*, *studiere*, *studieren*, *studierend*, *studierest*, *studieret*, *studierst*, *studiert*, *studiert*, *studierte*, *studierte*, *studierten*, *studiertest*, *studiertest*, *studiertet* — "study";
Tag, *Tage*, *Tagen*, *Tages*, *Tags* — "day".

Applying Algorithm 5.22 to this list of words, we obtain the following clustering results:

{*Apfel*, *Äpfel*, *Äpfelchen*, *Äpfeln*, *Apfels*}, {*Arbeit*, *arbeite*, *arbeiten*, *Arbeiten*, *Arbeiter*, *Arbeitern*, *Arbeiters*, *arbeitest*, *arbeitet*, *arbeitete*, *arbeiteten*, *gearbeitet*}, {*Arbeitsplatz*, *Arbeitsplatze*, *Arbeitsplätze*, *Arbeitsplätzen*, *Arbeitsplatzes*}, {*Arbeitstag*, *Arbeitstage*, *Arbeitstagen*, *Arbeitstages*, *Arbeitstags*}, {*Bus*, *Busse*, *Bußе*, *Bussen*, *Bußen*, *Busses*}, {*dunkel*, *dunkelste*, *dunkelstem*, *dunkelsten*, *dunkelster*, *dunkelstes*, *dunkle*, *dunklem*, *dunklen*, *dunkler*, *dunklere*, *dunklerem*, *dunkleren*, *dunklerer*, *dunkleres*, *dunkles*}, {*Ei*, *Eie*, *Eier*, *Eiern*, *Eies*}, {*Eigelb*, *Eigelbe*, *Eigelben*, *Eigelbs*}, {*Eis*, *Eise*, *Eises*}, {*frei*, *freie*, *freiem*, *freien*, *freier*, *freiere*, *freierem*, *freieren*, *freierer*, *freieres*, *freies*, *Freiheit*, *Freiheiten*, *freiste*, *freistem*, *freisten*, *freister*, *freistes*}, {*gehaust*, *haus*, *Haus*, *Häuschen*, *hause*, *Hause*, *hausen*, *hausend*, *Häuser*, *Hauses*, *haust*, *hauste*, *haustest*}, {*geküsst*, *Kuss*, *Kuß*, *Küsschen*, *Kusse*, *küsse*, *Küsse*, *küssen*, *Küssen*, *Kusses*, *küssest*, *küsst*, *küsste*, *küsstest*, *küsstet*}, {*gelb*, *gelbe*, *gelbem*, *gelben*, *gelber*, *gelbere*, *gelberem*, *gelberen*, *gelberer*, *gelberes*, *gelbes*, *gelbste*, *gelbstem*, *gelbsten*, *gelbster*, *gelbstes*}, {*gross*, *groß*, *große*, *großem*, *großen*, *großer*, *größere*, *größerem*, *größeren*, *größerer*, *größeres*, *großes*, *größte*, *größtem*, *größten*, *größter*, *größtes*}, {*Öl*, *Öle*, *Ölen*, *Öles*, *Öls*}, {*Platz*, *Platze*, *Plätze*, *Plätzen*, *Platzes*}, {*studier*, *studiere*, *studieren*, *studierend*, *studierest*, *studieret*, *studierst*, *studiert*, *studierte*, *studierten*, *studiertest*, *studiertet*}, {*Tag*, *Tage*, *Tagen*, *Tages*, *Tags*}.

Example 5.22.2. To further test our algorithm against German irregular verbs with various vowel alternation patterns across different tenses and moods, we throw the following list:

befahl, *befähle*, *befahlen*, *befählen*, *befählest*, *befählet*, *befahlst*, *befahlt*, *befehle*, *befehlen*, *befehlend*, *befehlest*, *befehlet*, *befehlt*, *befiehlst*, *befiehlt*, *befohlen* — "command";
begann, *begänne*, *begannen*, *begännen*, *begännest*, *begännet*, *begannst*, *begannt*, *beginn*, *beginne*, *beginnen*, *beginnend*, *beginnest*, *beginnet*, *beginnst*, *beginnt*, *begönne*, *begonnen*, *begönnen*, *begönnest*, *begönnet* — "begin";
beiß, *beiße*, *beißen*, *beißend*, *beißest*, *beißet*, *beißt*, *biss*, *bisse*, *bissen*, *bissest*, *bisset*, *bisst*, *gebissen* — "bite";
blas, *blase*, *blasen*, *blasend*, *blasest*, *blaset*, *blast*, *bläst*, *blies*, *bliese*, *bliesen*, *bliesest*, *blieset*, *bliest*, *geblasen* — "blow";
brach, *bräche*, *brachen*, *brächen*, *brächest*, *brächet*, *brachst*, *bracht*, *breche*, *brechen*, *brechend*, *brechest*, *brechet*, *brecht*, *brich*, *brichst*, *bricht*, *gebrochen* — "break";
brät, *brate*, *braten*, *bratend*, *bratest*, *bratet*, *brätst*, *briet*, *briete*, *brieten*, *brietest*, *brietet*, *gebraten* — "fry";
fahr, *fahre*, *fahren*, *fahrend*, *fahrest*, *fahret*, *fährst*, *fahrt*, *fährt*, *fuhr*, *führe*, *fuhren*, *führen*, *führest*, *führet*, *fuhrst*, *fuhrt*, *gefahren* — "drive";
fall, *falle*, *fallen*, *fallend*, *fallest*, *fallet*, *fällst*, *fallt*, *fällt*, *fiel*, *fiele*, *fielen*, *fielest*, *fielet*, *fielst*, *fielt*, *gefallen* — "fall";
fand, *fände*, *fanden*, *fänden*, *fandest*, *fändest*, *fandet*, *fändet*, *finde*, *finden*, *findend*, *findest*, *findet*, *gefunden* — "find";
fechte, *fechten*, *fechtend*, *fechtest*, *fechtet*, *ficht*, *fichtst*, *focht*, *föchte*, *fochten*, *föchten*, *fochtest*, *föchtest*, *fochtet*, *föchtet*, *gefochten* — "fence";
flieg, *fliege*, *fliegen*, *fliegend*, *fliegest*, *flieget*, *fliegst*, *fliegt*, *flog*, *flöge*, *flogen*, *flögen*, *flögest*, *flöget*, *flogst*, *flogt*, *geflogen* — "fly";
fraß, *fräße*, *fraßen*, *fräßen*, *fräßest*, *fräßet*, *fraßt*, *fresse*, *fressen*, *fressend*, *fressest*, *fresset*, *fresst*, *friss*, *frisst*, *gefressen* — "feed";
galt, *gälte*, *galten*, *gälten*, *galtest*, *gältest*, *galtet*, *gältet*, *gegolten*, *gelte*, *gelten*, *geltend*, *geltest*, *geltet*, *gilt*, *giltst*, *gölte*, *gölten*, *göltest*, *göltet* — "count";
geglichen, *gleich*, *gleiche*, *gleichen*, *gleichend*, *gleichest*, *gleichet*, *gleichst*, *gleicht*, *glich*, *gliche*, *glichen*, *glichest*, *glichet*, *glichst*, *glicht* — "resemble";
gegriffen, *greif*, *greife*, *greifen*, *greifend*, *greifest*, *greifet*, *greifst*, *greift*, *griff*, *griffe*, *griffen*, *griffest*, *griffet*, *griffst*, *grifft* — "grab";
gehangen, *häng*, *hänge*, *hängen*, *hängend*, *hängest*, *hänget*, *hängst*, *hängt*, *hing*, *hinge*, *hingen*, *hingest*, *hinget*, *hingst*, *hingt* — "hang";
gehoben, *haben*, *heb*, *hebe*, *heben*, *hebend*, *hebest*, *hebet*, *hebst*, *hebt*, *hob*, *höbe*, *hoben*, *höben*, *höbest*, *höbet*, *hobst*, *hobt*, *hübe*, *hüben*, *hübest*, *hübet* — "heave";
gekrochen, *kriech*, *krieche*, *kriechen*, *kriechend*, *kriechest*, *kriechet*, *kriechst*, *kriecht*, *kroch*, *kröche*, *krochen*, *kröchen*, *kröchest*, *kröchet*, *krochst*, *krocht* — "creep";
gelaufen, *lauf*, *laufe*, *laufen*, *laufend*, *laufest*, *laufet*, *läufst*, *lauft*, *läuft*, *lief*, *liefe*, *liefen*, *liefest*, *liefet*, *liefst*, *liefst* — "walk";
gelesen, *las*, *läse*, *lasen*, *läsen*, *läsest*, *läset*, *last*, *lese*, *lesen*, *lesend*, *lesest*, *leset*, *lest*, *lies*, *liest* — "read";
genommen, *nahm*, *nähme*, *nahmen*, *nähmen*, *nähmest*, *nähmet*, *nahmst*, *nahmt*, *nehme*, *nehmen*, *nehmend*, *nehmest*, *nehmet*, *nehmt*, *nimm*, *nimmst*, *nimmt* — "take";
geritten, *reite*, *reiten*, *reitend*, *reitest*, *reitet*, *ritt*, *ritte*, *ritten*, *rittest*, *rittet* — "ride";
gerufen, *rief*, *riefe*, *riefen*, *riefest*, *riefet*, *riefst*, *rieft*, *ruf*, *rufe*, *rufen*, *rufend*, *rufest*, *rufet*, *rufst*, *ruft* — "shriek";

geschlafen, *schlaf*, *schlafe*, *schlafen*, *schlafend*, *schlafest*, *schlafet*, *schläfst*, *schlaft*, *schläft*, *schlief*, *schliefe*, *schliefen*, *schliefest*, *schliefet*, *schliefst*, *schlieft* — “sleep”;
geschlossen, *schließ*, *schließe*, *schließen*, *schließend*, *schließest*, *schließet*, *schließt*, *schloss*, *schlösse*, *schlossen*, *schlössen*, *schlossest*, *schlössest*, *schlosset*, *schlösset*, *schlosst* — “close”;
geschrieben, *schreibe*, *schreiben*, *schreibend*, *schreibest*, *schreibet*, *schreibst*, *schreibt*, *schrieb*, *schriebe*, *schrieben*, *schriebest*, *schriebet*, *schriebst*, *schriebt* — “write”;
gesoffen, *sauf*, *saufe*, *saufen*, *saufend*, *saufest*, *saufet*, *säufst*, *sauft*, *säuft*, *soff*, *söffe*, *soffen*, *söffen*, *söffest*, *söffet*, *soffst*, *sofft* — “drink (of an animal), booze”;
gestoßen, *stieß*, *stieße*, *stießen*, *stießest*, *stießet*, *stießt*, *stoß*, *stoße*, *stoßen*, *stoßend*, *stoßest*, *stoßet*, *stoßt*, *stößt* — “shove”;
getroffen, *traf*, *träfe*, *trafen*, *träfen*, *träfest*, *träfet*, *trafst*, *traft*, *treffe*, *treffen*, *treffend*, *treffest*, *treffet*, *trefft*, *triff*, *triffst*, *trifft* — “meet”;
gezogen, *zieh*, *ziehe*, *ziehen*, *ziehend*, *ziehest*, *ziehet*, *ziehst*, *zieht*, *zog*, *zöge*, *zogen*, *zögen*, *zögest*, *zöget*, *zogst*, *zogt* — “pull”

into our algorithm, and obtain

{*befahl*, *befähle*, *befahlen*, *befählen*, *befählest*, *befählet*, *befahlst*, *befahlt*, *befehle*, *befehlen*, *befehlend*, *befehlest*, *befehlet*, *befehlt*, *befiehlst*, *befiehlt*, *befohlen*}, {*begann*, *begänne*, *begannen*, *begännen*, *begännest*, *begännet*, *begannst*, *begannt*, *beginn*, *beginne*, *beginnen*, *beginnend*, *beginnest*, *beginnet*, *beginnst*, *beginnt*, *begönne*, *begonnen*, *begönnen*, *begönnest*, *begönnet*}, {*beiß*, *beiße*, *beißen*, *beißend*, *beißest*, *beißet*, *beißt*, *biss*, *bisse*, *bissen*, *bissest*, *bisset*, *bisst*, *gebissen*}, {*blas*, *blase*, *blasen*, *blasend*, *blasest*, *blaset*, *blast*, *bläst*, *blies*, *bliese*, *bliesen*, *bliesest*, *blieset*, *bliest*, *geblasen*}, {*brach*, *bräche*, *brachen*, *brächen*, *brächest*, *brächet*, *brachst*, *bracht*, *breche*, *brechen*, *brechend*, *brechest*, *brechet*, *brecht*, *brich*, *brichst*, *bricht*, *gebrochen*}, {*brät*, *brate*, *braten*, *bratend*, *bratest*, *bratet*, *brätst*, *briet*, *briete*, *brieten*, *brietest*, *brietet*, *gebraten*}, {*fahr*, *fahre*, *fahren*, *fahrend*, *fahrest*, *fahret*, *fährst*, *fahrt*, *fährt*, *fuhr*, *führe*, *fuhren*, *führen*, *führest*, *führet*, *fuhrst*, *fuhrt*, *gefahren*}, {*fall*, *falle*, *fallen*, *fallend*, *fallest*, *fallet*, *fällst*, *fallt*, *fällt*, *fiel*, *fiele*, *fielen*, *fielest*, *fielet*, *fielst*, *fielt*, *gefallen*}, {*fand*, *fände*, *fanden*, *fänden*, *fandest*, *fändest*, *fandet*, *fändet*, *finde*, *finden*, *findend*, *findest*, *findet*, *gefunden*}, {*fechte*, *fechten*, *fechtend*, *fechtest*, *fechtet*, *ficht*, *fichtst*, *focht*, *föchte*, *fochten*, *föchten*, *fochtest*, *föchtest*, *fochtet*, *föchtet*, *gefochten*}, {*flieg*, *fliege*, *fliegen*, *fliegend*, *fliegest*, *flieget*, *fliegst*, *fliegt*, *flog*, *flöge*, *flogen*, *flögen*, *flögest*, *flöget*, *flogst*, *flogt*, *geflogen*}, {*fraß*, *fräße*, *fraßen*, *fräßen*, *fräßest*, *fräßet*, *fraßt*, *fresse*, *fressen*, *fressend*, *fressest*, *fresset*, *fresst*, *friss*, *frisst*, *gefressen*}, {*galt*, *gälte*, *galten*, *gälten*, *galtest*, *gältest*, *galtet*, *gältet*, *gegolten*, *gelte*, *gelten*, *geltend*, *geltest*, *geltet*, *gilt*, *giltst*, *gölte*, *gölten*, *göltest*, *göltet*}, {*geglichen*, *gleich*, *gleiche*, *gleichen*, *gleichend*, *gleichest*, *gleichet*, *gleichst*, *gleicht*, *glich*, *gliche*, *glichen*, *glichest*, *glichet*, *glichst*, *glicht*}, {*gegriffen*, *greif*, *greife*, *greifen*, *greifend*, *greifest*, *greifet*, *greifst*, *greift*, *griff*, *griffe*, *griffen*, *griffest*, *griffet*, *griffst*, *grifft*}, {*gehangen*, *häng*, *hänge*, *hängen*, *hängend*, *hängest*, *hänget*, *hängst*, *hängt*, *hing*, *hinge*, *hingen*, *hingest*, *hinget*, *hingst*, *hingt*}, {*gehoben*, *haben*, *heb*, *hebe*, *heben*, *hebend*, *hebest*, *hebet*, *hebst*, *hebt*, *hob*, *höbe*, *hoben*, *höben*, *höbest*, *höbet*, *hobst*, *hobt*, *hübe*, *hüben*, *hübest*, *hübet*}, {*gekrochen*, *kriech*, *krieche*, *kriechen*, *kriechend*, *kriechest*, *kriechet*, *kriechst*, *kriecht*, *kroch*, *kröche*, *krochen*, *kröchen*, *kröchest*, *kröchet*, *krochst*, *krocht*}, {*gelaufen*, *lauf*, *laufe*, *laufen*, *laufend*, *laufest*, *laufet*, *läufst*, *lauft*, *läuft*, *lief*, *liefe*, *liefen*, *liefest*, *liefet*, *liefst*, *lieft*}, {*gelesen*, *las*, *läse*, *lasen*, *läsen*, *läsest*, *läset*, *last*, *lese*, *lesen*, *lesend*, *lesest*, *leset*, *lest*, *lies*, *liest*}, {*genommen*, *nahm*, *nähme*, *nahmen*, *nähmen*, *nähmest*, *nähmet*, *nahmst*, *nahmt*, *nehme*, *nehmen*, *nehmend*, *nehmest*, *nehmet*, *nehmt*, *nimm*, *nimmst*, *nimmt*}, {*geritten*, *reite*, *reiten*, *reitend*, *reitest*, *reitet*, *ritt*, *ritte*, *ritten*, *rittest*, *rittet*}, {*gerufen*, *rief*, *riefe*, *riefen*, *riefest*, *riefet*, *riefst*, *rieft*, *ruf*, *rufe*, *rufen*, *rufend*, *rufest*, *rufet*, *rufst*, *ruft*}, {*geschlafen*, *schlaf*, *schlafe*, *schlafen*, *schlafend*, *schlafest*, *schlafet*, *schläfst*, *schlaft*, *schläft*, *schlief*, *schliefe*, *schliefen*, *schliefest*, *schliefet*, *schliefst*, *schlieft*}, {*geschlossen*, *schließ*, *schließe*, *schließen*, *schließend*, *schließest*, *schließet*, *schließt*, *schloss*, *schlösse*, *schlossen*, *schlössen*, *schlossest*, *schlössest*, *schlosset*, *schlösset*, *schlosst*}, {*geschrieben*, *schreibe*, *schreiben*, *schreibend*, *schreibest*, *schreibet*, *schreibst*, *schreibt*, *schrieb*, *schriebe*, *schrieben*, *schriebest*, *schriebet*, *schriebst*, *schriebt*}, {*gesoffen*, *sauf*, *saufe*, *saufen*, *saufend*, *saufest*, *saufet*, *säufst*, *sauft*, *säuft*, *soff*, *söffe*, *soffen*, *söffen*, *söffest*, *söffet*, *soffst*, *sofft*}, {*gestoßen*, *stieß*, *stieße*, *stießen*, *stießest*, *stießet*, *stießt*, *stoß*, *stoße*, *stoßen*, *stoßend*, *stoßest*, *stoßet*, *stoßt*, *stößt*}, {*getroffen*, *traf*, *träfe*, *trafen*, *träfen*, *träfest*, *träfet*, *trafst*, *traft*, *treffe*, *treffen*, *treffend*, *treffest*, *treffet*, *trefft*, *triff*, *triffst*, *trifft*}, {*gezogen*, *zieh*, *ziehe*, *ziehen*, *ziehend*, *ziehest*, *ziehet*, *ziehst*, *zieht*, *zog*, *zöge*, *zogen*, *zögen*, *zögest*, *zöget*, *zogst*, *zogt*}.

It might be noted that our input list above also includes some irregular subjunctive forms, which are rarely encountered in modern documents.

5.2.3 Heuristic detection of compounds

The following algorithm for heuristic detection of German compounds differs from the Danish version (Algorithm 5.12) only in some specific details. To make the context clear, we still state the algorithm in full. (In what follows, the string minus operation $\hat{\beta} \ominus \hat{\alpha}$ is prescribed by Definition 5.11.)

Algorithm 5.23 (Heuristic identification of German binary compounds)**.** *Let $\Lambda^{\hat{\rho}} = \{\hat{\rho}_1, \ldots, \hat{\rho}_Q\}$ be a list of distinct German essential roots (without vowel blotting) that contain at least one instance of* $\mathbf{V} = (a|e|\hat{e}|i|\hat{i}|o|u)$ *and DO NOT match the following string patterns:*

$$(ab|al|am|ar|arg|bar|be|beg|bek|be\kappa|\hat{e}n|fal|ge|gros|isç|los|mal|man|miss|naç|nis|sam|sçaf|sçaft|tal|u|un|ung|ver|vers|xüber).$$

The the output of the function $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is obtained through the following procedures:*

(1) *Construct a list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$ *where* $\lambda_q^{\hat{\rho}} = \{\hat{\rho}_{(q,1)}, \ldots, \hat{\rho}_{(q,n_q)}\}$ *is a subset of $\Lambda^{\hat{\rho}}$ whose members all match the string pattern $\hat{\rho}_q{\sim}$, for $q \in \mathbb{Z} \cap [1, Q]$.*

(2) *Expand the aforementioned entry* $(\hat{\rho}_q, \lambda_q^{\hat{\rho}})$ *into a list of triplets* $\{(\hat{\rho}_{(q,1)}, \hat{\rho}_q, \hat{\rho}_{(q,1)} \ominus \hat{\rho}_q), \ldots, (\hat{\rho}_{(q,n_q)}, \hat{\rho}_q, \hat{\rho}_{(q,n_q)} \ominus \hat{\rho}_q)\}$ *for every $q \in \mathbb{Z} \cap [1, Q]$ such that $\lambda_q^{\hat{\rho}} \neq \varnothing$. Collect all these triplets as one runs through the list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$*. The list of these triplets* $\{(\hat{\rho}_{\langle 1\rangle}, \hat{\eta}_{\langle 1\rangle}, \hat{\rho}_{\langle 1\rangle} \ominus \hat{\eta}_{\langle 1\rangle}), \ldots, (\hat{\rho}_{\langle Q'\rangle}, \hat{\eta}_{\langle Q'\rangle}, \hat{\rho}_{\langle Q'\rangle} \ominus \hat{\eta}_{\langle Q'\rangle})\}$ *contains potentially valid decompositions of compounds.*

(3) *Screen the aforementioned list of triplets as follows: for every $q' \in \mathbb{Z} \cap [1, Q']$, if* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \hat{\tau}_{\langle q'\rangle} = \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle})$ *satisfies*

$$\ell(\hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle}) \geq 2 \quad \textbf{AND} \quad \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle} = \mathbf{X}_1(a|e|\hat{e}|i|\hat{i}|o|u)\mathbf{X}_2,$$

then construct $\hat{\tau}^{*}_{\langle q'\rangle}$ *by performing* $(e|n|s){\sim}\to\varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$ *and* $\hat{\tau}^{**}_{\langle q'\rangle}$ *by doing* $(en|er|es|ge){\sim}\to\varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$*, before generating a list* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *by members of $\Lambda^{\hat{\rho}}$ that match the pattern* $\underline{(\hat{\tau}_{\langle q'\rangle}|\hat{\tau}^{*}_{\langle q'\rangle}|\hat{\tau}^{**}_{\langle q'\rangle})}$*; otherwise, set* $\lambda^{\hat{\tau}}_{\langle q'\rangle} = \varnothing$*.*

(4) *Collect all the triplets* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \lambda^{\hat{\tau}}_{\langle q'\rangle})$ *where* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *is non-void and* $\hat{\tau}_{\langle q'\rangle}$ *DOES NOT match the following string patterns:*

$$(h\hat{e}t\mathbf{X}|li\varsigma\mathbf{X}|k\hat{e}t\mathbf{X}|um|ung\mathbf{X}).$$

This list of triplets $\mathrm{CpdDet}(\Lambda_{\hat{\rho}})$ *contains the heuristic decompositions of all the identified binary compounds.*

Algorithm 5.24 (Approximate clustering of German words with heuristic detection of compounds)**.** *The procedure runs essentially the same way as Algorithm 5.24, except that German rules replace Danish rules.*

Example 5.24.1. Testing the algorithm above against the combined inputs from Examples 5.22.1 and 5.22.2, we obtain the following result:

{*Apfel*, *Äpfel*, *Äpfelchen*, *Äpfeln*, *Apfels*}, {*Arbeit*, *arbeite*, *arbeiten*, *Arbeiten*, *Arbeiter*, *Arbeitern*, *Arbeiters*, *arbeitest*, *arbeitet*, *arbeitete*, *arbeiteten*, *Arbeitsplatz*, *Arbeitsplatze*, *Arbeitsplätze*, *Arbeitsplätzen*, *Arbeitsplatzes*, *Arbeitstag*, *Arbeitstage*, *Arbeitstagen*, *Arbeitstages*, *Arbeitstags*, *gearbeitet*}, {*Arbeitsplatz*, *Arbeitsplatze*, *Arbeitsplätze*, *Arbeitsplätzen*, *Arbeitsplatzes*, *Platz*, *Platze*, *Plätze*, *Plätzen*, *Platzes*}, {*Arbeitstag*, *Arbeitstage*, *Arbeitstagen*, *Arbeitstages*, *Arbeitstags*, *Tag*, *Tage*, *Tagen*, *Tages*, *Tags*}, {*befahl*, *befähle*, *befahlen*, *befählen*, *befählest*, *befählet*, *befahlst*, *befahlt*, *befehle*, *befehlen*, *befehlend*, *befehlest*, *befehlet*, *befehlt*, *befiehlst*, *befiehlt*, *befohlen*}, {*begann*, *begänne*, *begannen*, *begännen*, *begännest*, *begännet*, *begannst*, *begannt*, *beginn*, *beginne*, *beginnen*, *beginnend*, *beginnest*, *beginnet*, *beginnst*, *beginnt*, *begönne*, *begonnen*, *begönnen*, *begönnest*, *begönnet*}, {*beiß*, *beiße*, *beißen*, *beißend*, *beißest*, *beißet*, *beißt*, *biss*, *bisse*, *bissen*, *bissest*, *bisset*, *bisst*, *gebissen*}, {*blas*, *blase*, *blasen*, *blasend*, *blasest*, *blaset*, *blast*, *bläst*, *blies*, *bliese*, *bliesen*, *bliesest*, *blieset*, *bliest*, *geblasen*}, {*brach*, *bräche*, *brachen*, *brächen*, *brächest*, *brächet*, *brachst*, *bracht*, *breche*, *brechen*, *brechend*, *brechest*, *brechet*, *brecht*, *brich*, *brichst*, *bricht*, *gebrochen*}, {*brät*, *brate*, *braten*, *bratend*, *bratest*, *bratet*, *brätst*, *briet*, *briete*, *brieten*, *brietest*, *brietet*, *gebraten*}, {*Bus*, *Busse*, *Buße*, *Bussen*, *Bußen*, *Busses*}, {*dunkel*, *dunkelste*, *dunkelstem*, *dunkelsten*, *dunkelster*, *dunkelstes*, *dunkle*, *dunklem*, *dunklen*, *dunkler*, *dunklere*, *dunklerem*, *dunkleren*, *dunklerer*, *dunkleres*, *dunkles*}, {*Ei*, *Eie*, *Eier*, *Eiern*, *Eies*, *Eigelb*, *Eigelbe*, *Eigelben*, *Eigelbs*}, {*Eigelb*, *Eigelbe*, *Eigelben*, *Eigelbs*, *gelb*, *gelbe*, *gelbem*, *gelben*, *gelber*, *gelbere*, *gelberem*, *gelberen*, *gelberer*, *gelberes*, *gelbes*, *gelbste*, *gelbstem*, *gelbsten*, *gelbster*, *gelbstes*}, {*Eis*, *Eise*, *Eises*}, {*fahr*, *fahre*, *fahren*, *fahrend*, *fahrest*, *fahret*, *fährst*, *fahrt*, *fährt*, *fuhr*, *führe*, *fuhren*, *führen*, *führest*, *führet*, *fuhrst*, *fuhrt*, *gefahren*}, {*fall*, *falle*, *fallen*, *fallend*, *fallest*, *fallet*, *fällst*, *fallt*, *fällt*, *fiel*, *fiele*, *fielen*, *fielest*, *fielet*, *fielst*, *fielt*, *gefallen*}, {*fand*, *fände*, *fanden*, *fänden*, *fandest*, *fändest*, *fandet*, *fändet*, *finde*, *finden*, *findend*, *findest*, *findet*, *gefunden*}, {*fechte*, *fechten*, *fechtend*, *fechtest*, *fechtet*, *ficht*, *fichtst*, *focht*, *föchte*, *fochten*, *föchten*, *fochtest*, *föchtest*, *fochtet*, *föchtet*, *gefochten*}, {*flieg*, *fliege*, *fliegen*, *fliegend*, *fliegest*, *flieget*, *fliegst*, *fliegt*, *flog*, *flöge*, *flogen*, *flögen*, *flögest*, *flöget*, *flogst*, *flogt*, *geflogen*}, {*fraß*, *fräße*,

fraßen, *fräßen*, *fräßest*, *fräßet*, *fraßt*, *fresse*, *fressen*, *fressend*, *fressest*, *fresset*, *fresst*, *friss*, *frisst*, *gefressen*}, {*frei*, *freiem*, *freien*, *freier*, *freiere*, *freierem*, *freieren*, *freierer*, *freieres*, *freies*, *Freiheit*, *Freiheiten*}, {*freie*, *freiste*, *freistem*, *freisten*, *freister*, *freistes*}, {*galt*, *gälte*, *galten*, *gälten*, *galtest*, *gältest*, *galtet*, *gältet*, *gegolten*, *gelte*, *gelten*, *geltend*, *geltest*, *geltet*, *gilt*, *giltst*, *gölte*, *gölten*, *göltest*, *göltet*}, {*geglichen*, *gleich*, *gleiche*, *gleichen*, *gleichend*, *gleichest*, *gleichet*, *gleichst*, *gleicht*, *glich*, *gliche*, *glichen*, *glichest*, *glichet*, *glichst*, *glicht*}, {*gegriffen*, *greif*, *greife*, *greifen*, *greifend*, *greifest*, *greifet*, *greifst*, *greift*, *griff*, *griffe*, *griffen*, *griffest*, *griffet*, *griffst*, *grifft*}, {*gehangen*, *häng*, *hänge*, *hängen*, *hängend*, *hängest*, *hänget*, *hängst*, *hängt*, *hing*, *hinge*, *hingen*, *hingest*, *hinget*, *hingst*, *hingt*}, {*gehaust*, *haus*, *Haus*, *Häuschen*, *hause*, *Hause*, *hausen*, *hausend*, *Häuser*, *Hauses*, *haust*, *hauste*, *haustest*}, {*gehoben*, *haben*, *heb*, *hebe*, *heben*, *hebend*, *hebest*, *hebet*, *hebst*, *hebt*, *hob*, *höbe*, *hoben*, *höben*, *höbest*, *höbet*, *hobst*, *hobt*, *hübe*, *hüben*, *hübest*, *hübet*}, {*gekrochen*, *kriech*, *krieche*, *kriechen*, *kriechend*, *kriechest*, *kriechet*, *kriechst*, *kriecht*, *kroch*, *kröche*, *krochen*, *kröchen*, *kröchest*, *kröchet*, *krochst*, *krocht*}, {*geküsst*, *Kuss*, *Kuß*, *Küsschen*, *Kusse*, *küsse*, *Küsse*, *küssen*, *Küssen*, *Kusses*, *küssest*, *küsst*, *küsste*, *küsstest*, *küsstet*}, {*gelaufen*, *lauf*, *laufe*, *laufen*, *laufend*, *laufest*, *laufet*, *läufst*, *lauft*, *läuft*, *lief*, *liefe*, *liefen*, *liefest*, *liefet*, *liefst*, *lieft*}, {*gelesen*, *las*, *läse*, *lasen*, *läsen*, *läsest*, *läset*, *last*, *lese*, *lesen*, *lesend*, *lesest*, *leset*, *lest*, *lies*, *liest*}, {*genommen*, *nahm*, *nähme*, *nahmen*, *nähmen*, *nähmest*, *nähmet*, *nahmst*, *nahmt*, *nehme*, *nehmen*, *nehmend*, *nehmest*, *nehmet*, *nehmt*, *nimm*, *nimmst*, *nimmt*}, {*geritten*, *reite*, *reiten*, *reitend*, *reitest*, *reitet*, *ritt*, *ritte*, *ritten*, *rittest*, *rittet*}, {*gerufen*, *rief*, *riefe*, *riefen*, *riefest*, *riefet*, *riefst*, *rieft*, *ruf*, *rufe*, *rufen*, *rufend*, *rufest*, *rufet*, *rufst*, *ruft*}, {*geschlafen*, *schlaf*, *schlafe*, *schlafen*, *schlafend*, *schlafest*, *schlafet*, *schläfst*, *schlaft*, *schläft*, *schlief*, *schliefe*, *schliefen*, *schliefest*, *schliefet*, *schliefst*, *schlieft*}, {*geschlossen*, *schließ*, *schließe*, *schließen*, *schließend*, *schließest*, *schließet*, *schließt*, *schloss*, *schlösse*, *schlossen*, *schlössen*, *schlossest*, *schlössest*, *schlosset*, *schlösset*, *schlosst*}, {*geschrieben*, *schreibe*, *schreiben*, *schreibend*, *schreibest*, *schreibet*, *schreibst*, *schreibt*, *schrieb*, *schriebe*, *schrieben*, *schriebest*, *schriebet*, *schriebst*, *schriebt*}, {*gesoffen*, *sauf*, *saufe*, *saufen*, *saufend*, *saufest*, *saufet*, *säufst*, *sauft*, *säuft*, *soff*, *söffe*, *soffen*, *söffen*, *söffest*, *söffet*, *soffst*, *sofft*}, {*gestoßen*, *stieß*, *stieße*, *stießen*, *stießest*, *stießet*, *stießt*, *stoß*, *stoße*, *stoßen*, *stoßend*, *stoßest*, *stoßet*, *stoßt*, *stößt*}, {*getroffen*, *traf*, *träfe*, *trafen*, *träfen*, *träfest*, *träfet*, *trafst*, *traft*, *treffe*, *treffen*, *treffend*, *treffest*, *treffet*, *trefft*, *triff*, *triffst*, *trifft*}, {*gezogen*, *zieh*, *ziehe*, *ziehen*, *ziehend*, *ziehest*, *ziehet*, *ziehst*, *zieht*, *zog*, *zöge*, *zogen*, *zögen*, *zögest*, *zöget*, *zogst*, *zogt*}, {*gross*, *groß*, *große*, *großem*, *großen*, *großer*, *größere*, *größerem*, *größeren*, *größerer*, *größeres*, *großes*, *größte*, *größtem*, *größten*, *größter*, *größtes*}, {*Öl*, *Öle*, *Ölen*, *Öles*, *Öls*}, {*studier*, *studiere*, *studieren*, *studierend*, *studierest*, *studieret*, *studierst*, *studiert*, *studierte*, *studierten*, *studiertest*, *studiertet*}.

Here, the compounds *Arbeitsplatz* "workplace", *Arbeitstag* "workday" and *Eigelb* "egg yolk" are correctly dissolved into their respective constituting components. In the meantime, we note that *bisse* "bite (a subjunctive form)" is not confused with the noun *Busse* "buses (nominative, genitive or accusative)", because the vowel alternation pattern [*i*, *u*] is disallowed in German tags (effective spellings or essential roots), unlike the case in English. Actual German verb conjugations do exhibit alternations between the vowels *i* and *u*, when the vowel in question is followed by *nd*, *ng* or *nk*. Our Algorithm 5.15(3) takes care of this by relabeling *i* in such scenarios as *î*, so that the vowel alternation pattern [*î*, *u*] will be later permitted in Algorithm 5.20.

Example 5.24.2. In Fig. S5, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text sources).

According to German orthography, diacritical marks in the letters *ä*, *ö*, *ü* are ignored during alphabetical sorting. This rule, if implemented literally, sometimes reduces readability of certain stacked vowels in our word clusters. Therefore, in Fig. S5, we make a compromise: the letters *ä*, *ö*, *ü* are sorted after *z* during alphabetization, as in Swedish orthography.

It should be noted that we have used a German version for *Jane Eyre* that contains pre-1902 spellings. For example, in Fig. S5a′–B′, we see *Thee* "tea" and *Thür* "door" instead of their modern counterparts *Tee* and *Tür*. Fortunately, our stemming algorithm is not adversely affected by such archaic orthography.

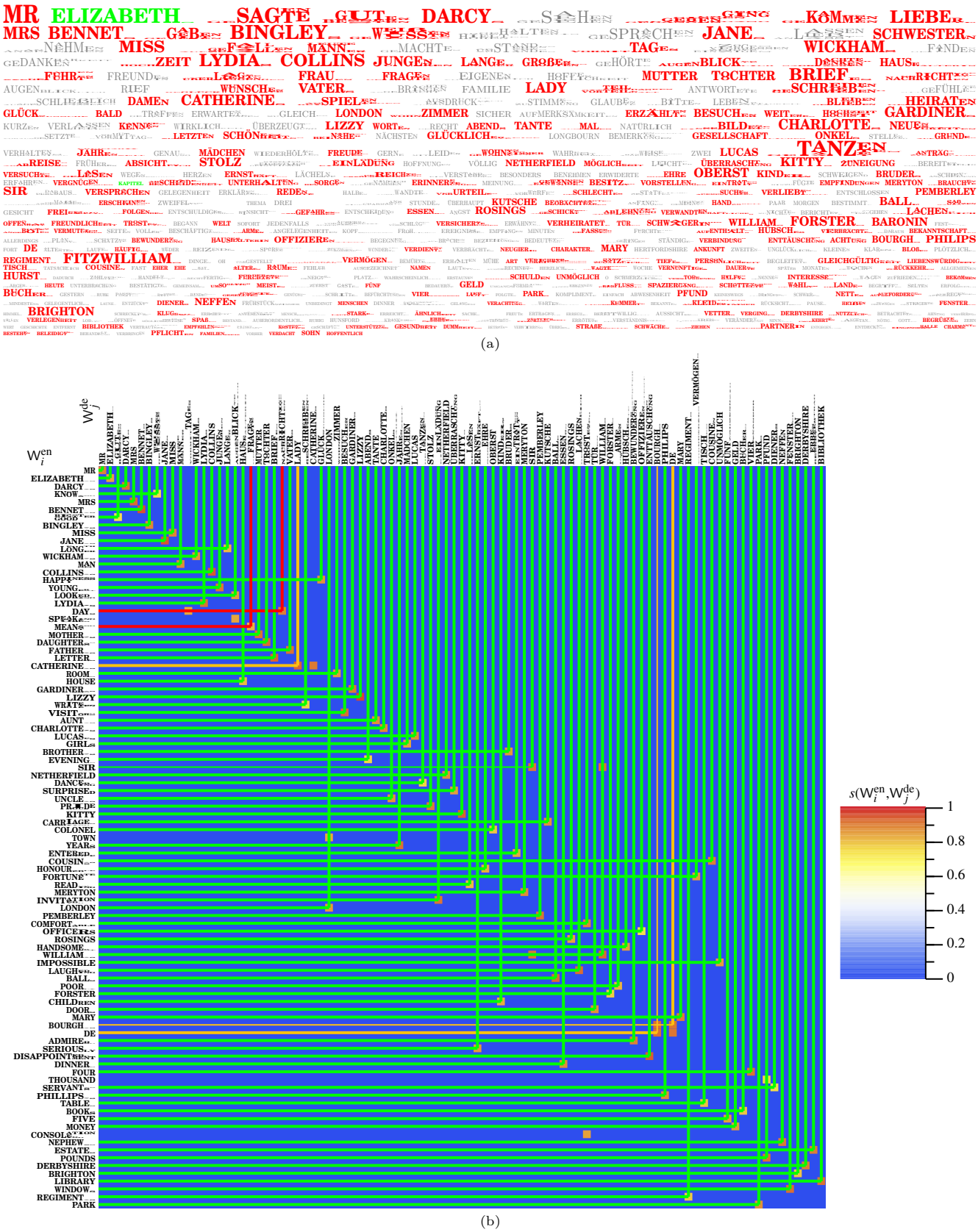


Fig. S5. Text mining in German. (a) Statistically identified topics ($n_{ii} \geq 20$) in a German version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{de})$ between selected topics in English and German versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.

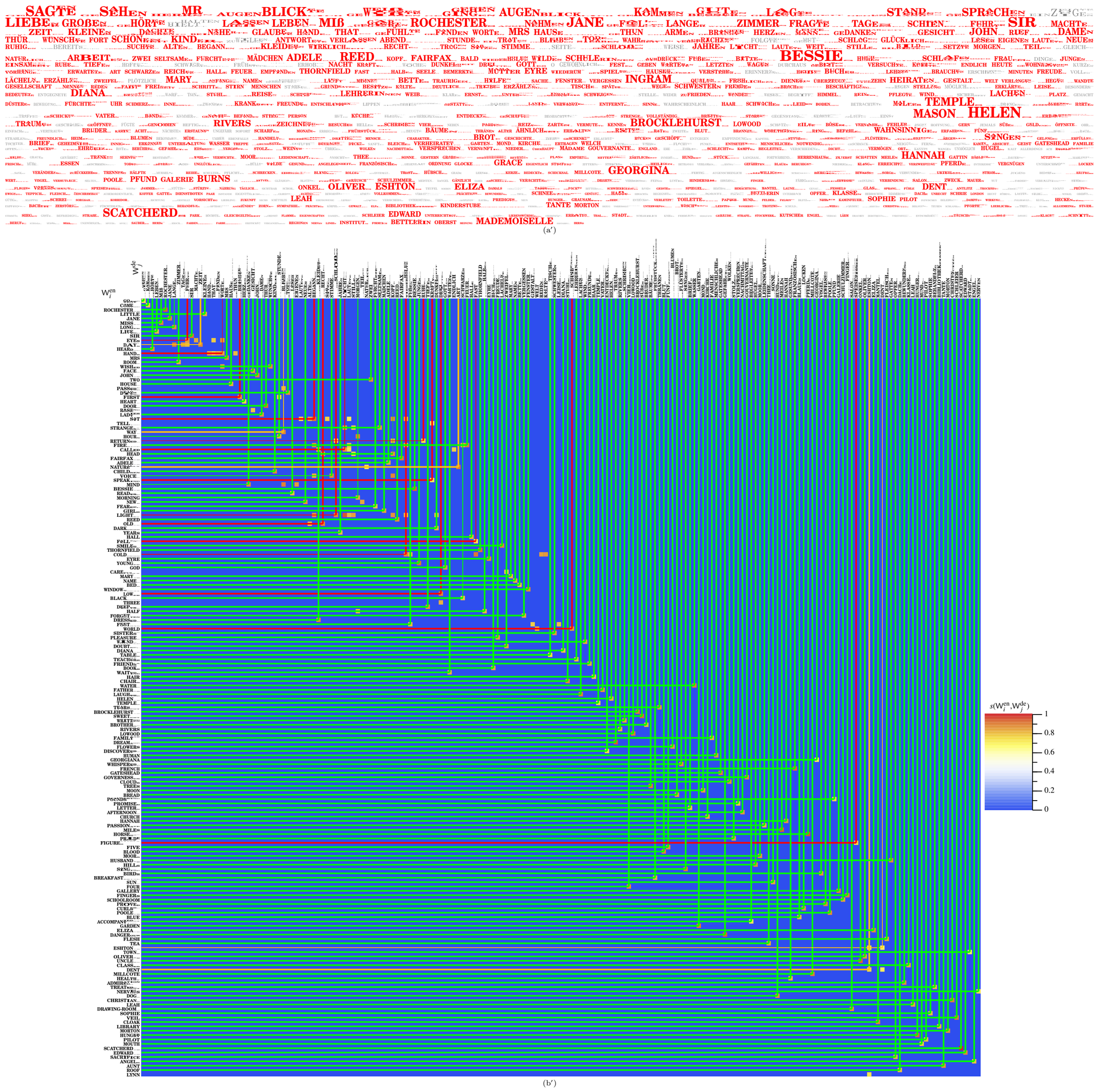


Fig. S5. Text mining in German. (*Continued*) (a′) Statistically identified topics ($n_{ii} \geq 20$) in a German version of *Jane Eyre*, with the same color encoding scheme as Fig. S3. (b′) Semantic similarities $s(\mathrm{W}_i^{\mathrm{en}}, \mathrm{W}_j^{\mathrm{de}})$ between selected topics in English and German versions of *Jane Eyre*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c′) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

5.3 Modified Porter stemming algorithm for Dutch

Definition 5.25 (Dutch stop words)**.** If a word belongs to the following list[64]:

> *'k*, *'n*, *'ne*, *'ns*, *'s*, *aan*, *achter*, *achteruit*, *af*, *al*, *allang*, *alle*, *allebei*, *allemaal*, *allen*, *alles*, *als*, *altijd*, *ander*, *andere*, *anders*, *behalve*, *beide*, *beiden*, *ben*, *beneden*, *bent*, *boven*, *bovendien*, *bij*, *bijna*, *d'n*, *daar*, *daardoor*, *daarin*, *daarmee*, *daarna*, *daarnaast*, *daarom*, *daaronder*, *daarop*, *daartoe*, *daarvan*, *dan*, *dat*, *datgene*, *datgenen*, *de*, *deden*, *deed*, *deedt*, *degene*, *degenen*, *den*, *der*, *dergelijk*, *dergelijke*, *dergelijks*, *des*, *desondanks*, *deze*, *dezelfde*, *dezen*, *dezer*, *dezes*, *die*, *diegene*, *diegenen*, *dien*, *diens*, *dier*, *dikwijls*, *dit*, *doch*, *doe*, *doen*, *doend*, *doet*, *door*, *dus*, *echt*, *echter*, *een*, *eenieder*, *eenmaal*, *eens*, *eigenlijk*, *elders*, *elk*, *elkaar*, *elkaars*, *elke*, *en*, *ene*, *enen*, *ener*, *enig*, *enige*, *enkel*, *enkele*, *enkels*, *er*, *erg*, *ergens*, *erin*, *ermee*, *ernaast*, *erom*, *eronder*, *erop*, *ertoe*, *ervan*, *even*, *eveneens*, *evengoed*, *evenmin*, *eventjes*, *eventueel*, *evenwel*, *evenzeer*, *ge*, *gedaan*, *geen*, *gehad*, *gekund*, *gemoeten*, *gemogen*, *genoeg*, *geweest*, *geworden*, *gij*, *haar*, *had*, *hadden*, *hadt*, *hare*, *heb*, *hebbe*, *hebben*, *hebbend*, *hebt*, *heeft*, *heel*, *heen*, *hele*, *helemaal*, *hem*, *hen*, *het*, *hetgeen*, *hetwelk*, *hetzelfde*, *hier*, *hierin*, *hiermee*, *hiernaast*, *hierom*, *hieronder*, *hierop*, *hiertoe*, *hiervan*, *hiervoor*, *hoe*, *hoelang*, *hoeveel*, *hoewel*, *hun*, *hunne*, *hunner*, *hij*, *ieder*, *iedere*, *iedereen*, *iemand*, *iets*, *ik*, *in*, *intussen*, *is*, *ja*, *je*, *jezelf*, *jou*, *jouw*, *jouwe*, *jouwer*, *jullie*, *jij*, *jijzelf*, *k*, *kan*, *kon*, *konden*, *kondt*, *kunnen*, *kunnend*, *kunt*, *later*, *luttel*, *luttele*, *maar*, *mag*, *me*, *mee*, *meer*, *meest*, *meeste*, *men*, *menig*, *menige*, *met*, *meteen*, *minder*, *mindere*, *minderen*, *minst*, *minste*, *minstens*, *misschien*, *mocht*, *mochte*, *mochten*, *moest*, *moeste*, *moesten*, *moet*, *moete*, *moeten*, *moetend*, *moge*, *mogen*, *mogend*, *moogt*, *mij*, *mijn*, *mijne*, *mijner*, *mijzelf*, *n*, *na*, *naar*, *naartoe*, *naast*, *nauwelijks*, *nee*, *neen*, *nen*, *nergens*, *net*, *niemand*, *niet*, *niets*, *niettemin*, *noch*, *nog*, *nooit*, *nou*, *nu*, *of*, *om*, *omdat*, *onder*, *onmiddellijk*, *ons*, *onszelf*, *onze*, *onzer*, *ooit*, *ook*, *op*, *over*, *per*, *reeds*, *sindsdien*, *sint*, *sommige*, *sommigen*, *soms*, *straks*, *t*, *te*, *tegen*, *ten*, *tenminste*, *tenslotte*, *ter*, *terug*, *tevens*, *toch*, *toe*, *toen*, *tot*, *trouwens*, *u*, *uit*, *uw*, *uwe*, *uwer*, *uzelf*, *vaak*, *vaaks*, *vaakst*, *vake*, *vaker*, *van*, *vanaf*, *vandaan*, *vanwaar*, *veel*, *vele*, *voor*, *vooral*, *vooruit*, *vroeger*, *waar*, *waarheen*, *waarin*, *waarmee*, *waarnaast*, *waarom*, *waaronder*, *waarop*, *waart*, *waartoe*, *waarvandaan*, *wanneer*, *want*, *waren*, *was*, *wat*, *we*, *weer*, *wees*, *weest*, *weinig*, *weinige*, *wel*, *welk*, *welke*, *wellicht*, *werd*, *werden*, *werdt*, *wezen*, *wie*, *wil*, *word*, *worden*, *wordend*, *wordt*, *wij*, *zal*, *ze*, *zeer*, *zelden*, *zelf*, *zich*, *zichzelf*, *zo*, *zoals*, *zojuist*, *zolang*, *zonder*, *zou*, *zoude*, *zouden*, *zoudt*, *zoveel*, *zulk*, *zulke*, *zulks*, *zulle*, *zullen*, *zullend*, *zult*, *zij*, *zijn*, *zijnd*, *zijne*, *zijner*, *zijt*,

then we consider it a Dutch stop word. All the Dutch stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

To accommodate to pre-1946 (Flanders)/pre-1947 (Netherlands) Dutch written standard, which employed a sophisticated case system as in modern German (see `https://en.wikipedia.org/wiki/Archaic_Dutch_declension`), we need to add inflected forms of the definite and indefinite articles, as well as those of the demonstratives to the list of stop words. It should be noted that archaic Dutch declensions still survive in some stock phrases in modern Dutch, such as the genitive construction in *het Koninkrijk der Nederlanden* "the Kingdom of Netherlands".

The 1946/1947 orthography reform changed the endings of certain words: *bosch* "forest", *mensch* "human" and *visch* "fish" (singular noun) are now spelt *bos*, *mens* and *vis*, respectively [47, §2.1]. Meanwhile, the spellings of adjectival endings in *mathematisch* "mathematical" and *typisch* "typical" are not affected, despite having the same silent *ch* at word final positions. So long as a document consistently employs an orthographical standard, we do not need to worry about these changes in our clustering algorithm, and we will not forcibly change all word final *sch* to *s*.

5.3.1 Effective spelling and essential root

Vowel lengths are indicated in Dutch spellings according to certain rules. For example, all these three words *boom* "tree", *boomen* "trees (pre-1946/pre-1947)" and *bomen* "trees (post-1946/post-1947)" contain the same long vowel in their first syllable [47, §2.1]. The penultimate step in the following algorithm helps us detect the vowel length in a Dutch spelling heuristically.

Algorithm 5.26 (Dutch effective spelling)**.** *Set* $\mathbf{V} = (a|A|e|\hat{e}|E|i|\hat{\imath}|o|\hat{o}|O|u|\hat{u}|\ddot{u}|U|y)$*,* $\mathbf{V}_* = (a|e|o|u)$ *and* $\mathbf{V}_0 = (a|e|i|o|u)$*. Set*

$$\mathbf{c} = (b|c|d|f|g|h|j|k|l|m|n|p|q|r|s|t|v|w|x|z) \text{ and } \mathbf{C} = (B|C|D|F|G|H|J|K|L|M|N|P|Q|R|S|T|V|W|X|Z).$$

For a Dutch word $\hat{\sigma}$*, its effective spelling* $\mathrm{EffSpell}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Convert to lowercase, and replace*

[64] Our list of Dutch stop words is based on `snowball.tartarus.org/algorithms/dutch/stop.txt`, with extensive additions that roughly match their counterparts in English. We have also included these words: *d'n*, *den*, *der*, *des*, *deze*, *dezen*, *dezer*, *dezes*, *dien*, *diens*, *dier*, *eens*, *eens*, *ene*, *enen*, *ener*, *'n*, *'ne*, *nen*, *'ns*, *'s*, which are involved in archaic Dutch declensions and records of colloquial speech.

(∅|ge)zegen~	aanwezig**X**	afwezig**X**	bezet~	compl~	diner	eerb~	eerder~	ellend	forst~	gardiner
βλεσn	πρεsεnτ	αβσεnτ	βζet	κμλ	δineρ	ehrβ	vroeg	μισρβd	φορστ	φarδiner

gelo$^{\mathbf{X}\in}$(of|v)~	genegen~	gezag	gezicht	jone~	kilo~	kind~	langza~	ma(∅|m)ma(∅|a)~	manier~
βλlo**X**	nijgen	αυτρg	φασt	γοṽε	κιλο	κiλδd	σλοωza	moeder	μaṽierρr

meester	negen~	o(or|re)~	pappa(∅|a)~	ped~	plant	schok	schoonh~	tafel	triest~	vier~	wit
μeeστer	9e	εaρ	vader	peδ	plantτ	σχok	mooih	ταβλl	τριeστt	4er	ωiτt

wonder	(∅|ge)rend(∅|e|en)	jane	ma('s|'s|atje)(∅|s)	mamme(n|tje)(∅|s)	oom(∅|s)	~mooie**X**
wonδer	rennen	γιοṽα	moeder	moeder	oṽκλ	mooi

(2) *Replace*[65]

(∅|ge)hoor	(∅|ge)zegd	$^{\mathbf{X}\in}$(bra(af|v))~	$^{\mathbf{X}\in}$(do(d|od))~	á	ä	bruik	docht~	é	è	ë	geëe~	geget	georg~
χoor	zeg	p**X**	t**X**	a	a	βruik	dtocht	e	e	e	geE	et	jgeorg

gezond	hor	í	ï	janu~	koning	leraar	lie$^{\mathbf{X}\in}$(f|v)(∅|st)~	lof	loof	los	lov	moe(∅|ie)~	ó	ö
jgezond	χor	i	i	jnu	κoning	leren	xlî**X**	λof	λoof	xlos	λov	mμoe	o	o

oe	ou	placht	su~	tien	tuin	ú	ü	voeld	ij	zicht	ziek~	zom~	zoon~	best(∅|e)
ô	û	pleeg	xsu	tzien	τuin	u	u	fôl	y	zien	xziek	ζom	xxzoon	gôd

beter(∅|e|s)	sir	~(isme|ist)(∅|en|tje)(∅|s)	~dje(∅|s)	~ga(∅|an|and|at)
gôd	ssir	e	d	ging

~ga(aft|f|ve|ven)	~kom(∅|t)	~kwa(amt|m|me|men)	~la(agt|g|ge|gen)	~la(ast|s|ze|zen)
geven	komen	komen	liggen	lezen

~sta(∅|an|and|at)	~uk	~wist(∅|e|en)	~zaagt	~zag(∅|e|en)	~zat(∅|e|en)	~zei(∅|de|den|dt)
stond	ukk	weten	zîn	zîn	zitten	zeg

(3) *Do ei → ê, list → λist, ui → ü;*

(4) *Do ie → î, aan(∅|ge)n~ → n, door(∅|ge)r~ → r, dwars(∅|ge)s~ → s, om(∅|ge)m~ → m, op(∅|ge)p~ → p, vort(∅|ge)t~ → t, weg(∅|ge)g~ → g;*

(5) *Do ~end → e,* $\sim_{\hat{\chi}_1 \notin} \mathbf{V}s \to \hat{\chi}_1$, $\hat{\chi}_2(dom|vol)(e|l|r|s|t)_{m_0} \to \hat{\chi}_2$, $\hat{\chi}_3(ery|vaardig)\mathbf{X} \to \hat{\chi}_3$;

(6) *Do* $\hat{\chi}_1 v \to \hat{\chi}_1 f$, $\hat{\chi}_2 z \to \hat{\chi}_2 s$ *before turning double letters into the capital form of the same letter (i.e. ee → E etc.);*

(7) *Do ~en → e;*

(8) *Do* $\sim \mathbf{V}_* \mathbf{c} e \to \mathbf{V}_*^+ \mathbf{c}$, $\sim \mathbf{V}_0 \mathbf{C} e \to \mathbf{V}_0 \mathbf{C}_-$, *where* $\mathbf{V}_*^+$ *is the upper case form of* $\mathbf{V}_*$ *and* $\mathbf{C}_-$ *is the lower case form of* $\mathbf{C}$*;*

(9) *Do lerares → lEr, vrE~ → βrE, ~îr(∅|es|s|tje)(∅|s) → ∅, ~An(∅|en|tje)(∅|s) → a, ~sje(∅|s) → s, ~tje(∅|s) → ∅, ~Tje(∅|s) → t.*

Definition 5.27 (Dutch protected range)**.** Set $\mathbf{V}^* = (a|A|ê|E|i|î|o|ô|O|u|û|ü|U|y)$, and write $\overline{\mathbf{V}}_m$ (resp. $\overline{\mathbf{V}}_{m_0}$) for one (resp. zero) or more repeats of any character other than $\mathbf{V} = (a|A|e|ê|E|i|î|o|ô|O|u|û|ü|U|y)$. Define $\mathbf{C}^{**}_{m_0} = \overline{(a|A|e|ê|E|i|î|ô|O|r|u|û|ü|U)}_{m_0}$. Let $\hat{\sigma}$ be the effective spelling of a Dutch word, its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \max\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is determined by two non-negative integers $\lambda_1(\hat{\sigma})$ and $\lambda_2(\hat{\sigma})$ specified through the following procedures:

- Look for the string pattern $((\varnothing|a|A|be|e|E|ge|her|i|mede|na|O|ô|u|U|û|ü|ver)\overline{\mathbf{V}}_{m_0}\mathbf{V}^*|\overline{\mathbf{V}}_m(e|E))\mathbf{C}^{**}_{m_0}\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Look for the pattern $(a|A|o|ô|O|u|û|U)$ in the string $\hat{\sigma}$;
- If a letter in the pattern above is found, the last position occupied by such a letter defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = 0$. □

Unlike Danish, but similar to German, our Dutch essential root is free from common separable verb prefixes (see step (3) in the algorithm below).

[65] Here, the string pattern *ij* may refer to two Latin letters *i* and *j* sitting next to each other, or a single Unicode character Latin Small Ligature IJ U+0133.

Algorithm 5.28 (Dutch essential root)**.** *Set* $\mathbf{V} = (a|A|e|ê|E|i|î|o|ô|O|u|û|U|y)$ *and* $\mathbf{C} = \overline{(a|A|e|ê|E|i|î|o|ô|O|u|û|U|y)}$*. Let* $\hat{\sigma}$ *be the effective spelling of a Dutch word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$*.*

(2) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*[66]

(2.1) *Do* $\sim(Er|s) \to \varnothing$*;*

(2.2) *Do* $\sim e(d|e|m|n|r|st|t)_{m_0} \to \varnothing$*.*

The result after these two steps of operations is called $\hat{\sigma}'_2$*.*

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$*, before doing* $(\varnothing|af|An|by|dOr|dwars|mE|mede|na|om|op|tô|vOrt|weg)(\varnothing|ge)\mathbf{XV}\sim \to \mathbf{XV}$*;*

(4) *Do* $\sim\mathbf{C}t \to \mathbf{C}$*;*

(5) *Do* $v \to f$*,* $z \to s$*;*

(6) *Do* $\sim\hat{\chi} \to \hat{\chi}_-$ *(i.e. reduce final letter to lowercase);*

(7) *Do* $ag \to Ag$*,* $ak \to Ak$*,* $al \to Al$*,* $at \to At$*,* $breng \to brach$*,* $denk \to dach$*,* $Ef \to ef$*,* $ep \to Ep$*,* $kOp \to koch$*,* $nam \to nAm$*,* $sôk \to soch$*.*

5.3.2 Admissible mutation and approximate clustering

The Dutch vowel blotting mechanism is similar to the Danish version (Algorithm 5.5).

Algorithm 5.29 (Dutch vowel blotting)**.** *Set* $\mathbf{V}_m = (a|A|e|ê|E|i|î|o|ô|O|u|û|ü|U|y)_m$ *and* $\mathbf{C}_{m_0} = \overline{(a|A|e|ê|E|i|î|o|ô|O|u|û|ü|U|y)}_{m_0}$*. For a token string* $\hat{\sigma}$*, its blotted form* $\mathrm{BlotV}_1(\hat{\sigma})$ *is constructed as follows:*

- *If the string pattern* $(\varnothing|a|be|e|ê|i|o|u|ü)\mathbf{C}_{m_0}\mathbf{V}_m\sim$ *can be found in the string* $\hat{\sigma}$*, then the last position occupied by such a pattern is replaced by the letter "a".*
- *Otherwise, leave the string* $\hat{\sigma}$ *intact.*

Similar to what we did in §5.1.2 for the case of Danish, we will construct a bivariate Boolean-valued function $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ on a "simple heredity test function" in Algorithm 5.30, and a set of "admissible suffix mismatch and vowel alternation" rules in Algorithm 5.31.

Algorithm 5.30 (Simple heredity test)**.** *Set* $\mathbf{V} = (a|A|e|ê|E|i|î|o|ô|O|u|û|ü|U|y)$*. The Boolean-valued function* $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if* $\hat{\alpha}$ *contains at least one instance of* $\mathbf{V}$*,* **AND** *at least one of the following six conditions holds:*[67]

(i) $\hat{\alpha} = \hat{\beta}$*;*

(ii) $\hat{\beta} = \hat{\alpha}d$*;*

(iii) $\hat{\beta} = \hat{\alpha}Er$*;*

(iv) $\hat{\beta} = \hat{\alpha}s$*;*

(v) $\hat{\beta} = \hat{\alpha}t$*;*

(vi) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \ell(\hat{\beta}) - 2$ **AND** $\hat{\alpha} = \hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{[\ell(\hat{\alpha})+1\}} = (e|n|s)$*. (See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{[n\}}$*.)*

In what follows, we define $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})$, $\mathrm{RootNW}(\hat{\alpha},\hat{\beta})$, $\mathrm{NW}^*(\hat{\alpha},\hat{\beta})$ and $\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta})$, $\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta})$, $\mathrm{SW}^*(\hat{\alpha},\hat{\beta})$ as what is done in the Danish case (Algorithm 5.7).

[66] In other words, the core algorithm for essential root extraction runs as follows: keep the last "strong" vowel *a*, *ï*, *o* or *u* in non-final position, plus one subsequent letter; delete final *a*; erase the final appearance of *e* and all the letters thereafter.

[67] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 5.31 (Admissible suffix mismatch and vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{SuffixNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if*

$$\begin{aligned}&\text{SuffixNW}(\hat{\alpha},\hat{\beta})\\ &= [(\varnothing|(e|E|n|t)_m|((acht|haft|mat)ig|(\varnothing|b)Ar|e(ling|lyk|nis|r)|h(e|\hat{e})d|i(n|ng|sch)|je|l(ing|yk)|nis|s(Am|e|t)|t(\hat{e}t|je)|yk)\mathbf{X})),\\ &\quad(\varnothing|(e|E|n|t)_m|((acht|haft|mat)ig|(\varnothing|b)Ar|e(ling|lyk|nis|r)|h(e|\hat{e})d|i(n|ng|sch)|je|l(ing|yk)|nis|s(Am|e|t)|t(\hat{e}t|je)|yk)\mathbf{X}))]\end{aligned}$$

AND *at least one of the following three conditions holds:*[68]

(i) $\text{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$ **AND** *the lowercase form of* $\text{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* $(a|e|\hat{e}|i|\hat{\imath}|o|\hat{o}|u)$*;*

(ii) $\text{NW}^*(\hat{\alpha},\hat{\beta}) = ([A, E]|[A, i]|[A, \hat{\imath}]|[A, \hat{o}]|[A, O]|[e, \hat{\imath}]|[e, o]|[E, i]|[E, O]|[E, y]|[i, o]|[\hat{\imath}, O]|[O, \ddot{u}])$*;*

(iii) $\text{NW}^*(\hat{\beta},\hat{\alpha}) = ([A, E]|[A, i]|[A, \hat{\imath}]|[A, \hat{o}]|[A, O]|[e, \hat{\imath}]|[e, o]|[E, i]|[E, O]|[E, y]|[i, o]|[\hat{\imath}, O]|[O, \ddot{u}])$*.*

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmMut}(\text{RootSW}(\hat{\alpha},\hat{\beta}), \text{SuffixSW}(\hat{\alpha},\hat{\beta}), \text{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW*.*

Algorithm 5.32 (Heredity test function)**.** *The structure of the Dutch heredity test function* $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ *is identical to the German version (Algorithm 8.1.2), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW**,* RootSW*,* SuffixSW*,* SW** must follow the Dutch rules stated above.*

Algorithm 5.33 (Approximate clustering of Dutch words)**.** *The algorithm is essentially the same as Algorithm 5.10, except that Dutch rules (instead of Danish rules) apply to all the tags (effective spelling, essential root, vowel blotting etc.).*

Example 5.33.1. As an illustration of our clustering algorithm, we pick the following families of Dutch words, where an "approximate translation" in English, enclosed in quotation marks, is appended to the end of each family:[69]

best, *beste*, *beter*, *betere*, *beters*, *goed*, *goede*, *goeden*, *goeder*, *goeds* — "good";
daad, *daden* — "deed";
dooier, *dooiers*, *dooiertje* — "yolk";
ei, *eieren*, *eitje*, *eitjes* — "egg";
eidooier, *eidooiers*, *eidooiertje*, *eierdooier*, *eierdooiers*, *eierdooiertje* — "egg yolk";
eiwit, *eiwitten*, *eiwittje* — "egg white";
geleerd, *leer*, *leerde*, *leerden*, *leert*, *lere*, *leren*, *lerend* — "learn";
gelost, *los*, *losse*, *lossen*, *lossend*, *lost*, *loste*, *losten* — "dump";
geraasd, *raas*, *raasde*, *raasden*, *raast*, *raze*, *razen*, *razend* — "rage";
gewerkt, *werk*, *werke*, *werken*, *werkend*, *werkt*, *werkte*, *werkten* — "work";
klein, *kleine*, *kleinen*, *kleiner*, *kleinere*, *kleiners*, *kleins*, *kleinst*, *kleinste* — "small";
maan, *manen*, *maantje* — "moon";
man, *manne*, *mannen*, *mans* — "man";
vrouw, *vrouwe*, *vrouwen* — "woman";
wit, *wits*, *witst*, *witste*, *witte*, *witter*, *wittere*, *witters* — "white".

Applying Algorithm 5.33 to this list of words, we obtain the following clustering results:

{*best*, *beste*, *beter*, *betere*, *beters*, *goed*, *goede*, *goeden*, *goeder*, *goeds*}, {*daad*, *daden*}, {*dooier*, *dooiers*, *dooiertje*}, {*ei*, *eieren*, *eitje*, *eitjes*}, {*eidooier*, *eidooiers*, *eidooiertje*}, {*eierdooier*, *eierdooiers*, *eierdooiertje*}, {*eiwit*, *eiwitten*, *eiwittje*}, {*geleerd*, *leer*, *leerde*, *leerden*, *leert*, *lere*, *leren*, *lerend*}, {*gelost*, *los*, *losse*, *lossen*, *lossend*, *lost*, *loste*, *losten*}, {*geraasd*, *raas*, *raasde*, *raasden*, *raast*, *raze*, *razen*, *razend*}, {*gewerkt*, *werk*, *werke*, *werken*, *werkend*, *werkt*, *werkte*, *werkten*}, {*klein*, *kleine*, *kleinen*, *kleiner*, *kleinere*, *kleiners*, *kleins*, *kleinst*, *kleinste*}, {*maan*, *maantje*, *man*, *manen*, *manne*, *mannen*, *mans*}, {*vrouw*, *vrouwe*, *vrouwen*}, {*wit*, *wits*, *witst*, *witste*, *witte*, *witter*, *wittere*, *witters*}.

[68] We note that $\text{NW}^*(\hat{\alpha},\hat{\beta})$ and $\text{NW}^*(\hat{\beta},\hat{\alpha})$ differ only in the order of the two components in the bracket.

[69] This list includes archaic declension forms found in the paradigms on the following web-page: `en.wikipedia.org/wiki/Archaic_Dutch_declension`.

We note that two alternative forms of "egg yolk" *eidooier* and *eierdooier* are split into two different clusters in this method.

Example 5.33.2. To further test our algorithm against Dutch irregular verbs with various vowel alternation patterns across different tenses and moods, we throw the following list:

bak, *bakke*, *bakken*, *bakkend*, *bakt*, *bakte*, *bakten*, *gebakken* — "bake";
ban, *bande*, *banden*, *banne*, *bannen*, *bannend*, *bant*, *gebannen* — "expel";
bederf, *bederft*, *bederve*, *bederven*, *bedervend*, *bedierf*, *bedierft*, *bedorve*, *bedorven* — "spoil";
bedrieg, *bedriege*, *bedriegen*, *bedriegend*, *bedriegt*, *bedroge*, *bedrogen*, *bedroog*, *bedroogt* — "deceive";
beet, *bete*, *beten*, *bijt*, *bijte*, *bijten*, *bijtend*, *gebeten* — "bite";
begin, *beginne*, *beginnen*, *beginnend*, *begint*, *begon*, *begonne*, *begonnen*, *begont* — "begin";
bied, *biede*, *bieden*, *biedend*, *biedt*, *bode*, *boden*, *bood*, *boodt*, *geboden* — "bid";
bind, *binde*, *binden*, *bindend*, *bindt*, *bond*, *bonde*, *bonden*, *bondt*, *gebonden* — "tie";
blaas, *blaast*, *blaze*, *blazen*, *blazend*, *blies*, *bliest*, *blieze*, *bliezen*, *geblazen* — "blow";
bleek, *bleekt*, *bleke*, *bleken*, *blijk*, *blijke*, *blijken*, *blijkend*, *blijkt*, *gebleken* — "appear";
boge, *bogen*, *boog*, *boogt*, *buig*, *buige*, *buigen*, *buigend*, *buigt*, *gebogen* — "bend";
braakt, *brak*, *brake*, *braken*, *breek*, *breekt*, *breke*, *breken*, *brekend*, *gebroken* — "break";
bracht, *brachte*, *brachten*, *breng*, *brenge*, *brengen*, *brengend*, *brengt*, *gebracht* — "bring";
dacht, *dachte*, *dachten*, *denk*, *denke*, *denken*, *denkend*, *denkt*, *gedacht* — "think";
draag, *draagt*, *drage*, *dragen*, *dragend*, *droeg*, *droege*, *droegen*, *droegt*, *gedragen* — "carry";
droop, *droopt*, *drope*, *dropen*, *druip*, *druipe*, *druipen*, *druipend*, *druipt*, *gedropen* — "drip";
gegolden, *geld*, *gelde*, *gelden*, *geldend*, *geldt*, *gold*, *golde*, *golden*, *goldt* — "apply";
geheven, *hef*, *heffe*, *heffen*, *heffend*, *heft*, *hief*, *hieft*, *hieve*, *hieven* — "heave";
geholpen, *help*, *helpe*, *helpen*, *helpend*, *helpt*, *hielp*, *hielpe*, *hielpen*, *hielpt* — "help";
gekocht, *kocht*, *kochte*, *kochten*, *koop*, *koopt*, *kope*, *kopen*, *kopend* — "buy";
gelegen, *laagt*, *lag*, *lage*, *lagen*, *lig*, *ligge*, *liggen*, *liggend*, *ligt* — "lie";
genomen, *naamt*, *nam*, *name*, *namen*, *neem*, *neemt*, *neme*, *nemen*, *nemend* — "take";
geschapen, *schep*, *scheppe*, *scheppen*, *scheppend*, *schept*, *schiep*, *schiepe*, *schiepen*, *schiept* — "create";
gescholden, *scheld*, *schelde*, *schelden*, *scheldend*, *scheldt*, *schold*, *scholde*, *scholden*, *scholdt* — "scold";
gevraagd, *vraag*, *vraagd*, *vraagde*, *vraagden*, *vraagt*, *vrage*, *vragen*, *vragend*, *vroeg*, *vroege*, *vroegen*, *vroegt* — "ask";
gewogen, *weeg*, *weegt*, *wege*, *wegen*, *wegend*, *woge*, *wogen*, *woog*, *woogt* — "weigh";
gezeten, *zat*, *zate*, *zaten*, *zit*, *zitte*, *zitten*, *zittend* — "sit";
gezocht, *zocht*, *zochte*, *zochten*, *zoek*, *zoeke*, *zoeken*, *zoekend*, *zoekt* — "seek"

into our algorithm, and obtain

{*bak*, *bakke*, *bakken*, *bakkend*, *bakt*, *bakte*, *bakten*, *gebakken*}, {*ban*, *bande*, *banden*, *banne*, *bannen*, *bannend*, *bant*, *gebannen*}, {*bederf*, *bederft*, *bederve*, *bederven*, *bedervend*, *bedierf*, *bedierft*, *bedorve*, *bedorven*}, {*bedrieg*, *bedriege*, *bedriegen*, *bedriegend*, *bedriegt*, *bedroge*, *bedrogen*, *bedroog*, *bedroogt*}, {*beet*, *bete*, *beten*, *bijt*, *bijte*, *bijten*, *bijtend*, *gebeten*}, {*begin*, *beginne*, *beginnen*, *beginnend*, *begint*, *begon*, *begonne*, *begonnen*, *begont*}, {*bied*, *biede*, *bieden*, *biedend*, *biedt*, *bode*, *boden*, *bood*, *boodt*, *geboden*}, {*bind*, *binde*, *binden*, *bindend*, *bindt*, *bond*, *bonde*, *bonden*, *bondt*, *gebonden*}, {*blaas*, *blaast*, *blaze*, *blazen*, *blazend*, *blies*, *bliest*, *blieze*, *bliezen*, *geblazen*}, {*bleek*, *bleekt*, *bleke*, *bleken*, *blijk*, *blijke*, *blijken*, *blijkend*, *blijkt*, *gebleken*}, {*boge*, *bogen*, *boog*, *boogt*, *buig*, *buige*, *buigen*, *buigend*, *buigt*, *gebogen*}, {*braakt*, *brak*, *brake*, *braken*, *breek*, *breekt*, *breke*, *breken*, *brekend*, *gebroken*}, {*bracht*, *brachte*, *brachten*, *breng*, *brenge*, *brengen*, *brengend*, *brengt*, *gebracht*}, {*dacht*, *dachte*, *dachten*, *denk*, *denke*, *denken*, *denkend*, *denkt*, *gedacht*}, {*draag*, *draagt*, *drage*, *dragen*, *dragend*, *droeg*, *droege*, *droegen*, *droegt*, *gedragen*}, {*droop*, *droopt*, *drope*, *dropen*, *druip*, *druipe*, *druipen*, *druipend*, *druipt*, *gedropen*}, {*gegolden*, *geld*, *gelde*, *gelden*, *geldend*, *geldt*, *gold*, *golde*, *golden*, *goldt*}, {*geheven*, *hef*, *heffe*, *heffen*, *heffend*, *heft*, *hief*, *hieft*, *hieve*, *hieven*}, {*geholpen*, *help*, *helpe*, *helpen*, *helpend*, *helpt*, *hielp*, *hielpe*, *hielpen*, *hielpt*}, {*gekocht*, *kocht*, *kochte*, *kochten*, *koop*, *koopt*, *kope*, *kopen*, *kopend*}, {*gelegen*, *laagt*, *lag*, *lage*, *lagen*, *lig*, *ligge*, *liggen*, *liggend*, *ligt*}, {*genomen*, *naamt*, *nam*, *name*, *namen*, *neem*, *neemt*, *neme*, *nemen*, *nemend*}, {*geschapen*, *schep*, *scheppe*, *scheppen*, *scheppend*, *schept*, *schiep*, *schiepe*, *schiepen*, *schiept*}, {*gescholden*, *scheld*, *schelde*, *schelden*, *scheldend*, *scheldt*, *schold*, *scholde*, *scholden*, *scholdt*}, {*gevraagd*, *vraag*, *vraagd*, *vraagde*, *vraagden*, *vraagt*, *vrage*, *vragen*, *vragend*, *vroeg*, *vroege*, *vroegen*, *vroegt*}, {*gewogen*, *weeg*, *weegt*, *wege*, *wegen*, *wegend*, *woge*, *wogen*, *woog*, *woogt*}, {*gezeten*, *zat*, *zate*, *zaten*, *zit*, *zitte*, *zitten*, *zittend*}, {*gezocht*, *zocht*, *zochte*, *zochten*, *zoek*, *zoeke*, *zoeken*, *zoekend*, *zoekt*}.

5.3.3 Heuristic detection of compounds

The following algorithm for heuristic detection of Dutch compounds differs from the Danish version (Algorithm 5.12) only in some specific details. To make the context clear, we still state the algorithm in full. (In what follows, the string minus operation $\hat{\beta} \ominus \hat{\alpha}$ is prescribed by Definition 5.11.)

Algorithm 5.34 (Heuristic identification of Dutch binary compounds)**.** *Let $\Lambda^{\hat{\rho}} = \{\hat{\rho}_1, \ldots, \hat{\rho}_Q\}$ be a list of distinct Dutch essential roots (without vowel blotting) that contain at least one instance of* $\mathbf{V} = (a|A|e|\hat{e}|E|i|\hat{\imath}|o|\hat{o}|O|u|\hat{u}|\ddot{u}|U|y)$ *and DO NOT match the following string patterns:*

$$(acht\mathbf{X}|al|an|At|bar|be\mathbf{X}|con|el|en|Er|fer\mathbf{X}|fol|fOr\mathbf{X}|ge|ging|grOt|kOm|le|man|me|o|of|on|ond|onge|per|\ddot{u}|\ddot{u}t|wAr|xlos).$$

The output of the function $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is obtained through the following procedures:*

(1) *Construct a list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$ *where* $\lambda_q^{\hat{\rho}} = \{\hat{\rho}_{(q,1)}, \ldots, \hat{\rho}_{(q,n_q)}\}$ *is a subset of* $\Lambda^{\hat{\rho}}$ *whose members all match the string pattern* $\hat{\rho}_q\sim$*, for* $q \in \mathbb{Z} \cap [1, Q]$.

(2) *Expand the aforementioned entry* $(\hat{\rho}_q, \lambda_q^{\hat{\rho}})$ *into a list of triplets* $\{(\hat{\rho}_{(q,1)}, \hat{\rho}_q, \hat{\rho}_{(q,1)} \ominus \hat{\rho}_q), \ldots, (\hat{\rho}_{(q,n_q)}, \hat{\rho}_q, \hat{\rho}_{(q,n_q)} \ominus \hat{\rho}_q)\}$ *for every* $q \in \mathbb{Z} \cap [1, Q]$ *such that* $\lambda_q^{\hat{\rho}} \neq \varnothing$*. Collect all these triplets as one runs through the list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$*. The list of these triplets* $\{(\hat{\rho}_{\langle 1\rangle}, \hat{\eta}_{\langle 1\rangle}, \hat{\rho}_{\langle 1\rangle} \ominus \hat{\eta}_{\langle 1\rangle}), \ldots, (\hat{\rho}_{\langle Q'\rangle}, \hat{\eta}_{\langle Q'\rangle}, \hat{\rho}_{\langle Q'\rangle} \ominus \hat{\eta}_{\langle Q'\rangle})\}$ *contains potentially valid decompositions of compounds.*

(3) *Screen the aforementioned list of triplets as follows: for every* $q' \in \mathbb{Z} \cap [1, Q']$*, if* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \hat{\tau}_{\langle q'\rangle} = \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle})$ *satisfies*

$$\ell(\hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle}) \geq 2 \quad \textbf{AND} \quad \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle} = \mathbf{X}_1(a|A|e|\hat{e}|E|i|\hat{\imath}|o|\hat{o}|O|u|\hat{u}|\ddot{u}|U|y)\mathbf{X}_2,$$

then construct $\hat{\tau}^*_{\langle q'\rangle}$ *by performing* $(e|n|s)\sim\!\!\rightarrow \varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$ *and* $\hat{\tau}^{**}_{\langle q'\rangle}$ *by doing* $(en|er|es|ge)\sim\!\!\rightarrow \varnothing$ *on* $\hat{\tau}_{\langle q'\rangle}$*, before generating a list* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *by members of* $\Lambda^{\hat{\rho}}$ *that match the pattern* $\underline{(\hat{\tau}_{\langle q'\rangle}|\hat{\tau}^*_{\langle q'\rangle}|\hat{\tau}^{**}_{\langle q'\rangle})}$*; otherwise, set* $\lambda^{\hat{\tau}}_{\langle q'\rangle} = \varnothing$.

(4) *Collect all the triplets* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \lambda^{\hat{\tau}}_{\langle q'\rangle})$ *where* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *is non-void and* $\hat{\tau}_{\langle q'\rangle}$ *DOES NOT match the following string patterns:*

$$(h(e|\hat{e})d\mathbf{X}|ing\mathbf{X}|lyk\mathbf{X}|t\hat{e}t\mathbf{X}|um).$$

This list of triplets $\mathrm{CpdDet}(\Lambda_{\hat{\rho}})$ *contains the heuristic decompositions of all the identified binary compounds.*

Algorithm 5.35 (Approximate clustering of Dutch words with heuristic detection of compounds)**.** *The procedure runs essentially the same way as Algorithm 5.35, except that Dutch rules replace Danish rules.*

Example 5.35.1. Testing the algorithm above against the combined inputs from Examples 5.33.1 and 5.33.2, we obtain the following result:

{*bak*, *bakke*, *bakken*, *bakkend*, *bakt*, *bakte*, *bakten*, *gebakken*}, {*ban*, *bande*, *banden*, *banne*, *bannen*, *bannend*, *bant*, *gebannen*}, {*bederf*, *bederft*, *bederve*, *bederven*, *bedervend*, *bedierf*, *bedierft*, *bedorve*, *bedorven*}, {*bedrieg*, *bedriege*, *bedriegen*, *bedriegend*, *bedriegt*, *bedroge*, *bedrogen*, *bedroog*, *bedroogt*}, {*beet*, *bete*, *beten*, *bijt*, *bijte*, *bijten*, *bijtend*, *gebeten*}, {*begin*, *beginne*, *beginnen*, *beginnend*, *begint*, *begon*, *begonne*, *begonnen*, *begont*}, {*best*, *beste*, *beter*, *betere*, *beters*, *goed*, *goede*, *goeden*, *goeder*, *goeds*}, {*bied*, *biede*, *bieden*, *biedend*, *biedt*, *bode*, *boden*, *bood*, *boodt*, *geboden*}, {*bind*, *binde*, *binden*, *bindend*, *bindt*, *bond*, *bonde*, *bonden*, *bondt*, *gebonden*}, {*blaas*, *blaast*, *blaze*, *blazen*, *blazend*, *blies*, *bliest*, *blieze*, *bliezen*, *geblazen*}, {*bleek*, *bleekt*, *bleke*, *bleken*, *blijk*, *blijke*, *blijken*, *blijkend*, *blijkt*, *gebleken*}, {*boge*, *bogen*, *boog*, *boogt*, *buig*, *buige*, *buigen*, *buigend*, *buigt*, *gebogen*}, {*braakt*, *brak*, *brake*, *braken*, *breek*, *breekt*, *breke*, *breken*, *brekend*, *gebroken*}, {*bracht*, *brachte*, *brachten*, *breng*, *brenge*, *brengen*, *brengend*, *brengt*, *gebracht*}, {*daad*, *daden*}, {*dacht*, *dachte*, *dachten*, *denk*, *denke*, *denken*, *denkend*, *denkt*, *gedacht*}, {*dooier*, *dooiers*, *dooiertje*, *eidooier*, *eidooiers*, *eidooiertje*, *eierdooier*, *eierdooiers*, *eierdooiertje*}, {*draag*, *draagt*, *drage*, *dragen*, *dragend*, *droeg*, *droege*, *droegen*, *droegt*, *gedragen*}, {*droop*, *droopt*, *drope*, *dropen*, *druip*, *druipe*, *druipen*, *druipend*, *druipt*, *gedropen*}, {*ei*, *eidooier*, *eidooiers*, *eidooiertje*, *eierdooier*, *eierdooiers*, *eierdooiertje*, *eieren*, *eitje*, *eitjes*, *eiwit*, *eiwitten*, *eiwittje*}, {*eiwit*, *eiwitten*, *eiwittje*, *wit*, *wits*, *witst*, *witste*, *witte*, *witter*, *wittere*, *witters*}, {*gegolden*, *geld*, *gelde*, *gelden*, *geldend*, *geldt*, *gold*, *golde*, *golden*, *goldt*}, {*geheven*, *hef*, *heffe*, *heffen*, *heffend*, *heft*, *hief*, *hieft*, *hieve*, *hieven*}, {*geholpen*, *help*, *helpe*, *helpen*, *helpend*, *helpt*, *hielp*, *hielpe*, *hielpen*, *hielpt*}, {*gekocht*, *kocht*, *kochte*, *kochten*, *koop*, *koopt*, *kope*, *kopen*, *kopend*}, {*geleerd*, *leer*, *leerde*, *leerden*, *leert*, *lere*, *leren*, *lerend*}, {*gelegen*, *laagt*, *lag*, *lage*, *lagen*, *lig*, *ligge*, *liggen*, *liggend*, *ligt*}, {*gelost*, *los*, *losse*, *lossen*, *lossend*, *lost*, *loste*, *losten*}, {*genomen*, *naamt*, *nam*, *name*, *namen*, *neem*, *neemt*, *neme*, *nemen*, *nemend*}, {*geraasd*, *raas*, *raasde*, *raasden*, *raast*, *raze*, *razen*, *razend*}, {*geschapen*, *schep*, *scheppe*, *scheppen*, *scheppend*, *schept*, *schiep*, *schiepe*, *schiepen*, *schiept*}, {*gescholden*, *scheld*,

schelde, *schelden*, *scheldend*, *scheldt*, *schold*, *scholde*, *scholden*, *scholdt*}, {*gevraagd*, *vraag*, *vraagd*, *vraagde*, *vraagden*, *vraagt*, *vrage*, *vragen*, *vragend*, *vroeg*, *vroege*, *vroegen*, *vroegt*}, {*gewerkt*, *werk*, *werke*, *werken*, *werkend*, *werkt*, *werkte*, *werkten*}, {*gewogen*, *weeg*, *weegt*, *wege*, *wegen*, *wegend*, *woge*, *wogen*, *woog*, *woogt*}, {*gezeten*, *zat*, *zate*, *zaten*, *zit*, *zitte*, *zitten*, *zittend*}, {*gezocht*, *zocht*, *zochte*, *zochten*, *zoek*, *zoeke*, *zoeken*, *zoekend*, *zoekt*}, {*klein*, *kleine*, *kleinen*, *kleiner*, *kleinere*, *kleiners*, *kleins*, *kleinst*, *kleinste*}, {*maan*, *maantje*, *man*, *manen*, *manne*, *mannen*, *mans*}, {*vrouw*, *vrouwe*, *vrouwen*}.

Here, the compounds *ei*(∅|*er*)*dooier* "egg yolk" and *eiwit* "egg white" are correctly dissolved into their respective constituting components.

Example 5.35.2. In Fig. S6, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).

Note that *de* "the" is a Dutch stop word, so the English *de* cannot be translated exactly by our algorithm.

The co-occurring words in *Lady Catherine (de Bourgh)* and *Sir William (Lucas)* also make it difficult to resolve them exactly in our numerical experiments. These pairs of mismatches are indicated by amber cross-hairs in Fig. S6b.

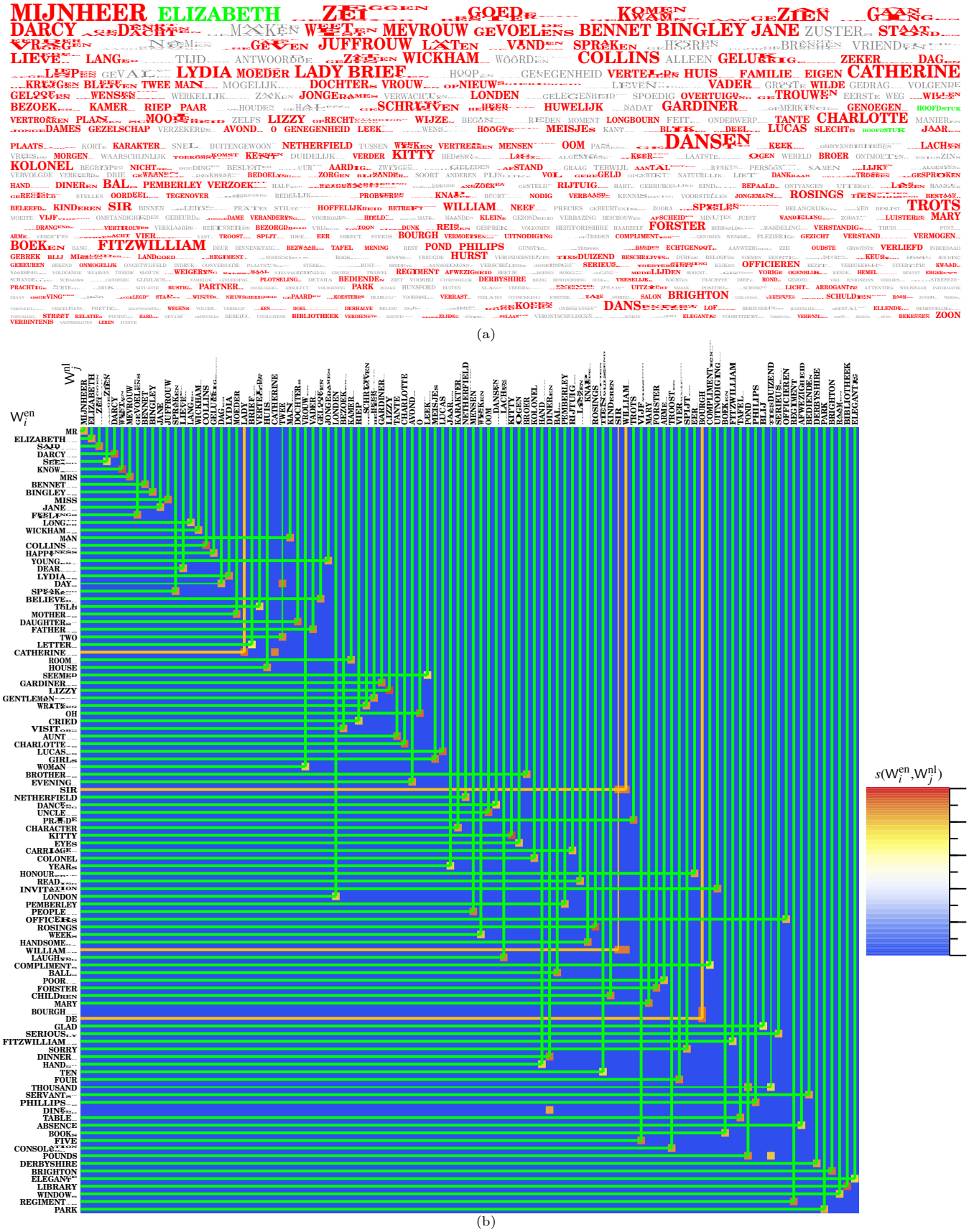


Fig. S6. Text mining in Dutch. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Dutch version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{nl})$ between selected topics in English and Dutch versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *amber*) cross-hair indicates an exact (resp. a close but non-exact) match.

6 Approximate word clustering in selected Romance languages

In this section, we present the approximate clustering algorithms for two representative modern Romance languages with a large number of speakers: Spanish and French. Their ancestral language, Latin, is also treated in this section.

Due to heavy loads of Latinate loanwords in English (borrowed either directly from Latin, or through Norman French), the word formation of English shares a lot of features with modern Romance languages. However, some critical differences must be noted:

- Unlike English, apostrophes and hyphens are used extensively in French to mark word boundaries.
- Unlike English, verbs in all these three languages conjugate in all persons for all tenses and aspects, thus making the verb morphology highly complicated.
- Unlike English, adjectives in all these three languages inflect in both gender and number.
- The classical language Latin additionally have declensions of nouns and adjectives in six cases.

The descriptions above (except the case declension of Latin) also extend to modern Italian, together with some Romance languages spoken in the Iberian Peninsula other than Spanish (namely, Catalan, Galician, Portuguese and Valencian), with minor modifications. Romanian, a Romance language spoken in Eastern Europe, has more conservative morphological structures (such as three genders and five cases), which are closer to the complexity of Latin (which has three genders and six cases) than all the other modern Romance languages.

6.1 Modified Porter stemming algorithm for Spanish

Before defining Spanish stop words, we need to define Spanish vowel extensions, and construct an algorithm to stress the penultimate vowel in a Spanish word, by heuristics.

Definition 6.1 (Spanish Vowel Extensions)**.** Hereafter in §6.1, the symbol $\mathbf{V}^*$ stands for any member from the list {*a*, *e*, *i*, *î*, *o*, *u*}, the so-called Spanish vowel extensions. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}_m^*$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Spanish vowel extensions.

Dual to the notations above, the symbol $\mathbf{C}^*$ stands for any character that does not belong to the list {*a*, *e*, *i*, *î*, *o*, *u*}, and $\mathbf{C}_{m_0}^*$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list. □

Algorithm 6.2 (Penultimate Stress)**.** *We define* StrPenult($\hat{\sigma}$) *through the following operations on* $\hat{\sigma}$*:*

(1) *Remove diacritic marks from* $\hat{\sigma}$*, except that the Spanish letter ñ must be kept intact.*

(2) *Do* $\sim(\mathbf{V}^*)^{\mathbf{X}\in}(\mathbf{C}_m^*\mathbf{V}_m^*\mathbf{C}_{m_0}^*) \to \mathbf{V}'\mathbf{X}$*, where one derives* $\mathbf{V}'$ *from* $\mathbf{V}^*$ *by adding stress marks to* (*a*|*e*|*i*|*o*|*u*)*. The result so far is called* $\hat{\sigma}'$*.*

(3) *If* $\Omega(\hat{\sigma}') = (d|r)$ *or* $\hat{\sigma}' =$ *está, then perform the operation in Step* (1) *on* $\hat{\sigma}'$ *and output* StrPenult($\hat{\sigma}$)*. Otherwise, check whether* $\hat{\sigma}' = \mathbf{X}$*ea. If so,* StrPenult($\hat{\sigma}$) *results from doing* $\sim$*ea* $\to$ *éa on* $\hat{\sigma}'$*; if not,* StrPenult($\hat{\sigma}$) = $\hat{\sigma}'$*.*

Definition 6.3 (Spanish stop words)**.** If a word belongs to the following list[70]:

> *a*, *abajo*, *acá*, *acaso*, *adentro*, *afuera*, *ahí*, *ahora*, *al*, *algo*, *alguien*, *algun*, *algún*, *alguna*, *algunas*, *alguno*, *algunos*, *allá*, *allí*, *alrededor*, *ambas*, *ambos*, *ante*, *antes*, *aquel*, *aquél*, *aquella*, *aquélla*, *aquellas*, *aquéllas*, *aquello*, *aquellos*, *aquéllos*, *aquí*, *arriba*, *así*, *atrás*, *aun*, *aún*, *aunque*, *bajo*, *bien*, *cada*, *casi*, *cerca*, *como*, *cómo*, *con*, *conmigo*, *connosco*, *consigo*, *contigo*, *contra*, *convusco*, *cual*, *cuál*, *cuales*, *cualesquier*, *cualesquiera*, *cualquier*, *cualquiera*, *cuan*, *cuán*, *cuando*, *cuándo*, *cuanta*, *cuánta*, *cuantas*, *cuántas*, *cuanto*, *cuánto*, *cuantos*, *cuántos*, *cuya*, *cúya*, *cuyas*, *cúyas*, *cuyo*, *cúyo*, *cuyos*, *cúyos*, *de*, *debajo*, *del*, *delante*, *dentro*, *desde*, *después*, *detrás*, *donde*, *dónde*, *durante*, *durea*, *e*, *el*, *él*, *ella*, *ellas*, *ello*, *ellos*, *empero*, *en*, *encima*, *enfrente*, *entonces*, *entre*, *esa*, *ésa*, *esas*, *ésas*, *ese*, *ése*, *eses*, *éses*, *eso*, *esos*, *esta*, *ésta*, *estas*, *éstas*, *este*, *éste*, *éstes*, *esto*, *estos*, *frente*, *hacia*, *hasta*, *jamás*, *junto*, *la*, *las*, *le*, *lejos*, *les*, *lo*, *los*, *luego*, *mas*, *más*, *me*, *mediante*, *menos*, *mi*, *mí*, *mía*, *mías*, *mío*, *míos*, *mis*, *misma*, *mismas*, *mismísima*, *mismísimas*, *mismísimo*, *mismísimos*, *mismo*, *mismos*, *mucha*, *muchas*, *mucho*, *muchos*, *muy*, *nada*, *nadie*, *ni*, *ningún*, *ninguna*, *ningunas*, *ninguno*, *ningunos*, *no*, *nos*, *nosotras*, *nosotros*, *nuestra*, *nuestras*, *nuestro*, *nuestros*, *nunca*, *o*, *os*, *otra*, *otras*, *otro*, *otros*, *para*, *pero*, *poco*, *por*, *porque*, *pronto*, *pues*, *que*, *qué*, *quien*, *quién*, *quienes*, *quizá*, *según*, *si*, *sí*, *siempre*, *sin*, *sino*, *siquiera*, *so*, *sobre*, *su*, *sus*, *suya*, *suyas*, *suyo*, *suyos*,

[70] Our list of Spanish stop words is based on `snowball.tartarus.org/algorithms/spanish/stop.txt`, with extensive additions.

también, *tampoco*, *tan*, *tanta*, *tantas*, *tanto*, *tantos*, *te*, *ti*, *toda*, *todas*, *todavía*, *todo*, *todos*, *tras*, *tu*, *tú*, *tus*, *tuya*, *tuyas*, *tuyo*, *tuyos*, *u*, *un*, *una*, *uno*, *unos*, *usted*, *ustedes*, *vosotras*, *vosotros*, *vuestra*, *vuestras*, *vuestro*, *vuestros*, *y*, *ya*, *yo*,

then we consider it a Spanish stop non-verb (notation: **StopNonVerbSpanish**). If a word belongs to the following list:

era, *erais*, *éramos*, *eran*, *eras*, *eres*, *es*, *está*, *estaba*, *estabais*, *estábamos*, *estaban*, *estabas*, *estad*, *estada*, *estadas*, *estado*, *estados*, *estáis*, *estamos*, *están*, *estando*, *estar*, *estará*, *estarán*, *estarás*, *estaré*, *estaréis*, *estaremos*, *estaría*, *estaríais*, *estaríamos*, *estarían*, *estarías*, *estás*, *esté*, *estéis*, *estemos*, *estén*, *estés*, *estoy*, *estuve*, *estuviera*, *estuvierais*, *estuviéramos*, *estuvieran*, *estuvieras*, *estuviere*, *estuviereis*, *estuviéremos*, *estuvieren*, *estuvieres*, *estuvieron*, *estuviese*, *estuvieseis*, *estuviésemos*, *estuviesen*, *estuvieses*, *estuvimos*, *estuviste*, *estuvisteis*, *estuvo*, *fue*, *fuera*, *fuerais*, *fuéramos*, *fueran*, *fueras*, *fuere*, *fuereis*, *fuéremos*, *fueren*, *fueres*, *fueron*, *fuese*, *fueseis*, *fuésemos*, *fuesen*, *fueses*, *fui*, *fuimos*, *fuiste*, *fuisteis*, *ha*, *habe*, *habed*, *habéis*, *haber*, *había*, *habíais*, *habíamos*, *habían*, *habías*, *habida*, *habidas*, *habido*, *habidos*, *habiendo*, *habrá*, *habrán*, *habrás*, *habré*, *habréis*, *habremos*, *habría*, *habríais*, *habríamos*, *habrían*, *habrías*, *hace*, *hacé*, *haced*, *hacéis*, *hacemos*, *hacen*, *hacer*, *hacestú*, *hacésvos*, *hacía*, *hacíais*, *hacíamos*, *hacían*, *hacías*, *haciendo*, *haga*, *hagáis*, *hagamos*, *hagan*, *hagas*, *hago*, *han*, *hará*, *harán*, *harás*, *haré*, *haréis*, *haremos*, *haría*, *haríais*, *haríamos*, *harían*, *harías*, *has*, *hay*, *haya*, *hayáis*, *hayamos*, *hayan*, *hayas*, *haz*, *he*, *hé*, *hecha*, *hechas*, *hecho*, *hechos*, *hemos*, *hice*, *hiciera*, *hicierais*, *hiciéramos*, *hicieran*, *hicieras*, *hiciere*, *hiciereis*, *hiciéremos*, *hicieren*, *hicieres*, *hicieron*, *hiciese*, *hicieseis*, *hiciésemos*, *hiciesen*, *hicieses*, *hicimos*, *hiciste*, *hicisteis*, *hizo*, *hube*, *hubiera*, *hubierais*, *hubiéramos*, *hubieran*, *hubieras*, *hubiere*, *hubiereis*, *hubiéremos*, *hubieren*, *hubieres*, *hubieron*, *hubiese*, *hubieseis*, *hubiésemos*, *hubiesen*, *hubieses*, *hubimos*, *hubiste*, *hubisteis*, *hubo*, *podáis*, *podamos*, *poded*, *podéis*, *podemos*, *poder*, *podía*, *podíais*, *podíamos*, *podían*, *podías*, *podida*, *podidas*, *podido*, *podidos*, *podrá*, *podrán*, *podrás*, *podré*, *podréis*, *podremos*, *podría*, *podríais*, *podríamos*, *podrían*, *podrías*, *pude*, *pudiendo*, *pudiera*, *pudierais*, *pudiéramos*, *pudieran*, *pudieras*, *pudiere*, *pudiereis*, *pudiéremos*, *pudieren*, *pudieres*, *pudieron*, *pudiese*, *pudieseis*, *pudiésemos*, *pudiesen*, *pudieses*, *pudimos*, *pudiste*, *pudisteis*, *pudo*, *pueda*, *puedan*, *puedas*, *puede*, *pueden*, *puedes*, *puedo*, *se*, *sé*, *sea*, *seáis*, *seamos*, *sean*, *seas*, *sed*, *ser*, *será*, *serán*, *serás*, *seré*, *seréis*, *seremos*, *sería*, *seríais*, *seríamos*, *serían*, *serías*, *sida*, *sidas*, *sido*, *sidos*, *siendo*, *sois*, *somos*, *son*, *sos*, *soy*, *ten*, *tendrá*, *tendrán*, *tendrás*, *tendré*, *tendréis*, *tendremos*, *tendría*, *tendríais*, *tendríamos*, *tendrían*, *tendrías*, *tené*, *tened*, *tenéis*, *tenemos*, *tener*, *tenés*, *tenga*, *tengáis*, *tengamos*, *tengan*, *tengas*, *tengo*, *tenía*, *teníais*, *teníamos*, *tenían*, *tenías*, *tenida*, *tenidas*, *tenido*, *tenidos*, *teniendo*, *tiene*, *tienen*, *tienes*, *tuve*, *tuviera*, *tuvierais*, *tuviéramos*, *tuvieran*, *tuvieras*, *tuviere*, *tuviereis*, *tuviéremos*, *tuvieren*, *tuvieres*, *tuvieron*, *tuviese*, *tuvieseis*, *tuviésemos*, *tuviesen*, *tuvieses*, *tuvimos*, *tuviste*, *tuvisteis*, *tuvo*,

then we consider it a Spanish stop verb (notation: **StopVerbSpanish**). A Spanish stop word matches the following string pattern:

(**StopNonVerbSpanish**|**StopVerbSpanish**|StrPenult(**StopVerbSpanish**)**SpanishVerbCombo**),

where the string pattern **SpanishVerbCombo** = (*la*|*las*|*le*|*les*|*lo*|*los*|*me*|*nos*|*os*|*se*|*te*) points to pronouns that may attach to the end of Spanish verbs to form "combined forms". All the Spanish stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

The following Spanish verbs have highly irregular conjugations:

da, *daba*, *dabais*, *dábamos*, *daban*, *dabas*, *dada*, *dadas*, *dado*, *dados*, *dais*, *damos*, *dan*, *dando*, *dar*, *dará*, *darán*, *darás*, *daré*, *daréis*, *daremos*, *daría*, *daríais*, *daríamos*, *darían*, *darías*, *das*, *dé*, *deis*, *demos*, *den*, *des*, *di*, *diera*, *dierais*, *diéramos*, *dieran*, *dieras*, *diere*, *diereis*, *diéremos*, *dieren*, *dieres*, *dieron*, *diese*, *dieseis*, *diésemos*, *diesen*, *dieses*, *dimos*, *dio*, *diste*, *disteis*, *doy* — "give";

fue, *fuera*, *fuerais*, *fuéramos*, *fueran*, *fueras*, *fuere*, *fuereis*, *fuéremos*, *fueren*, *fueres*, *fueron*, *fuese*, *fueseis*, *fuésemos*, *fuesen*, *fueses*, *fui*, *fuimos*, *fuiste*, *fuisteis*, *iba*, *ibais*, *íbamos*, *iban*, *ibas*, *id*, *ida*, *idas*, *ido*, *idos*, *ir*, *irá*, *irán*, *irás*, *iré*, *iréis*, *iremos*, *iría*, *iríais*, *iríamos*, *irían*, *irías*, *va*, *vais*, *vamos*, *van*, *vas*, *vaya*, *vayáis*, *vayamos*, *vayan*, *vayas*, *ve*, *voy*, *yendo* — "go".

We are not going to treat them as stop words, so as to be consistent with their English counterparts. Nevertheless, we will identify these string patterns as **giveSpanish** and **goSpanish**, to facilitate the clustering of content words.

6.1.1 Effective spelling and essential root

Algorithm 6.4 (Spanish effective spelling)**.** *For a Spanish word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in six sequential steps:*

(1) *Replace*

af~	*ag~*	*amab(∅|i)l~*	*ami(g|stad|stos)~*	*año~*	*at~*	*biblio~*	*caus~*	*charlott*	*cinc~*	*com(ú|u)n~*
αf	*αg*	*φρeṽδ*	*φρeṽδ*	*ηαρο*	*αt*	*βiβλiω*	*καuσ*	*šarlοττ*	*qi5*	*κωmuṽ*

coment~	*comprom*	*contr~*	*dis~*	*feli(c|z)~*	*gard~*	*inter~*	*list~*	*long~*	*mer(a|o)(∅|s)~*	*mili(c|t)~*	*mir~*
κομεητ	*κπρομ*	*κτρ*	*δ*	*φελiζ*	*γarδ*	*ι*	*λiστ*	*λοṽγ*	*πuρ*	*αarμ*	*μiρ*

pare(∅|z)c	*posib~*	*posic~*	*positiv~*	*primaver*	*primit*	*puerta*	*vener~*	*vino~*	*carta(∅|s)*	*jueves*	*mary*
πeαρ	*ποσiβ*	*πωσiζ*	*ρωsiτiβ*	*σπριṽγ*	*πριμiτ*	*pγortτ*	*βeṽeρ*	*ωiṽε*	*λεττρ*	*j4ue4*	*μarη*

(2) *Replace*

(á|a)nim~	*(di|díga)***SpanishVerbCombo**~	*(dia|día)~*	*(ie|ié)*	*(ndr|ng)*	*(zc|zg)*	*á*	*abuelo~*	*ac~*
αniμ	*decir*	*δia*	*î*	*n*	*z*	*a*	*abuelω*	*αc*

dece~	*dif~*	*dij~*	*diremos*	*dire~*	*edad~*	*edu~*	*en~*	*es~*	*ex~*	*hermano*	*hermos~*	*hijo~*	*in~*	*lady~*
δece	*διf*	*dich*	*decir*	*δire*	*εdaδ*	*εδu*	*εṽ*	*εs*	*εξ*	*hermanω*	*hermωσ*	*hijω*	*ιṽ*	*λady*

ldr	*marid~*	*mejor(∅|es)~*	*oc~*	*ón*	*par(e|ie)nt~*	*qu*	*sg*	*sir~*	*su*$^{\mathbf{X}\in}$*((f|g|p|r)χ̂)~*
l	*μarid*	*bueno*	*ωc*	*one*	*παριent*	*k*	*s*	*σiρ*	*sω***X**

tío~	*ye~*	StrPenult(**giveSpanish**)**SpanishVerbCombo**	StrPenult(**goSpanish**)**SpanishVerbCombo**
tíω	*e*	*σγibσ*	*σgoσ*

(miss|señorita|srta)(∅|s)	*(mr|señor)(∅|es)*	*(mrs|señora|sra)(∅|s)*	**giveSpanish**	**goSpanish**
mdseñorita	*mrseñor*	*fseñora*	*σγibσ*	*σgoσ*

mam(a|á|i)(∅|s)	*pap(a|á|acito|aíto|i)(∅|s)*	*~χ̂(í|i)s(mo|ta|tica|tico)(∅|s)*	*~dos*	*~tri(ces|z)*
madre	*padre*	*χ̂e*	*das*	*tor*

(3) *Replace*

(∅|h)(u|ü)e	*c(c|t)*	*co*$^{\mathbf{X}\in}$*(g|j)*	*én*	*í*
o	*z*	*raco***X**	*ena*	*i*

(4) *Do é → e, ñ → n, ó → o, (ú\|ü) → u.*

(5) *Do* $c^{\mathbf{V}\in}$*(a\|o\|u) → k***V**, $c^{\mathbf{V}\in}$*(e\|i\|î) → z***V**, $g^{\mathbf{V}\in}$*(e\|i\|î) → j***V**, *gu → g.*

(6) *Do ven → vis, ~(d\|mo\|mos\|n\|r)***SpanishVerbCombo** *→ ∅, ~idad(∅\|es) → ∅.*

Definition 6.5 (Spanish protected range)**.** Let $\hat{\sigma}$ be a text string derived from a Spanish word, its protected range ProtRg($\hat{\sigma}$) is an integer determined as follows:

- Try to find the string pattern $(\varnothing|al|des|es|ko|pre|pro|ra|re|tran|tras)\mathbf{C}^*_{m_0}\mathbf{V}^*(\varnothing|(b|d|g|m|n|r|s|t|v|z)_m)\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines ProtRg($\hat{\sigma}$); otherwise, set ProtRg($\hat{\sigma}$) = 0. □

Algorithm 6.6 (Spanish essential root)**.** *Let $\hat{\sigma}$ be the effective spelling of a Spanish word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do kom***V**$^*\sim \to kkom$**V*** *on $\hat{\sigma}$ and call the result $\hat{\sigma}_*$.*

(2) *Break down $\hat{\sigma}_* = \hat{\sigma}_1\hat{\sigma}_2$ into the concatenation of two strings $\hat{\sigma}_1 = \hat{\sigma}_*^{[\mathrm{ProtRg}(\hat{\sigma}_*)]}$ (see the notation in Definition 3.1) and $\hat{\sigma}_2$, where the length of the first string $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma}_*)$ is equal to the protected range of $\hat{\sigma}_*$.*

(3) *On $\hat{\sigma}_2$, perform the following substitutions in a sequel:*

(3.1) *Do* $\sim(d|t)(a|o)(\varnothing|s)(\varnothing|$**SpanishVerbCombo**$) \to \varnothing$, $\sim d(\varnothing|es) \to \varnothing$.

(3.2) *Do* $(ba|(b(\varnothing|i)(\varnothing|l))|nt|(r$**V**$^*|\sim r)|st|\sim n)(\varnothing|$**SpanishVerbCombo**$) \to \varnothing$.

(3.3) *Do* $(dor|g|ion|ka|ko|log|m|s|tiv|uv|y|z) \to \varnothing$.[71]

(3.4) *Do* $\sim\mathbf{V}^*_m \to \varnothing$.

[71] We note that longer matches take priority over shorter matches: if *log* is found, then delete these three letters altogether, instead of just a single letter *g*.

The result after these four steps of operations is called $\hat{\sigma}'_2$.

(4) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(5) *Perform the following replacements on* $\hat{\sigma}_1\hat{\sigma}'_2$*:*

$\underline{\mathbf{X}^{\in}(\varnothing\|kon\|re)di}$	$\underline{durm}$	$\underline{o}$	~d(ez|ich|ig|ij|iz)	~duj	~jog	~k(e|u)p	~kerr	~kis	~p(ost|us)
Xdir	dorm	oxx	dir	duz	jug	kab	ker	kîr	pon
~pud	~traj		~tuv	~v(e|ed|em|er|es|im|înd|îr|îs|ist)					~yaz
pod	tra		ten	vis					yag

6.1.2 Admissible mutation and approximate clustering

Algorithm 6.7 (Spanish vowel blotting)**.** *Set* $\mathbf{V}^*_m$ *and* $\mathbf{C}^*_{m_0}$ *as in Definition 6.1. For a token string* $\hat{\sigma}$*, its blotted form* $\mathrm{BlotV}_1(\hat{\sigma})$ *is constructed as follows:*

- *If the string pattern* $(\varnothing|a|des|e|es|i|î|ko|o|pre|pro|ra|re|tran|tras|u)\mathbf{C}^*_{m_0}\mathbf{V}^*_m\sim$ *can be found in the string* $\hat{\sigma}$*, then the last position occupied by such a pattern is replaced by the letter "a".*
- *Otherwise, leave the string* $\hat{\sigma}$ *intact.*

Similar to what we did in §5 for three representative Germanic languages, we will construct a bivariate Boolean-valued function $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ on a "simple heredity test function" in Algorithm 6.8, and a set of "admissible suffix mismatch and vowel alternation" rules in Algorithm 6.9.

Algorithm 6.8 (Simple heredity test)**.** *The Boolean-valued function* $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if the lowercase form of* $\hat{\alpha}$ *contains at least one instance of* $\mathbf{V}^*$ *(Definition 6.1)* **AND** *at least one of the following four conditions holds:*[72]

(i) *After doing* $\sim s\rightarrow\varnothing$ *on* $\hat{\beta}$*, we obtain* $\hat{\alpha}$*;*

(ii) *After doing* $\sim a(\varnothing|s)\rightarrow o$ *on* $\hat{\alpha}$*, we obtain* $\hat{\beta}$*;*

(iii) $\hat{\beta}=\hat{\alpha}k$*;*

(iv) $\ell(\hat{\beta})>\ell(\hat{\alpha})\geq\frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha}=\hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha}')+1\}}=\mathbf{V}^*$*. (See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

In what follows, we define $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}),\mathrm{RootNW}(\hat{\alpha},\hat{\beta}),\mathrm{NW}^*(\hat{\alpha},\hat{\beta})$ and $\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\mathrm{SW}^*(\hat{\alpha},\hat{\beta})$ as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5.

Algorithm 6.9 (Admissible suffix mismatch and vowel alternation)**.** *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function*

$$\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}),\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}),\mathrm{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if*

$$\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})=[(\varnothing|\mathbf{V}^*\mathbf{X}(d|t)|(r|s)\mathbf{X}|tiv\mathbf{X}),(\varnothing|\mathbf{V}^*\mathbf{X}(d|t)|(r|s)\mathbf{X}|tiv\mathbf{X})]$$

AND *at least one of the following two conditions holds:*

(i) $\mathrm{NW}^*(\hat{\alpha},\hat{\beta})=\varnothing$ **AND** $\mathrm{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* $\mathbf{V}^*$*;*

(ii) $\mathrm{NW}^*(\hat{\alpha},\hat{\beta})=([e,\,i]||[e,\,î]||[i,\,î]||[î,\,i])$.

Similarly, one can evaluate another Boolean-valued function

$$\mathrm{AdmSM}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}),\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}),\mathrm{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW*.*

Algorithm 6.10 (Heredity test function)**.** *The structure of the Spanish heredity test function* $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ *is identical to the German version (Algorithm 8.1.2), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW***, RootSW*,* SuffixSW*,* SW** must follow the Spanish rules stated above.*

[72] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 6.11 (Approximate clustering of Spanish words)**.** *The algorithm is essentially the same as Algorithm 5.10, except that Spanish rules (instead of Danish rules) apply to all the tags (effective spelling, essential root, vowel blotting etc.).*

Example 6.11.1. We first select some "simple" Spanish word families with (more or less) regular inflections. For the selected verbs, we have not only included conjugated forms (in different moods, persons and tenses) but also derived forms (such as adjectives and nouns associated with a verb). An "approximate translation" in English, enclosed in quotation marks, is appended to the end of each family. The readers should be reminded that such an "approximate translation" only matches certain members with a word family.

> *concluí*, *concluía*, *concluíais*, *concluíamos*, *concluían*, *concluías*, *concluid*, *concluida*, *concluidas*, *concluido*, *concluidos*, *concluimos*, *concluir*, *concluirá*, *concluirán*, *concluirás*, *concluiré*, *concluiréis*, *concluiremos*, *concluiría*, *concluiríais*, *concluiríamos*, *concluirían*, *concluirías*, *concluís*, *concluiste*, *concluisteis*, *conclusión*, *conclusivo*, *concluso*, *concluya*, *concluyáis*, *concluyamos*, *concluyan*, *concluyas*, *concluye*, *concluyen*, *concluyendo*, *concluyente*, *concluyera*, *concluyerais*, *concluyéramos*, *concluyeran*, *concluyeras*, *concluyere*, *concluyereis*, *concluyéremos*, *concluyeren*, *concluyeres*, *concluyeron*, *concluyes*, *concluyese*, *concluyeseis*, *concluyésemos*, *concluyesen*, *concluyeses*, *concluyo*, *concluyó* — "conclude";
>
> *continua*, *continúa*, *continuaba*, *continuabais*, *continuábamos*, *continuaban*, *continuabas*, *continuación*, *continuad*, *continuada*, *continuadamente*, *continuadas*, *continuado*, *continuador*, *continuadora*, *continuados*, *continuáis*, *continuamente*, *continuamiento*, *continuamos*, *continúan*, *continuando*, *continuar*, *continuara*, *continuará*, *continuarais*, *continuáramos*, *continuaran*, *continuarán*, *continuaras*, *continuarás*, *continuare*, *continuaré*, *continuareis*, *continuaréis*, *continuaremos*, *continuáremos*, *continuaren*, *continuares*, *continuaría*, *continuaríais*, *continuaríamos*, *continuarían*, *continuarías*, *continuaron*, *continúas*, *continuase*, *continuaseis*, *continuásemos*, *continuasen*, *continuases*, *continuaste*, *continuasteis*, *continuativa*, *continuativo*, *continué*, *continué*, *continúe*, *continuéis*, *continuemos*, *continúen*, *continúes*, *continuidad*, *continuo*, *continuó*, *continúo* — "continue";
>
> *rápida*, *rápidamente*, *rápidas*, *rápido*, *rápidos* — "fast";
>
> *rapideces*, *rapidez* — "speed".

Applying Algorithm 6.11 to the list of words above, we arrive at the following result:

> {*concluí*, *concluía*, *concluíais*, *concluíamos*, *concluían*, *concluías*, *concluid*, *concluida*, *concluidas*, *concluido*, *concluidos*, *concluimos*, *concluir*, *concluirá*, *concluirán*, *concluirás*, *concluiré*, *concluiréis*, *concluiremos*, *concluiría*, *concluiríais*, *concluiríamos*, *concluirían*, *concluirías*, *concluís*, *concluiste*, *concluisteis*, *conclusión*, *conclusivo*, *concluso*, *concluya*, *concluyáis*, *concluyamos*, *concluyan*, *concluyas*, *concluye*, *concluyen*, *concluyendo*, *concluyente*, *concluyera*, *concluyerais*, *concluyéramos*, *concluyeran*, *concluyeras*, *concluyere*, *concluyereis*, *concluyéremos*, *concluyeren*, *concluyeres*, *concluyeron*, *concluyes*, *concluyese*, *concluyeseis*, *concluyésemos*, *concluyesen*, *concluyeses*, *concluyo*, *concluyó*},
>
> {*continua*, *continúa*, *continuaba*, *continuabais*, *continuábamos*, *continuaban*, *continuabas*, *continuación*, *continuad*, *continuada*, *continuadamente*, *continuadas*, *continuado*, *continuador*, *continuadora*, *continuados*, *continuáis*, *continuamente*, *continuamiento*, *continuamos*, *continúan*, *continuando*, *continuar*, *continuara*, *continuará*, *continuarais*, *continuáramos*, *continuaran*, *continuarán*, *continuaras*, *continuarás*, *continuare*, *continuaré*, *continuareis*, *continuaréis*, *continuaremos*, *continuáremos*, *continuaren*, *continuares*, *continuaría*, *continuaríais*, *continuaríamos*, *continuarían*, *continuarías*, *continuaron*, *continúas*, *continuase*, *continuaseis*, *continuásemos*, *continuasen*, *continuases*, *continuaste*, *continuasteis*, *continuativa*, *continuativo*, *continué*, *continúe*, *continuéis*, *continuemos*, *continúen*, *continúes*, *continuidad*, *continuo*, *continuó*, *continúo*},
>
> {*rápida*, *rápidamente*, *rápidas*, *rapideces*, *rapidez*, *rápido*, *rápidos*}.

Example 6.11.2. The representative verbs from the three regular conjugations of Spanish are given below.

> *coma*, *comáis*, *comamos*, *coman*, *comas*, *comás*, *come*, *comé*, *comed*, *coméis*, *comemos*, *comen*, *comer*, *comerá*, *comerán*, *comerás*, *comeré*, *comeréis*, *comeremos*, *comería*, *comeríais*, *comeríamos*, *comerían*, *comerías*, *comes*, *comés*, *comí*, *comía*, *comíais*, *comíamos*, *comían*, *comías*, *comida*, *comidas*, *comido*, *comidos*, *comiendo*, *comiera*, *comierais*, *comiéramos*, *comieran*, *comieras*, *comiere*, *comiereis*, *comiéremos*, *comieren*, *comieres*, *comieron*, *comiese*, *comieseis*, *comiésemos*, *comiesen*, *comieses*, *comimos*, *comió*, *comiste*, *comisteis*, *como* — "eat";
>
> *habla*, *hablá*, *hablaba*, *hablabais*, *hablábamos*, *hablaban*, *hablabas*, *hablad*, *hablada*, *habladas*, *hablado*, *hablados*, *habláis*, *hablamos*, *hablan*, *hablando*, *hablar*, *hablara*, *hablará*, *hablarais*, *hablaríamos*, *hablaran*, *hablarán*,

hablaras, *hablarás*, *hablare*, *hablaré*, *hablareis*, *hablaréis*, *hablaremos*, *hablaremos*, *hablaren*, *hablares*, *hablaría*, *hablaríais*, *hablaríamos*, *hablarían*, *hablarías*, *hablaron*, *hablas*, *hablás*, *hablase*, *hablaseis*, *hablásemos*, *hablasen*, *hablases*, *hablaste*, *hablasteis*, *hable*, *hablé*, *habléis*, *hablemos*, *hablen*, *hables*, *hablés*, *hablo*, *habló* — "speak";

viva, *viváis*, *vivamos*, *vivan*, *vivas*, *vivás*, *vive*, *viven*, *vives*, *viví*, *vivía*, *vivíais*, *vivíamos*, *vivían*, *vivías*, *vivid*, *vivida*, *vividas*, *vivido*, *vividos*, *viviendo*, *viviera*, *vivierais*, *viviéramos*, *vivieran*, *vivieras*, *viviere*, *viviereis*, *viviéremos*, *vivieren*, *vivieres*, *vivieron*, *viviese*, *vivieseis*, *viviésemos*, *viviesen*, *vivieses*, *vivimos*, *vivió*, *vivir*, *vivirá*, *vivirán*, *vivirás*, *viviré*, *viviréis*, *viviremos*, *viviría*, *viviríais*, *viviríamos*, *vivirían*, *vivirías*, *vivís*, *viviste*, *vivisteis*, *vivo* — "live".

Without exceptions, the conjugated forms of these regular verbs follow the pattern **root** + **ending**, where the root remains invariant, while the ending starts with a vowel, and consists of at most two syllables (discounting final *~mos*). We shall refer to this pattern as the "canonical root-ending dichotomy" hereafter.

Applying the clustering algorithm to the conjugated forms of the three Spanish verbs above, we obtain

{*coma*, *comáis*, *comamos*, *coman*, *comas*, *comás*, *come*, *comé*, *comed*, *coméis*, *comemos*, *comen*, *comer*, *comerá*, *comerán*, *comerás*, *comeré*, *comeréis*, *comeremos*, *comería*, *comeríais*, *comeríamos*, *comerían*, *comerías*, *comes*, *comés*, *comí*, *comía*, *comíais*, *comíamos*, *comían*, *comías*, *comida*, *comidas*, *comido*, *comidos*, *comiendo*, *comiera*, *comierais*, *comiéramos*, *comieran*, *comieras*, *comiere*, *comiereis*, *comiéremos*, *comieren*, *comieres*, *comieron*, *comiese*, *comieseis*, *comiésemos*, *comiesen*, *comieses*, *comimos*, *comió*, *comiste*, *comisteis*, *como*},

{*habla*, *hablá*, *hablaba*, *hablabais*, *hablábamos*, *hablaban*, *hablabas*, *hablad*, *hablada*, *habladas*, *hablado*, *hablados*, *habláis*, *hablamos*, *hablan*, *hablando*, *hablar*, *hablara*, *hablará*, *hablarais*, *hablaríamos*, *hablaran*, *hablarán*, *hablaras*, *hablarás*, *hablare*, *hablaré*, *hablareis*, *hablaréis*, *hablaremos*, *hablaremos*, *hablaren*, *hablares*, *hablaría*, *hablaríais*, *hablaríamos*, *hablarían*, *hablarías*, *hablaron*, *hablas*, *hablás*, *hablase*, *hablaseis*, *hablásemos*, *hablasen*, *hablases*, *hablaste*, *hablasteis*, *hable*, *hablé*, *habléis*, *hablemos*, *hablen*, *hables*, *hablés*, *hablo*, *habló*},

{*viva*, *viváis*, *vivamos*, *vivan*, *vivas*, *vivás*, *vive*, *viven*, *vives*, *viví*, *vivía*, *vivíais*, *vivíamos*, *vivían*, *vivías*, *vivid*, *vivida*, *vividas*, *vivido*, *vividos*, *viviendo*, *viviera*, *vivierais*, *viviéramos*, *vivieran*, *vivieras*, *viviere*, *viviereis*, *viviéremos*, *vivieren*, *vivieres*, *vivieron*, *viviese*, *vivieseis*, *viviésemos*, *viviesen*, *vivieses*, *vivimos*, *vivió*, *vivir*, *vivirá*, *vivirán*, *vivirás*, *viviré*, *viviréis*, *viviremos*, *viviría*, *viviríais*, *viviríamos*, *vivirían*, *vivirías*, *vivís*, *viviste*, *vivisteis*, *vivo*}.

Example 6.11.3. Our typology for irregular Spanish verbs is largely based on Wiktionary (`https://en.wiktionary.org/wiki/Wiktionary:Spanish_verb_inflection-table_templates`). There are about one hundred different types of irregularities on this offical template list. We are not going to exhaust all of them in our sample words.

In our clustering test in this example, we ignore some "trivially irregular verbs", whose roots are at most off by a stress mark (which will be dropped in the effective spelling, anyway) in conjugated forms. We also ignore irregular types whose conjugation template follows the "canonical root-ending dichotomy", even though the detailed endings may deviate from the three aforementioned regular types (for example, the conjugations of *distinguir* and *plañir*). Some highly irregular verbs in common use are also excluded from our test.

The typology on the aforementioned Wiktionary page regards [*o*, *ue*]**X***ar*, [*o*, *ue*]**X***er* ending and [*o*, *ue*]**X***ir* ending as three separate cases. We pick only one of these three scenarios for our test below, because our algorithm treats vowel alternations and endings independently. The same principle applies to other types of vowel alternations and consonant changes. What remain in our sample pool are the irregular verbs exhibiting systematic vowel alternation and/or consonant changes, which account for the overwhelming majority of irregular factors in Spanish verbs.

Concretely speaking, we send the following verbs

adquiera, *adquieran*, *adquieras*, *adquiere*, *adquieren*, *adquieres*, *adquiero*, *adquiráis*, *adquiramos*, *adquirí*, *adquiría*, *adquiríais*, *adquiríamos*, *adquirían*, *adquirías*, *adquirid*, *adquirida*, *adquiridas*, *adquirido*, *adquiridos*, *adquiriendo*, *adquiriera*, *adquirierais*, *adquiriéramos*, *adquirieran*, *adquirieras*, *adquiriere*, *adquiriereis*, *adquiriéremos*, *adquirieren*, *adquirieres*, *adquirieron*, *adquiriese*, *adquirieseis*, *adquiriésemos*, *adquiriesen*, *adquirieses*, *adquirimos*, *adquirió*, *adquirir*, *adquirirá*, *adquirirán*, *adquirirás*, *adquiriré*, *adquiriréis*, *adquiriremos*, *adquiriría*, *adquiriríais*, *adquiriríamos*, *adquirirían*, *adquirirías*, *adquirís*, *adquiriste*, *adquiristeis* — "acquire";

anda, *andá*, *andaba*, *andabais*, *andábamos*, *andaban*, *andabas*, *andad*, *andada*, *andadas*, *andado*, *andados*, *andáis*, *andamos*, *andan*, *andando*, *andar*, *andará*, *andarán*, *andarás*, *andaré*, *andaréis*, *andaremos*, *andaría*, *andaríais*, *andaríamos*, *andarían*, *andarías*, *andas*, *andás*, *ande*, *andéis*, *andemos*, *anden*, *andes*, *ando*, *anduve*, *anduviera*, *anduvierais*, *anduviéramos*, *anduvieran*, *anduvieras*, *anduviere*, *anduviereis*, *anduviéremos*, *anduvieren*, *anduvieres*, *anduvieron*, *anduviese*, *anduvieseis*, *anduviésemos*, *anduviesen*, *anduvieses*, *anduvimos*, *anduviste*, *anduvisteis*, *anduvo* — "walk";

argüí, *argüía*, *argüíais*, *argüíamos*, *argüían*, *argüías*, *argüid*, *argüida*, *argüidas*, *argüido*, *argüidos*, *argüimos*, *argüir*, *argüirá*, *argüirán*, *argüirás*, *argüiré*, *argüiréis*, *argüiremos*, *argüiría*, *argüiríais*, *argüiríamos*, *argüirían*, *argüirías*, *argüís*, *argüiste*, *argüisteis*, *arguya*, *arguyáis*, *arguyamos*, *arguyan*, *arguyas*, *arguye*, *arguyen*, *arguyendo*, *arguyera*, *arguyerais*, *arguyéramos*, *arguyeran*, *arguyeras*, *arguyere*, *arguyereis*, *arguyéremos*, *arguyeren*, *arguyeres*, *arguyeron*, *arguyes*, *arguyese*, *arguyeseis*, *arguyésemos*, *arguyesen*, *arguyeses*, *arguyo*, *arguyó* — "infer";

ase, *asen*, *ases*, *asga*, *asgáis*, *asgamos*, *asgan*, *asgas*, *asgo*, *así*, *asía*, *asíais*, *asíamos*, *asían*, *asías*, *asid*, *asida*, *asidas*, *asido*, *asidos*, *asiendo*, *asiera*, *asierais*, *asiéramos*, *asieran*, *asieras*, *asiere*, *asiereis*, *asiéremos*, *asieren*, *asieres*, *asieron*, *asiese*, *asieseis*, *asiésemos*, *asiesen*, *asieses*, *asimos*, *asió*, *asir*, *asirá*, *asirán*, *asirás*, *asiré*, *asiréis*, *asiremos*, *asiría*, *asiríais*, *asiríamos*, *asirían*, *asirías*, *asís*, *asiste*, *asisteis* — "grab";

avergoncé, *avergoncéis*, *avergoncemos*, *avergonzaba*, *avergonzabais*, *avergonzábamos*, *avergonzaban*, *avergonzabas*, *avergonzad*, *avergonzada*, *avergonzadas*, *avergonzado*, *avergonzados*, *avergonzáis*, *avergonzamos*, *avergonzando*, *avergonzar*, *avergonzara*, *avergonzará*, *avergonzarais*, *avergonzáramos*, *avergonzaran*, *avergonzarán*, *avergonzaras*, *avergonzarás*, *avergonzare*, *avergonzaré*, *avergonzareis*, *avergonzaréis*, *avergonzaremos*, *avergonzáremos*, *avergonzaren*, *avergonzares*, *avergonzaría*, *avergonzaríais*, *avergonzaríamos*, *avergonzarían*, *avergonzarías*, *avergonzaron*, *avergonzase*, *avergonzaseis*, *avergonzásemos*, *avergonzasen*, *avergonzases*, *avergonzaste*, *avergonzasteis*, *avergonzó*, *avergüence*, *avergüencen*, *avergüences*, *avergüenza*, *avergüenzan*, *avergüenzas*, *avergüenzo* — "embarrass";

cabe, *cabed*, *cabéis*, *cabemos*, *caben*, *caber*, *cabes*, *cabía*, *cabíais*, *cabíamos*, *cabían*, *cabías*, *cabida*, *cabidas*, *cabido*, *cabidos*, *cabiendo*, *cabrá*, *cabrán*, *cabrás*, *cabré*, *cabréis*, *cabremos*, *cabría*, *cabríais*, *cabríamos*, *cabrían*, *cabrías*, *cupe*, *cupiera*, *cupierais*, *cupiéramos*, *cupieran*, *cupieras*, *cupiere*, *cupiereis*, *cupiéremos*, *cupieren*, *cupieres*, *cupieron*, *cupiese*, *cupieseis*, *cupiésemos*, *cupiesen*, *cupieses*, *cupimos*, *cupiste*, *cupisteis*, *cupo*, *quepa*, *quepáis*, *quepamos*, *quepan*, *quepas*, *quepo* — "fit";

cae, *caed*, *caéis*, *caemos*, *caen*, *caer*, *caerá*, *caerán*, *caerás*, *caeré*, *caeréis*, *caeremos*, *caería*, *caeríais*, *caeríamos*, *caerían*, *caerías*, *caes*, *caí*, *caía*, *caíais*, *caíamos*, *caían*, *caías*, *caída*, *caídas*, *caído*, *caídos*, *caiga*, *caigáis*, *caigamos*, *caigan*, *caigas*, *caigo*, *caímos*, *caíste*, *caísteis*, *cayendo*, *cayera*, *cayerais*, *cayéramos*, *cayeran*, *cayeras*, *cayere*, *cayereis*, *cayéremos*, *cayeren*, *cayeres*, *cayeron*, *cayese*, *cayeseis*, *cayésemos*, *cayesen*, *cayeses*, *cayó* — "fall";

coge, *coged*, *cogéis*, *cogemos*, *cogen*, *coger*, *cogerá*, *cogerán*, *cogerás*, *cogeré*, *cogeréis*, *cogeremos*, *cogería*, *cogeríais*, *cogeríamos*, *cogerían*, *cogerías*, *coges*, *cogí*, *cogía*, *cogíais*, *cogíamos*, *cogían*, *cogías*, *cogida*, *cogidas*, *cogido*, *cogidos*, *cogiendo*, *cogiera*, *cogierais*, *cogiéramos*, *cogieran*, *cogieras*, *cogiere*, *cogiereis*, *cogiéremos*, *cogieren*, *cogieres*, *cogieron*, *cogiese*, *cogieseis*, *cogiésemos*, *cogiesen*, *cogieses*, *cogimos*, *cogió*, *cogiste*, *cogisteis*, *coja*, *cojáis*, *cojamos*, *cojan*, *cojas*, *cojo* — "catch";

conoce, *conocé*, *conoced*, *conocéis*, *conocemos*, *conocen*, *conocer*, *conocerá*, *conocerán*, *conocerás*, *conoceré*, *conoceréis*, *conoceremos*, *conocería*, *conoceríais*, *conoceríamos*, *conocerían*, *conocerías*, *conoces*, *conocés*, *conocí*, *conocía*, *conocíais*, *conocíamos*, *conocían*, *conocías*, *conocida*, *conocidas*, *conocido*, *conocidos*, *conociendo*, *conociera*, *conocierais*, *conociéramos*, *conocieran*, *conocieras*, *conociere*, *conociereis*, *conociéremos*, *conocieren*, *conocieres*, *conocieron*, *conociese*, *conocieseis*, *conociésemos*, *conociesen*, *conocieses*, *conocimos*, *conoció*, *conociste*, *conocisteis*, *conozca*, *conozcáis*, *conozcamos*, *conozcan*, *conozcas*, *conozco* — "know";

cruce, *crucé*, *crucéis*, *crucemos*, *crucen*, *cruces*, *crucés*, *cruza*, *cruzá*, *cruzaba*, *cruzabais*, *cruzábamos*, *cruzaban*, *cruzabas*, *cruzad*, *cruzada*, *cruzadas*, *cruzado*, *cruzados*, *cruzáis*, *cruzamos*, *cruzan*, *cruzando*, *cruzar*, *cruzara*, *cruzará*, *cruzarais*, *cruzáramos*, *cruzaran*, *cruzarán*, *cruzaras*, *cruzarás*, *cruzare*, *cruzaré*, *cruzareis*, *cruzaréis*, *cruzaremos*, *cruzáremos*, *cruzaren*, *cruzares*, *cruzaría*, *cruzaríais*, *cruzaríamos*, *cruzarían*, *cruzarías*, *cruzaron*, *cruzas*, *cruzás*, *cruzase*, *cruzaseis*, *cruzásemos*, *cruzasen*, *cruzases*, *cruzaste*, *cruzasteis*, *cruzo*, *cruzó* — "cross";

decía, *decíais*, *decíamos*, *decían*, *decías*, *decid*, *decimos*, *decir*, *decís*, *di*, *dice*, *dicen*, *dices*, *dicha*, *dichas*, *dicho*, *dichos*, *diciendo*, *diga*, *digáis*, *digamos*, *digan*, *digas*, *digo*, *dije*, *dijera*, *dijerais*, *dijéramos*, *dijeran*, *dijeras*, *dijere*, *dijereis*, *dijéremos*, *dijeren*, *dijeres*, *dijeron*, *dijese*, *dijeseis*, *dijésemos*, *dijesen*, *dijeses*, *dijimos*, *dijiste*, *dijisteis*, *dijo*, *dirá*, *dirán*, *dirás*, *diré*, *diréis*, *diremos*, *diría*, *diríais*, *diríamos*, *dirían*, *dirías* — "say";

delinca, *delincáis*, *delincamos*, *delincan*, *delincas*, *delinco*, *delinque*, *delinquen*, *delinques*, *delinquí*, *delinquía*, *delinquíais*, *delinquíamos*, *delinquían*, *delinquías*, *delinquid*, *delinquida*, *delinquidas*, *delinquido*, *delinquidos*, *delinquiendo*, *delinquiera*, *delinquierais*, *delinquiéramos*, *delinquieran*, *delinquieras*, *delinquiere*, *delinquiereis*,

delinquiéremos, *delinquieren*, *delinquieres*, *delinquieron*, *delinquiese*, *delinquieseis*, *delinquiésemos*, *delinquiesen*, *delinquieses*, *delinquimos*, *delinquió*, *delinquir*, *delinquirá*, *delinquirán*, *delinquirás*, *delinquiré*, *delinquiréis*, *delinquiremos*, *delinquiría*, *delinquiríais*, *delinquiríamos*, *delinquirían*, *delinquirías*, *delinquís*, *delinquiste*, *delinquisteis* — "commit a crime";

deshuesa, *deshuesá*, *deshuesan*, *deshuesas*, *deshuese*, *deshuesen*, *deshueses*, *deshueso*, *desosaba*, *desosabais*, *desosábamos*, *desosaban*, *desosabas*, *desosad*, *desosada*, *desosadas*, *desosado*, *desosados*, *desosáis*, *desosamos*, *desosando*, *desosar*, *desosara*, *desosará*, *desosarais*, *desosáramos*, *desosaran*, *desosarán*, *desosaras*, *desosarás*, *desosare*, *desosaré*, *desosareis*, *desosaréis*, *desosaremos*, *desosáremos*, *desosaren*, *desosares*, *desosaría*, *desosaríais*, *desosaríamos*, *desosarían*, *desosarías*, *desosaron*, *desosás*, *desosase*, *desosaseis*, *desosásemos*, *desosasen*, *desosases*, *desosaste*, *desosasteis*, *desosé*, *desoséis*, *desosemos*, *desosó* — "debone";

discernáis, *discernamos*, *discerní*, *discernía*, *discerníais*, *discerníamos*, *discernían*, *discernías*, *discernid*, *discernida*, *discernidas*, *discernido*, *discernidos*, *discerniendo*, *discerniera*, *discernierais*, *discerniéramos*, *discernieran*, *discernieras*, *discerniere*, *discerniereis*, *discerniéremos*, *discernieren*, *discernieres*, *discernieron*, *discerniese*, *discernieseis*, *discerniésemos*, *discerniesen*, *discernieses*, *discernimos*, *discernió*, *discernir*, *discernirá*, *discernirán*, *discernirás*, *discerniré*, *discerniréis*, *discerniremos*, *discerniría*, *discerniríais*, *discerniríamos*, *discernirían*, *discernirías*, *discernís*, *discerniste*, *discernisteis*, *discierna*, *disciernan*, *disciernas*, *discierne*, *disciernen*, *disciernes*, *discierno* — "discern";

erra, *erraba*, *errabais*, *errábamos*, *erraban*, *errabas*, *errad*, *errada*, *erradas*, *errado*, *errados*, *erráis*, *erramos*, *erran*, *errando*, *errar*, *errara*, *errará*, *errarais*, *erráramos*, *erraran*, *errarán*, *erraras*, *errarás*, *errare*, *erraré*, *errareis*, *erraréis*, *erraremos*, *erráremos*, *erraren*, *errares*, *erraría*, *erraríais*, *erraríamos*, *errarían*, *errarías*, *erraron*, *erras*, *errase*, *erraseis*, *errásemos*, *errasen*, *errases*, *erraste*, *errasteis*, *erre*, *erré*, *erréis*, *erremos*, *erren*, *erres*, *erro*, *erró*, *yerra*, *yerran*, *yerras*, *yerre*, *yerren*, *yerres*, *yerro* — "miss";

juega, *juegan*, *juegas*, *juego*, *juegue*, *jueguen*, *juegues*, *jugá*, *jugaba*, *jugabais*, *jugábamos*, *jugaban*, *jugabas*, *jugad*, *jugada*, *jugadas*, *jugado*, *jugados*, *jugáis*, *jugamos*, *jugando*, *jugar*, *jugara*, *jugará*, *jugarais*, *jugáramos*, *jugaran*, *jugarán*, *jugaras*, *jugarás*, *jugare*, *jugaré*, *jugareis*, *jugaréis*, *jugaremos*, *jugáremos*, *jugaren*, *jugares*, *jugaría*, *jugaríais*, *jugaríamos*, *jugarían*, *jugarías*, *jugaron*, *jugás*, *jugase*, *jugaseis*, *jugásemos*, *jugasen*, *jugases*, *jugaste*, *jugasteis*, *jugó*, *jugué*, *juguéis*, *juguemos*, *jugués* — "play";

lea, *leáis*, *leamos*, *lean*, *leas*, *lee*, *leé*, *leed*, *leéis*, *leemos*, *leen*, *leer*, *leerá*, *leerán*, *leerás*, *leeré*, *leeréis*, *leeremos*, *leería*, *leeríais*, *leeríamos*, *leerían*, *leerías*, *lees*, *leés*, *leí*, *leía*, *leíais*, *leíamos*, *leían*, *leías*, *leída*, *leídas*, *leído*, *leídos*, *leímos*, *leíste*, *leísteis*, *leo*, *leyendo*, *leyera*, *leyerais*, *leyéramos*, *leyeran*, *leyeras*, *leyere*, *leyereis*, *leyéremos*, *leyeren*, *leyeres*, *leyeron*, *leyese*, *leyeseis*, *leyésemos*, *leyesen*, *leyeses*, *leyó* — "read";

oí, *oía*, *oíais*, *oíamos*, *oían*, *oías*, *oíd*, *oída*, *oídas*, *oído*, *oídos*, *oiga*, *oigáis*, *oigamos*, *oigan*, *oigas*, *oigo*, *oímos*, *oír*, *oirá*, *oirán*, *oirás*, *oiré*, *oiréis*, *oiremos*, *oiría*, *oiríais*, *oiríamos*, *oirían*, *oirías*, *oís*, *oíste*, *oísteis*, *oye*, *oyen*, *oyendo*, *oyera*, *oyerais*, *oyéramos*, *oyeran*, *oyeras*, *oyere*, *oyereis*, *oyéremos*, *oyeren*, *oyeres*, *oyeron*, *oyes*, *oyese*, *oyeseis*, *oyésemos*, *oyesen*, *oyeses*, *oyó* — "hear";

pedí, *pedía*, *pedíais*, *pedíamos*, *pedían*, *pedías*, *pedid*, *pedida*, *pedidas*, *pedido*, *pedidos*, *pedimos*, *pedir*, *pedirá*, *pedirán*, *pedirás*, *pediré*, *pediréis*, *pediremos*, *pediría*, *pediríais*, *pediríamos*, *pedirían*, *pedirías*, *pedís*, *pediste*, *pedisteis*, *pida*, *pidáis*, *pidamos*, *pidan*, *pidas*, *pide*, *piden*, *pides*, *pidiendo*, *pidiera*, *pidierais*, *pidiéramos*, *pidieran*, *pidieras*, *pidiere*, *pidiereis*, *pidiéremos*, *pidieren*, *pidieres*, *pidieron*, *pidiese*, *pidieseis*, *pidiésemos*, *pidiesen*, *pidieses*, *pidió*, *pido* — "request";

pervertí, *pervertía*, *pervertíais*, *pervertíamos*, *pervertían*, *pervertías*, *pervertid*, *pervertida*, *pervertidas*, *pervertido*, *pervertidos*, *pervertimos*, *pervertir*, *pervertirá*, *pervertirán*, *pervertirás*, *pervertiré*, *pervertiréis*, *pervertiremos*, *pervertiría*, *pervertiríais*, *pervertiríamos*, *pervertirían*, *pervertirías*, *pervertís*, *pervertiste*, *pervertisteis*, *pervierta*, *perviertan*, *perviertas*, *pervierte*, *pervierten*, *perviertes*, *pervierto*, *pervirtáis*, *pervirtamos*, *pervirtás*, *pervirtiendo*, *pervirtiera*, *pervirtierais*, *pervirtiéramos*, *pervirtieran*, *pervirtieras*, *pervirtiere*, *pervirtiereis*, *pervirtiéremos*, *pervirtieren*, *pervirtieres*, *pervirtieron*, *pervirtiese*, *pervirtieseis*, *pervirtiésemos*, *pervirtiesen*, *pervirtieses*, *pervirtió* — "corrupt";

podáis, *podamos*, *poded*, *podéis*, *podemos*, *poder*, *podía*, *podíais*, *podíamos*, *podían*, *podías*, *podida*, *podidas*, *podido*, *podidos*, *podrá*, *podrán*, *podrás*, *podré*, *podréis*, *podremos*, *podría*, *podríais*, *podríamos*, *podrían*, *podrías*, *pude*, *pudiendo*, *pudiera*, *pudierais*, *pudiéramos*, *pudieran*, *pudieras*, *pudiere*, *pudiereis*, *pudiéremos*, *pudieren*, *pudieres*, *pudieron*, *pudiese*, *pudieseis*, *pudiésemos*, *pudiesen*, *pudieses*, *pudimos*, *pudiste*, *pudisteis*, *pudo*, *pueda*, *puedan*, *puedas*, *puede*, *pueden*, *puedes*, *puedo* — "can";

prevén, *prevendrá*, *prevendrán*, *prevendrás*, *prevendré*, *prevendréis*, *prevendremos*, *prevendría*, *prevendríais*, *prevendríamos*, *prevendrían*, *prevendrías*, *prevenga*, *prevengáis*, *prevengamos*, *prevengan*, *prevengas*, *prevengo*, *prevenía*, *preveníais*, *preveníamos*, *prevenían*, *prevenías*, *prevenid*, *prevenida*, *prevenidas*, *prevenido*, *prevenidos*, *prevenimos*, *prevenir*, *prevenís*, *previene*, *previenen*, *previenes*, *previne*, *previniendo*, *previniera*, *previnierais*, *previniéramos*, *previnieran*, *previnieras*, *previniere*, *previniereis*, *previniéremos*, *previnieren*, *previnieres*, *previnieron*, *previniese*, *previnieseis*, *previniésemos*, *previniesen*, *previnieses*, *previnimos*, *previniste*, *previnisteis*, *previno* — "prevent";

produce, *producen*, *produces*, *producí*, *producía*, *producíais*, *producíamos*, *producían*, *producías*, *producid*, *producida*, *producidas*, *producido*, *producidos*, *produciendo*, *producimos*, *producir*, *producirá*, *producirán*, *producirás*, *produciré*, *produciréis*, *produciremos*, *produciría*, *produciríais*, *produciríamos*, *producirían*, *producirías*, *producís*, *produje*, *produjera*, *produjerais*, *produjéramos*, *produjeran*, *produjeras*, *produjere*, *produjereis*, *produjéremos*, *produjeren*, *produjeres*, *produjeron*, *produjese*, *produjeseis*, *produjésemos*, *produjesen*, *produjeses*, *produjimos*, *produjiste*, *produjisteis*, *produjo*, *produzca*, *produzcáis*, *produzcamos*, *produzcan*, *produzcas*, *produzcás*, *produzco* — "produce";

queráis, *queramos*, *querás*, *queré*, *quered*, *queréis*, *queremos*, *querer*, *querés*, *quería*, *queríais*, *queríamos*, *querían*, *querías*, *querida*, *queridas*, *querido*, *queridos*, *queriendo*, *querrá*, *querrán*, *querrás*, *querré*, *querréis*, *querremos*, *querría*, *querríais*, *querríamos*, *querrían*, *querrías*, *quiera*, *quieran*, *quieras*, *quiere*, *quieren*, *quieres*, *quiero*, *quise*, *quisiera*, *quisierais*, *quisiéramos*, *quisieran*, *quisieras*, *quisiere*, *quisiereis*, *quisiéremos*, *quisieren*, *quisieres*, *quisieron*, *quisiese*, *quisieseis*, *quisiésemos*, *quisiesen*, *quisieses*, *quisimos*, *quisiste*, *quisisteis*, *quiso* — "desire";

rehuí, *rehuía*, *rehuíais*, *rehuíamos*, *rehuían*, *rehuías*, *rehuid*, *rehuida*, *rehuidas*, *rehuido*, *rehuidos*, *rehuimos*, *rehuir*, *rehuirá*, *rehuirán*, *rehuirás*, *rehuiré*, *rehuiréis*, *rehuiremos*, *rehuiría*, *rehuiríais*, *rehuiríamos*, *rehuirían*, *rehuirías*, *rehuís*, *rehuiste*, *rehuisteis*, *rehúya*, *rehuyáis*, *rehuyamos*, *rehúyan*, *rehúyas*, *rehúye*, *rehúyen*, *rehuyendo*, *rehuyera*, *rehuyerais*, *rehuyéramos*, *rehuyeran*, *rehuyeras*, *rehuyere*, *rehuyereis*, *rehuyéremos*, *rehuyeren*, *rehuyeres*, *rehuyeron*, *rehúyes*, *rehuyese*, *rehuyeseis*, *rehuyésemos*, *rehuyesen*, *rehuyeses*, *rehuyó*, *rehúyo* — "avoid";

repón, *repondrá*, *repondrán*, *repondrás*, *repondré*, *repondréis*, *repondremos*, *repondría*, *repondríais*, *repondríamos*, *repondrían*, *repondrías*, *repone*, *reponed*, *reponéis*, *reponemos*, *reponen*, *reponer*, *repones*, *reponga*, *repongáis*, *repongamos*, *repongan*, *repongas*, *repongo*, *reponía*, *reponíais*, *reponíamos*, *reponían*, *reponías*, *reponiendo*, *repuesta*, *repuestas*, *repuesto*, *repuestos*, *repuse*, *repusiera*, *repusierais*, *repusiéramos*, *repusieran*, *repusieras*, *repusiere*, *repusiereis*, *repusiéremos*, *repusieren*, *repusieres*, *repusieron*, *repusiese*, *repusieseis*, *repusiésemos*, *repusiesen*, *repusieses*, *repusimos*, *repusiste*, *repusisteis*, *repuso* — "replace";

retén, *retendrá*, *retendrán*, *retendrás*, *retendré*, *retendréis*, *retendremos*, *retendría*, *retendríais*, *retendríamos*, *retendrían*, *retendrías*, *retened*, *retenéis*, *retenemos*, *retener*, *retenga*, *retengáis*, *retengamos*, *retengan*, *retengas*, *retengo*, *retenía*, *reteníais*, *reteníamos*, *retenían*, *retenías*, *retenida*, *retenidas*, *retenido*, *retenidos*, *reteniendo*, *retiene*, *retienen*, *retienes*, *retuve*, *retuviera*, *retuvierais*, *retuviéramos*, *retuvieran*, *retuvieras*, *retuviere*, *retuviereis*, *retuviéremos*, *retuvieren*, *retuvieres*, *retuvieron*, *retuviese*, *retuvieseis*, *retuviésemos*, *retuviesen*, *retuvieses*, *retuvimos*, *retuviste*, *retuvisteis*, *retuvo* — "retain";

sal, *saldrá*, *saldrán*, *saldrás*, *saldré*, *saldréis*, *saldremos*, *saldría*, *saldríais*, *saldríamos*, *saldrían*, *saldrías*, *sale*, *salen*, *sales*, *salga*, *salgáis*, *salgamos*, *salgan*, *salgas*, *salgo*, *salí*, *salía*, *salíais*, *salíamos*, *salían*, *salías*, *salid*, *salida*, *salidas*, *salido*, *salidos*, *saliendo*, *saliera*, *salierais*, *saliéramos*, *salieran*, *salieras*, *saliere*, *saliereis*, *saliéremos*, *salieren*, *salieres*, *salieron*, *saliese*, *salieseis*, *saliésemos*, *saliesen*, *salieses*, *salimos*, *salió*, *salir*, *salís*, *saliste*, *salisteis* — "leave";

trae, *traed*, *traéis*, *traemos*, *traen*, *traer*, *traerá*, *traerán*, *traerás*, *traeré*, *traeréis*, *traeremos*, *traería*, *traeríais*, *traeríamos*, *traerían*, *traerías*, *traes*, *traía*, *traíais*, *traíamos*, *traían*, *traías*, *traída*, *traídas*, *traído*, *traídos*, *traiga*, *traigáis*, *traigamos*, *traigan*, *traigas*, *traigo*, *traje*, *trajera*, *trajerais*, *trajéramos*, *trajeran*, *trajeras*, *trajere*, *trajereis*, *trajéremos*, *trajeren*, *trajeres*, *trajeron*, *trajese*, *trajeseis*, *trajésemos*, *trajesen*, *trajeses*, *trajimos*, *trajiste*, *trajisteis*, *trajo*, *trayendo* — "bring";

ve, *vea*, *veáis*, *veamos*, *vean*, *veas*, *ved*, *veía*, *veíais*, *veíamos*, *veían*, *veías*, *veis*, *vemos*, *ven*, *veo*, *ver*, *verá*, *verán*, *verás*, *veré*, *veréis*, *veremos*, *vería*, *veríais*, *veríamos*, *verían*, *verías*, *ves*, *vi*, *viendo*, *viera*, *vierais*, *viéramos*, *vieran*, *vieras*, *viere*, *viereis*, *viéremos*, *vieren*, *vieres*, *vieron*, *viese*, *vieseis*, *viésemos*, *viesen*, *vieses*, *vimos*, *vio*, *vista*, *vistas*, *viste*, *visteis*, *visto*, *vistos* — "see";

yace, *yaced*, *yacéis*, *yacemos*, *yacen*, *yacer*, *yacerá*, *yacerán*, *yacerás*, *yaceré*, *yaceréis*, *yaceremos*, *yacería*, *yaceríais*, *yaceríamos*, *yacerían*, *yacerías*, *yaces*, *yací*, *yacía*, *yacíais*, *yacíamos*, *yacían*, *yacías*, *yacida*, *yacidas*, *yacido*, *yacidos*, *yaciendo*, *yaciera*, *yacierais*, *yaciéramos*, *yacieran*, *yacieras*, *yaciere*, *yaciereis*, *yaciéremos*, *yacieren*, *yacieres*, *yacieron*, *yaciese*, *yacieseis*, *yaciésemos*, *yaciesen*, *yacieses*, *yacimos*, *yació*, *yaciste*, *yacisteis*, *yaga*, *yagáis*, *yagamos*, *yagan*, *yagas*, *yago*, *yaz*, *yazca*, *yazcáis*, *yazcamos*, *yazcan*, *yazcas*, *yazco*, *yazga*, *yazgáis*, *yazgamos*, *yazgan*, *yazgas*, *yazgo* — "recline"

as input, and receive

{*adquiera*, *adquieran*, *adquieras*, *adquiere*, *adquieren*, *adquieres*, *adquiero*, *adquiráis*, *adquiramos*, *adquirí*, *adquiría*, *adquiríais*, *adquiríamos*, *adquirían*, *adquirías*, *adquirid*, *adquirida*, *adquiridas*, *adquirido*, *adquiridos*, *adquiriendo*, *adquiriera*, *adquirierais*, *adquiriéramos*, *adquirieran*, *adquirieras*, *adquiriere*, *adquiriereis*, *adquiriéremos*, *adquirieren*, *adquirieres*, *adquirieron*, *adquiriese*, *adquirieseis*, *adquiriésemos*, *adquiriesen*, *adquirieses*, *adquirimos*, *adquirió*, *adquirir*, *adquirirá*, *adquirirán*, *adquirirás*, *adquiriré*, *adquiriréis*, *adquiriremos*, *adquiriría*, *adquiriríais*, *adquiriríamos*, *adquirirían*, *adquirirías*, *adquirís*, *adquiriste*, *adquiristeis*},

{*anda*, *andá*, *andaba*, *andabais*, *andábamos*, *andaban*, *andabas*, *andad*, *andada*, *andadas*, *andado*, *andados*, *andáis*, *andamos*, *andan*, *andando*, *andar*, *andará*, *andarán*, *andarás*, *andaré*, *andaréis*, *andaremos*, *andaría*, *andaríais*, *andaríamos*, *andarían*, *andarías*, *andas*, *andás*, *ande*, *andéis*, *andemos*, *anden*, *andes*, *ando*, *anduve*, *anduviera*, *anduvierais*, *anduviéramos*, *anduvieran*, *anduvieras*, *anduviere*, *anduviereis*, *anduviéremos*, *anduvieren*, *anduvieres*, *anduvieron*, *anduviese*, *anduvieseis*, *anduviésemos*, *anduviesen*, *anduvieses*, *anduvimos*, *anduviste*, *anduvisteis*, *anduvo*},

{*argüí*, *argüía*, *argüíais*, *argüíamos*, *argüían*, *argüías*, *argüid*, *argüida*, *argüidas*, *argüido*, *argüidos*, *argüimos*, *argüir*, *argüirá*, *argüirán*, *argüirás*, *argüiré*, *argüiréis*, *argüiremos*, *argüiría*, *argüiríais*, *argüiríamos*, *argüirían*, *argüirías*, *argüís*, *argüiste*, *argüisteis*, *arguya*, *arguyáis*, *arguyamos*, *arguyan*, *arguyas*, *arguye*, *arguyen*, *arguyendo*, *arguyera*, *arguyerais*, *arguyéramos*, *arguyeran*, *arguyeras*, *arguyere*, *arguyereis*, *arguyéremos*, *arguyeren*, *arguyeres*, *arguyeron*, *arguyes*, *arguyese*, *arguyeseis*, *arguyésemos*, *arguyesen*, *arguyeses*, *arguyo*, *arguyó*},

{*ase*, *asen*, *ases*, *asga*, *asgáis*, *asgamos*, *asgan*, *asgas*, *asgo*, *así*, *asía*, *asíais*, *asíamos*, *asían*, *asías*, *asid*, *asida*, *asidas*, *asido*, *asidos*, *asiendo*, *asiera*, *asierais*, *asiéramos*, *asieran*, *asieras*, *asiere*, *asiereis*, *asiéremos*, *asieren*, *asieres*, *asieron*, *asiese*, *asieseis*, *asiésemos*, *asiesen*, *asieses*, *asimos*, *asió*, *asir*, *asirá*, *asirán*, *asirás*, *asiré*, *asiréis*, *asiremos*, *asiría*, *asiríais*, *asiríamos*, *asirían*, *asirías*, *asís*, *asiste*, *asisteis*},

{*avergoncé*, *avergoncéis*, *avergoncemos*, *avergonzaba*, *avergonzabais*, *avergonzábamos*, *avergonzaban*, *avergonzabas*, *avergonzad*, *avergonzada*, *avergonzadas*, *avergonzado*, *avergonzados*, *avergonzáis*, *avergonzamos*, *avergonzando*, *avergonzar*, *avergonzara*, *avergonzará*, *avergonzarais*, *avergonzáramos*, *avergonzaran*, *avergonzarán*, *avergonzaras*, *avergonzarás*, *avergonzare*, *avergonzaré*, *avergonzareis*, *avergonzaréis*, *avergonzaremos*, *avergonzáremos*, *avergonzaren*, *avergonzares*, *avergonzaría*, *avergonzaríais*, *avergonzaríamos*, *avergonzarían*, *avergonzarías*, *avergonzaron*, *avergonzase*, *avergonzaseis*, *avergonzásemos*, *avergonzasen*, *avergonzases*, *avergonzaste*, *avergonzasteis*, *avergonzó*, *avergüence*, *avergüencen*, *avergüences*, *avergüenza*, *avergüenzan*, *avergüenzas*, *avergüenzo*},

{*cabe*, *cabed*, *cabéis*, *cabemos*, *caben*, *caber*, *cabes*, *cabía*, *cabíais*, *cabíamos*, *cabían*, *cabías*, *cabida*, *cabidas*, *cabido*, *cabidos*, *cabiendo*, *cabrá*, *cabrán*, *cabrás*, *cabré*, *cabréis*, *cabremos*, *cabría*, *cabríais*, *cabríamos*, *cabrían*, *cabrías*, *cupe*, *cupiera*, *cupierais*, *cupiéramos*, *cupieran*, *cupieras*, *cupiere*, *cupiereis*, *cupiéremos*, *cupieren*, *cupieres*, *cupieron*, *cupiese*, *cupieseis*, *cupiésemos*, *cupiesen*, *cupieses*, *cupimos*, *cupiste*, *cupisteis*, *cupo*, *quepa*, *quepáis*, *quepamos*, *quepan*, *quepas*, *quepo*},

{*cae*, *caed*, *caéis*, *caemos*, *caen*, *caer*, *caerá*, *caerán*, *caerás*, *caeré*, *caeréis*, *caeremos*, *caería*, *caeríais*, *caeríamos*, *caerían*, *caerías*, *caes*, *caí*, *caía*, *caíais*, *caíamos*, *caían*, *caías*, *caída*, *caídas*, *caído*, *caídos*, *caiga*, *caigáis*, *caigamos*, *caigan*, *caigas*, *caigo*, *caímos*, *caíste*, *caísteis*, *cayendo*, *cayera*, *cayerais*, *cayéramos*, *cayeran*, *cayeras*, *cayere*, *cayereis*, *cayéremos*, *cayeren*, *cayeres*, *cayeron*, *cayese*, *cayeseis*, *cayésemos*, *cayesen*, *cayeses*, *cayó*},

{*coge*, *coged*, *cogéis*, *cogemos*, *cogen*, *coger*, *cogerá*, *cogerán*, *cogerás*, *cogeré*, *cogeréis*, *cogeremos*, *cogería*, *cogeríais*, *cogeríamos*, *cogerían*, *cogerías*, *coges*, *cogí*, *cogía*, *cogíais*, *cogíamos*, *cogían*, *cogías*, *cogida*, *cogidas*, *cogido*, *cogidos*, *cogiendo*, *cogiera*, *cogierais*, *cogiéramos*, *cogieran*, *cogieras*, *cogiere*, *cogiereis*, *cogiéremos*, *cogieren*, *cogieres*, *cogieron*, *cogiese*, *cogieseis*, *cogiésemos*, *cogiesen*, *cogieses*, *cogimos*, *cogió*, *cogiste*, *cogisteis*, *coja*, *cojáis*, *cojamos*, *cojan*, *cojas*, *cojo*},

{*conoce*, *conocé*, *conoced*, *conocéis*, *conocemos*, *conocen*, *conocer*, *conocerá*, *conocerán*, *conocerás*, *conoceré*, *conoceréis*, *conoceremos*, *conocería*, *conoceríais*, *conoceríamos*, *conocerían*, *conocerías*, *conoces*, *conocés*, *conocí*, *conocía*, *conocíais*, *conocíamos*, *conocían*, *conocías*, *conocida*, *conocidas*, *conocido*, *conocidos*, *conociendo*, *conociera*, *conocierais*, *conociéramos*, *conocieran*, *conocieras*, *conociere*, *conociereis*, *conociéremos*, *conocieren*, *conocieres*, *conocieron*, *conociese*, *conocieseis*, *conociésemos*, *conociesen*, *conocieses*, *conocimos*, *conoció*, *conociste*, *conocisteis*, *conozca*, *conozcáis*, *conozcamos*, *conozcan*, *conozcas*, *conozco*},

{*cruce*, *crucé*, *crucéis*, *crucemos*, *crucen*, *cruces*, *crucés*, *cruza*, *cruzá*, *cruzaba*, *cruzabais*, *cruzábamos*, *cruzaban*, *cruzabas*, *cruzad*, *cruzada*, *cruzadas*, *cruzado*, *cruzados*, *cruzáis*, *cruzamos*, *cruzan*, *cruzando*, *cruzar*, *cruzara*, *cruzará*, *cruzarais*, *cruzáramos*, *cruzaran*, *cruzarán*, *cruzaras*, *cruzarás*, *cruzare*, *cruzaré*, *cruzareis*, *cruzaréis*, *cruzaremos*, *cruzáremos*, *cruzaren*, *cruzares*, *cruzaría*, *cruzaríais*, *cruzaríamos*, *cruzarían*, *cruzarías*, *cruzaron*, *cruzas*, *cruzás*, *cruzase*, *cruzaseis*, *cruzásemos*, *cruzasen*, *cruzases*, *cruzaste*, *cruzasteis*, *cruzo*, *cruzó*},

{*decía*, *decíais*, *decíamos*, *decían*, *decías*, *decid*, *decimos*, *decir*, *decís*, *dice*, *dicen*, *dices*, *dicha*, *dichas*, *dicho*, *dichos*, *diciendo*, *diga*, *digáis*, *digamos*, *digan*, *digas*, *digo*, *dije*, *dijera*, *dijerais*, *dijéramos*, *dijeran*, *dijeras*, *dijere*, *dijereis*, *dijéremos*, *dijeren*, *dijeres*, *dijeron*, *dijese*, *dijeseis*, *dijésemos*, *dijesen*, *dijeses*, *dijimos*, *dijiste*, *dijisteis*, *dijo*, *dirá*, *dirán*, *dirás*, *diré*, *diréis*, *diremos*, *diría*, *diríais*, *diríamos*, *dirían*, *dirías*, *discernáis*},

{*delinca*, *delincáis*, *delincamos*, *delincan*, *delincas*, *delinco*, *delinque*, *delinquen*, *delinques*, *delinquí*, *delinquía*, *delinquíais*, *delinquíamos*, *delinquían*, *delinquías*, *delinquid*, *delinquida*, *delinquidas*, *delinquido*, *delinquidos*, *delinquiendo*, *delinquiera*, *delinquierais*, *delinquiéramos*, *delinquieran*, *delinquieras*, *delinquiere*, *delinquiereis*, *delinquiéremos*, *delinquieren*, *delinquieres*, *delinquieron*, *delinquiese*, *delinquieseis*, *delinquiésemos*, *delinquiesen*, *delinquieses*, *delinquimos*, *delinquió*, *delinquir*, *delinquirá*, *delinquirán*, *delinquirás*, *delinquiré*, *delinquiréis*, *delinquiremos*, *delinquiría*, *delinquiríais*, *delinquiríamos*, *delinquirían*, *delinquirías*, *delinquís*, *delinquiste*, *delinquisteis*},

{*deshuesa*, *deshuesá*, *deshuesan*, *deshuesas*, *deshuese*, *deshuesen*, *deshueses*, *deshueso*, *desosaba*, *desosabais*, *desosábamos*, *desosaban*, *desosabas*, *desosad*, *desosada*, *desosadas*, *desosado*, *desosados*, *desosáis*, *desosamos*, *desosando*, *desosar*, *desosara*, *desosará*, *desosarais*, *desosáramos*, *desosaran*, *desosarán*, *desosaras*, *desosarás*, *desosare*, *desosaré*, *desosareis*, *desosaréis*, *desosaremos*, *desosáremos*, *desosaren*, *desosares*, *desosaría*, *desosaríais*, *desosaríamos*, *desosarían*, *desosarías*, *desosaron*, *desosás*, *desosase*, *desosaseis*, *desosásemos*, *desosasen*, *desosases*, *desosaste*, *desosasteis*, *desosé*, *desoséis*, *desosemos*, *desosó*},

{*di*},

{*discernamos*, *discerní*, *discernía*, *discerníais*, *discerníamos*, *discernían*, *discernías*, *discernid*, *discernida*, *discernidas*, *discernido*, *discernidos*, *discerniendo*, *discerniera*, *discernierais*, *discerniéramos*, *discernieran*, *discernieras*, *discerniere*, *discerniereis*, *discerniéremos*, *discernieren*, *discernieres*, *discernieron*, *discerniese*, *discernieseis*, *discerniésemos*, *discerniesen*, *discernieses*, *discernimos*, *discernió*, *discernir*, *discernirá*, *discernirán*, *discernirás*, *discerniré*, *discerniréis*, *discerniremos*, *discerniría*, *discerniríais*, *discerniríamos*, *discernirían*, *discernirías*, *discernís*, *discerniste*, *discernisteis*, *discierna*, *disciernan*, *disciernas*, *discierne*, *disciernen*, *disciernes*, *discierno*},

{*erra*, *erraba*, *errabais*, *errábamos*, *erraban*, *errabas*, *errad*, *errada*, *erradas*, *errado*, *errados*, *erráis*, *erramos*, *erran*, *errando*, *errar*, *errara*, *errará*, *errarais*, *erráramos*, *erraran*, *errarán*, *erraras*, *errarás*, *errare*, *erraré*, *errareis*, *erraréis*, *erraremos*, *erráremos*, *erraren*, *errares*, *erraría*, *erraríais*, *erraríamos*, *errarían*, *errarías*, *erraron*, *erras*, *errase*, *erraseis*, *errásemos*, *errasen*, *errases*, *erraste*, *errasteis*, *erre*, *erré*, *erréis*, *erremos*, *erren*, *erres*, *erro*, *erró*, *yerra*, *yerran*, *yerras*, *yerre*, *yerren*, *yerres*, *yerro*},

{*juega*, *juegan*, *juegas*, *juego*, *juegue*, *jueguen*, *juegues*, *jugá*, *jugaba*, *jugabais*, *jugábamos*, *jugaban*, *jugabas*, *jugad*, *jugada*, *jugadas*, *jugado*, *jugados*, *jugáis*, *jugamos*, *jugando*, *jugar*, *jugara*, *jugará*, *jugarais*, *jugáramos*, *jugaran*, *jugarán*, *jugaras*, *jugarás*, *jugare*, *jugaré*, *jugareis*, *jugaréis*, *jugaremos*, *jugáremos*, *jugaren*, *jugares*, *jugaría*, *jugaríais*, *jugaríamos*, *jugarían*, *jugarías*, *jugaron*, *jugás*, *jugase*, *jugaseis*, *jugásemos*, *jugasen*, *jugases*, *jugaste*, *jugasteis*, *jugó*, *jugué*, *juguéis*, *juguemos*, *jugués*},

{*lea*, *leáis*, *leamos*, *lean*, *leas*, *lee*, *leé*, *leed*, *leéis*, *leemos*, *leen*, *leer*, *leerá*, *leerán*, *leerás*, *leeré*, *leeréis*, *leeremos*, *leería*, *leeríais*, *leeríamos*, *leerían*, *leerías*, *lees*, *leés*, *leí*, *leía*, *leíais*, *leíamos*, *leían*, *leías*, *leída*, *leídas*, *leído*, *leídos*, *leímos*, *leíste*, *leísteis*, *leo*, *leyendo*, *leyera*, *leyerais*, *leyéramos*, *leyeran*, *leyeras*, *leyere*, *leyereis*, *leyéremos*, *leyeren*, *leyeres*, *leyeron*, *leyese*, *leyeseis*, *leyésemos*, *leyesen*, *leyeses*, *leyó*},

{*oí*, *oía*, *oíais*, *oíamos*, *oían*, *oías*, *oíd*, *oída*, *oídas*, *oído*, *oídos*, *oiga*, *oigáis*, *oigamos*, *oigan*, *oigas*, *oigo*, *oímos*, *oír*, *oirá*, *oirán*, *oirás*, *oiré*, *oiréis*, *oiremos*, *oiría*, *oiríais*, *oiríamos*, *oirían*, *oirías*, *oís*, *oíste*, *oísteis*, *oye*, *oyen*, *oyendo*, *oyera*, *oyerais*, *oyéramos*, *oyeran*, *oyeras*, *oyere*, *oyereis*, *oyéremos*, *oyeren*, *oyeres*, *oyeron*, *oyes*, *oyese*, *oyeseis*, *oyésemos*, *oyesen*, *oyeses*, *oyó*},

{*pedí*, *pedía*, *pedíais*, *pedíamos*, *pedían*, *pedías*, *pedid*, *pedida*, *pedidas*, *pedido*, *pedidos*, *pedimos*, *pedir*, *pedirá*, *pedirán*, *pedirás*, *pediré*, *pediréis*, *pediremos*, *pediría*, *pediríais*, *pediríamos*, *pedirían*, *pedirías*, *pedís*, *pediste*, *pedisteis*, *pida*, *pidáis*, *pidamos*, *pidan*, *pidas*, *pide*, *piden*, *pides*, *pidiendo*, *pidiera*, *pidierais*, *pidiéramos*, *pidieran*, *pidieras*, *pidiere*, *pidiereis*, *pidiéremos*, *pidieren*, *pidieres*, *pidieron*, *pidiese*, *pidieseis*, *pidiésemos*, *pidiesen*, *pidieses*, *pidió*, *pido*},

{*pervertí*, *pervertía*, *pervertíais*, *pervertíamos*, *pervertían*, *pervertías*, *pervertid*, *pervertida*, *pervertidas*, *pervertido*, *pervertidos*, *pervertimos*, *pervertir*, *pervertirá*, *pervertirán*, *pervertirás*, *pervertiré*, *pervertiréis*, *pervertiremos*, *pervertiría*, *pervertiríais*, *pervertiríamos*, *pervertirían*, *pervertirías*, *pervertís*, *pervertiste*, *pervertisteis*, *pervierta*, *perviertan*, *perviertas*, *pervierte*, *pervierten*, *perviertes*, *pervierto*, *pervirtáis*, *pervirtamos*, *pervirtás*, *pervirtiendo*, *pervirtiera*, *pervirtierais*, *pervirtiéramos*, *pervirtieran*, *pervirtieras*, *pervirtiere*, *pervirtiereis*, *pervirtiéremos*, *pervirtieren*, *pervirtieres*, *pervirtieron*, *pervirtiese*, *pervirtieseis*, *pervirtiésemos*, *pervirtiesen*, *pervirtieses*, *pervirtió*},

{*podáis*, *podamos*, *poded*, *podéis*, *podemos*, *poder*, *podía*, *podíais*, *podíamos*, *podían*, *podías*, *podida*, *podidas*, *podido*, *podidos*, *podrá*, *podrán*, *podrás*, *podré*, *podréis*, *podremos*, *podría*, *podríais*, *podríamos*, *podrían*, *podrías*, *pude*, *pudiendo*, *pudiera*, *pudierais*, *pudiéramos*, *pudieran*, *pudieras*, *pudiere*, *pudiereis*, *pudiéremos*, *pudieren*, *pudieres*, *pudieron*, *pudiese*, *pudieseis*, *pudiésemos*, *pudiesen*, *pudieses*, *pudimos*, *pudiste*, *pudisteis*, *pudo*, *pueda*, *puedan*, *puedas*, *puede*, *pueden*, *puedes*, *puedo*},

{*prevén*, *prevendrá*, *prevendrán*, *prevendrás*, *prevendré*, *prevendréis*, *prevendremos*, *prevendría*, *prevendríais*, *prevendríamos*, *prevendrían*, *prevendrías*, *prevenga*, *prevengáis*, *prevengamos*, *prevengan*, *prevengas*, *prevengo*, *prevenía*, *preveníais*, *preveníamos*, *prevenían*, *prevenías*, *prevenid*, *prevenida*, *prevenidas*, *prevenido*, *prevenidos*, *prevenimos*, *prevenir*, *prevenís*, *previene*, *previenen*, *previenes*, *previne*, *previniendo*, *previniera*, *previnierais*, *previniéramos*, *previnieran*, *previnieras*, *previniere*, *previniereis*, *previniéremos*, *previnieren*, *previnieres*, *previnieron*, *previniese*, *previnieseis*, *previniésemos*, *previniesen*, *previnieses*, *previnimos*, *previniste*, *previnisteis*, *previno*},

{*produce*, *producen*, *produces*, *producí*, *producía*, *producíais*, *producíamos*, *producían*, *producías*, *producid*, *producida*, *producidas*, *producido*, *producidos*, *produciendo*, *producimos*, *producir*, *producirá*, *producirán*, *producirás*, *produciré*, *produciréis*, *produciremos*, *produciría*, *produciríais*, *produciríamos*, *producirían*, *producirías*, *producís*, *produje*, *produjera*, *produjerais*, *produjéramos*, *produjeran*, *produjeras*, *produjere*, *produjereis*, *produjéremos*, *produjeren*, *produjeres*, *produjeron*, *produjese*, *produjeseis*, *produjésemos*, *produjesen*, *produjeses*, *produjimos*, *produjiste*, *produjisteis*, *produjo*, *produzca*, *produzcáis*, *produzcamos*, *produzcan*, *produzcas*, *produzcás*, *produzco*},

{*queráis*, *queramos*, *querás*, *queré*, *quered*, *queréis*, *queremos*, *querer*, *querés*, *quería*, *queríais*, *queríamos*, *querían*, *querías*, *querida*, *queridas*, *querido*, *queridos*, *queriendo*, *querrá*, *querrán*, *querrás*, *querré*, *querréis*, *querremos*, *querría*, *querríais*, *querríamos*, *querrían*, *querrías*, *quiera*, *quieran*, *quieras*, *quiere*, *quieren*, *quieres*, *quiero*, *quise*, *quisiera*, *quisierais*, *quisiéramos*, *quisieran*, *quisieras*, *quisiere*, *quisiereis*, *quisiéremos*, *quisieren*, *quisieres*, *quisieron*, *quisiese*, *quisieseis*, *quisiésemos*, *quisiesen*, *quisieses*, *quisimos*, *quisiste*, *quisisteis*, *quiso*},

{*rehuí*, *rehuía*, *rehuíais*, *rehuíamos*, *rehuían*, *rehuías*, *rehuid*, *rehuida*, *rehuidas*, *rehuido*, *rehuidos*, *rehuimos*, *rehuir*, *rehuirá*, *rehuirán*, *rehuirás*, *rehuiré*, *rehuiréis*, *rehuiremos*, *rehuiría*, *rehuiríais*, *rehuiríamos*, *rehuirían*, *rehuirías*, *rehuís*, *rehuiste*, *rehuisteis*, *rehúya*, *rehuyáis*, *rehuyamos*, *rehúyan*, *rehúyas*, *rehúye*, *rehúyen*, *rehuyendo*, *rehuyera*, *rehuyerais*, *rehuyéramos*, *rehuyeran*, *rehuyeras*, *rehuyere*, *rehuyereis*, *rehuyéremos*, *rehuyeren*, *rehuyeres*, *rehuyeron*, *rehúyes*, *rehuyese*, *rehuyeseis*, *rehuyésemos*, *rehuyesen*, *rehuyeses*, *rehuyó*, *rehúyo*},

{*repón*, *repondrá*, *repondrán*, *repondrás*, *repondré*, *repondréis*, *repondremos*, *repondría*, *repondríais*, *repondríamos*, *repondrían*, *repondrías*, *repone*, *reponed*, *reponéis*, *reponemos*, *reponen*, *reponer*, *repones*, *reponga*, *repongáis*, *repongamos*, *repongan*, *repongas*, *repongo*, *reponía*, *reponíais*, *reponíamos*, *reponían*, *reponías*, *reponiendo*, *repuesta*, *repuestas*, *repuesto*, *repuestos*, *repuse*, *repusiera*, *repusierais*, *repusiéramos*, *repusieran*, *repusieras*, *repusiere*, *repusiereis*, *repusiéremos*, *repusieren*, *repusieres*, *repusieron*, *repusiese*, *repusieseis*, *repusiésemos*, *repusiesen*, *repusieses*, *repusimos*, *repusiste*, *repusisteis*, *repuso*},

{*retén*, *retendrá*, *retendrán*, *retendrás*, *retendré*, *retendréis*, *retendremos*, *retendría*, *retendríais*, *retendríamos*, *retendrían*, *retendrías*, *retened*, *retenéis*, *retenemos*, *retener*, *retenga*, *retengáis*, *retengamos*, *retengan*, *retengas*, *retengo*, *retenía*, *reteníais*, *reteníamos*, *retenían*, *retenías*, *retenida*, *retenidas*, *retenido*, *retenidos*, *reteniendo*, *retiene*, *retienen*, *retienes*, *retuve*, *retuviera*, *retuvierais*, *retuviéramos*, *retuvieran*, *retuvieras*, *retuviere*, *retuviereis*, *retuviéremos*, *retuvieren*, *retuvieres*, *retuvieron*, *retuviese*, *retuvieseis*, *retuviésemos*, *retuviesen*, *retuvieses*, *retuvimos*, *retuviste*, *retuvisteis*, *retuvo*},

{*sal*, *saldrá*, *saldrán*, *saldrás*, *saldré*, *saldréis*, *saldremos*, *saldría*, *saldríais*, *saldríamos*, *saldrían*, *saldrías*, *sale*, *salen*, *sales*, *salga*, *salgáis*, *salgamos*, *salgan*, *salgas*, *salgo*, *salí*, *salía*, *salíais*, *salíamos*, *salían*, *salías*, *salid*, *salida*, *salidas*, *salido*, *salidos*, *saliendo*, *saliera*, *salierais*, *saliéramos*, *salieran*, *salieras*, *saliere*, *saliereis*, *saliéremos*, *salieren*, *salieres*, *salieron*, *saliese*, *salieseis*, *saliésemos*, *saliesen*, *salieses*, *salimos*, *salió*, *salir*, *salís*, *saliste*, *salisteis*},

{*trae*, *traed*, *traéis*, *traemos*, *traen*, *traer*, *traerá*, *traerán*, *traerás*, *traeré*, *traeréis*, *traeremos*, *traería*, *traeríais*, *traeríamos*, *traerían*, *traerías*, *traes*, *traía*, *traíais*, *traíamos*, *traían*, *traías*, *traída*, *traídas*, *traído*, *traídos*, *traiga*, *traigáis*, *traigamos*, *traigan*, *traigas*, *traigo*, *traje*, *trajera*, *trajerais*, *trajéramos*, *trajeran*, *trajeras*, *trajere*, *trajereis*, *trajéremos*, *trajeren*, *trajeres*, *trajeron*, *trajese*, *trajeseis*, *trajésemos*, *trajesen*, *trajeses*, *trajimos*, *trajiste*, *trajisteis*, *trajo*, *trayendo*},

{*ve*},

{*vea*, *veáis*, *veamos*, *vean*, *veas*, *ved*, *veía*, *veíais*, *veíamos*, *veían*, *veías*, *veis*, *vemos*, *ven*, *veo*, *ver*, *verá*, *verán*, *verás*, *veré*, *veréis*, *veremos*, *vería*, *veríais*, *veríamos*, *verían*, *verías*, *ves*, *vi*, *viendo*, *viera*, *vierais*, *viéramos*, *vieran*, *vieras*, *viere*, *viereis*, *viéremos*, *vieren*, *vieres*, *vieron*, *viese*, *vieseis*, *viésemos*, *viesen*, *vieses*, *vimos*, *vio*, *vista*, *vistas*, *viste*, *visteis*, *visto*, *vistos*},

{*yace*, *yaced*, *yacéis*, *yacemos*, *yacen*, *yacer*, *yacerá*, *yacerán*, *yacerás*, *yaceré*, *yaceréis*, *yaceremos*, *yacería*, *yaceríais*, *yaceríamos*, *yacerían*, *yacerías*, *yaces*, *yací*, *yacía*, *yacíais*, *yacíamos*, *yacían*, *yacías*, *yacida*, *yacidas*, *yacido*, *yacidos*, *yaciendo*, *yaciera*, *yacierais*, *yaciéramos*, *yacieran*, *yacieras*, *yaciere*, *yaciereis*, *yaciéremos*, *yacieren*, *yacieres*, *yacieron*, *yaciese*, *yacieseis*, *yaciésemos*, *yaciesen*, *yacieses*, *yacimos*, *yació*, *yaciste*, *yacisteis*, *yaga*, *yagáis*, *yagamos*, *yagan*, *yagas*, *yago*, *yaz*, *yazca*, *yazcáis*, *yazcamos*, *yazcan*, *yazcas*, *yazco*, *yazga*, *yazgáis*, *yazgamos*, *yazgan*, *yazgas*, *yazgo*}

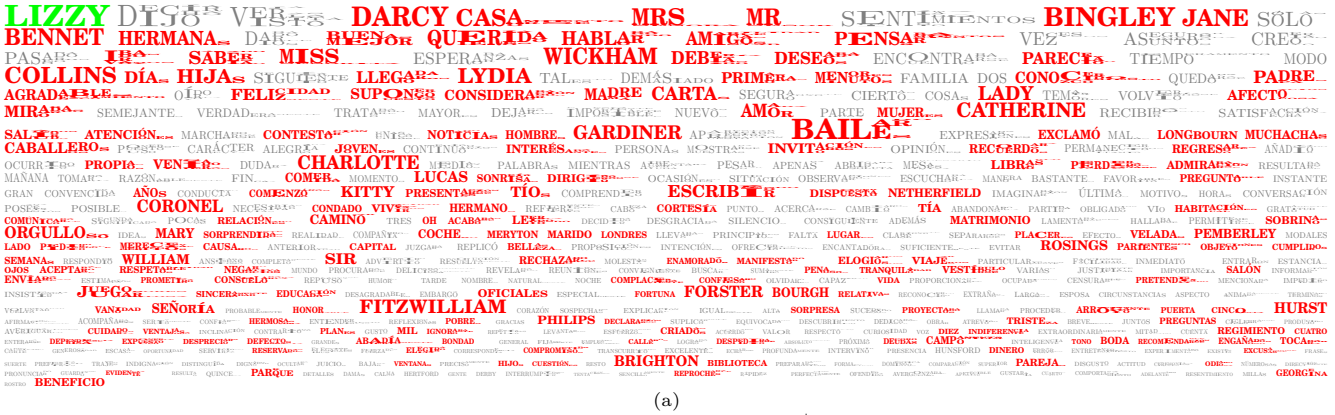
as output. Here, the ambiguous words *di* and *ve* become orphans in our algorithm: *di* is a conjugated form of both *dar* “give” and *decir* “say”; *ve* is a conjugated form of both *ir* “go” and *ver* “see”.

Example 6.11.4. In Fig. S7, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).

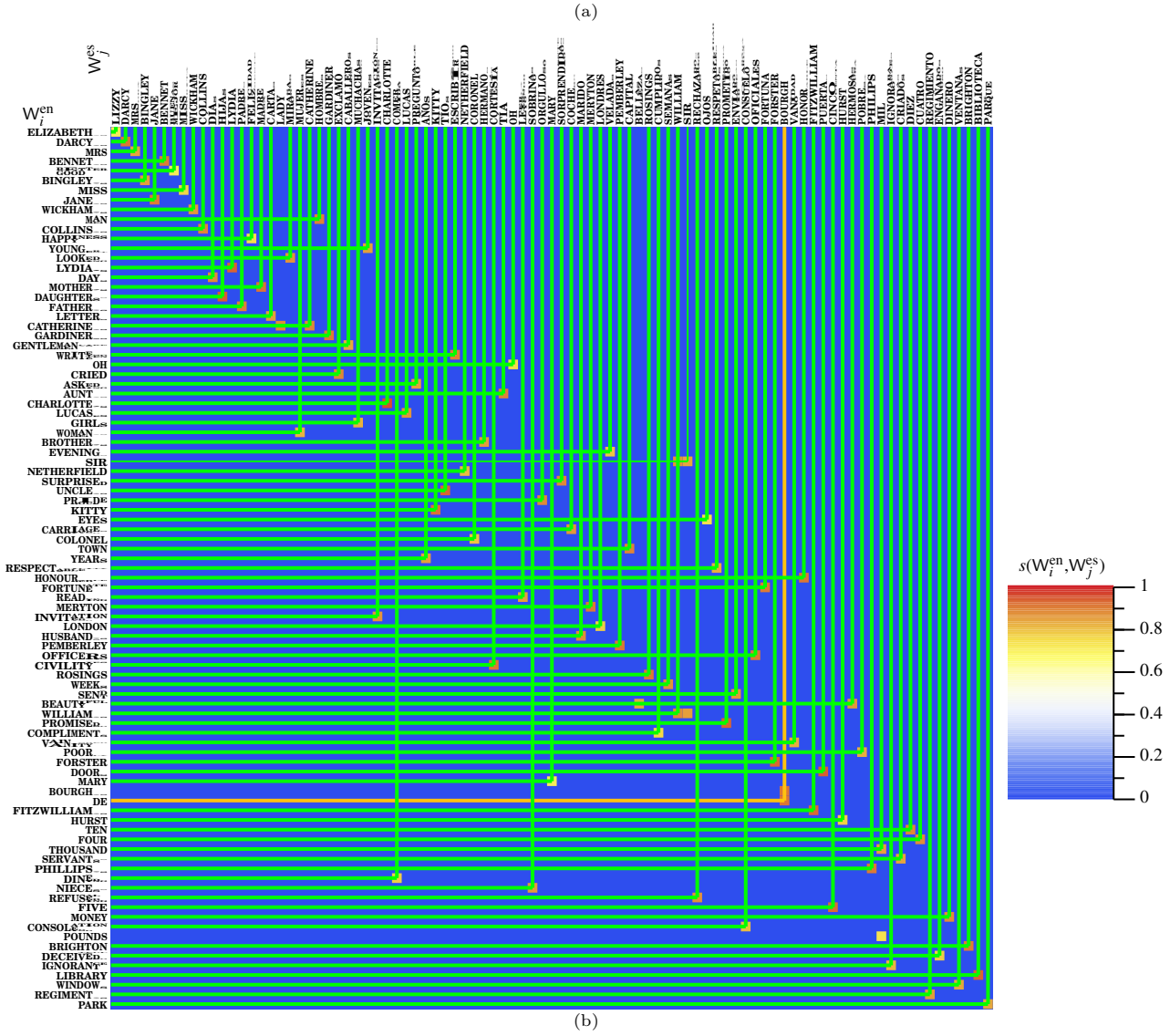
Note that the machine translation task is particularly challenging in the example shown, for at least two reasons: (1) The translator has merged the name *Elizabeth* and the nickname *Lizzy* of the protagonist into one; (2) The Spanish noun *casa* “house” is spelt the same way as a conjugated form of the Spanish verb *casar* “marry”. Despite these disadvantages, the topics that are unrelated to these challenges are correctly translated.

In Fig. S7b, we consider the Spanish word *capital* “capital (city)” an exact match to the English word *town*, because the latter refers to London in *Pride and Prejudice*.

In Spanish, both *bella* and *hermosa* mean “beautiful” and modify feminine nouns. It thus comes to no surprise that the row for “beautiful” in Fig. S7b contains two hot spots that correspond to these two Spanish adjectives.



(a)



(b)

Fig. S7. Text mining in Spanish. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Spanish version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{es})$ between selected topics in English and Spanish versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *amber*) cross-hair indicates an exact (resp. a close but non-exact) match.

6.2 Modified Porter stemming algorithm for French

The French verbs for "to be" and "to have" are highly irregular. These verbs are also used as auxiliaries in various tenses and moods of arbitrary verbs, similar to their counterparts in English. All their conjugated forms (including the infinitive forms) are treated as stop words.

> *a*, *ai*, *aie*, *aient*, *aies*, *ait*, *as*, *aura*, *aurai*, *auraient*, *aurais*, *aurait*, *auras*, *aurez*, *auriez*, *aurions*, *aurons*, *auront*, *avaient*, *avais*, *avait*, *avez*, *aviez*, *avions*, *avoir*, *avons*, *ayant*, *ayez*, *ayons*, *eu*, *eûmes*, *eurent*, *eus*, *eusse*, *eussent*, *eusses*, *eussiez*, *eussions*, *eut*, *eût*, *eûtes*, *ont* — "have" ;
>
> *es*, *est*, *étaient*, *étais*, *était*, *étant*, *été*, *êtes*, *étiez*, *étions*, *être*, *fûmes*, *furent*, *fus*, *fusse*, *fussent*, *fusses*, *fussiez*, *fussions*, *fut*, *fût*, *fûtes*, *sera*, *serai*, *seraient*, *serais*, *serait*, *seras*, *serez*, *seriez*, *serions*, *serons*, *seront*, *soient*, *sois*, *soit*, *sommes*, *sont*, *soyez*, *soyons*, *suis* — "be".

In addition, we treat the following verbs as stop words.

> *faire*, *fais*, *faisaient*, *faisais*, *faisait*, *faisant*, *faisiez*, *faisions*, *faisons*, *fait*, *faites*, *fasse*, *fassent*, *fasses*, *fassiez*, *fassions*, *fera*, *ferai*, *feraient*, *ferais*, *ferait*, *feras*, *ferez*, *feriez*, *ferions*, *ferons*, *feront*, *fîmes*, *firent*, *fis*, *fisse*, *fissent*, *fisses*, *fissiez*, *fissions*, *fit*, *fît*, *fîtes*, *font* — "do" ;
>
> *fallait*, *falloir*, *fallu*, *fallut*, *faudra*, *faudrait*, *faut* — "be necessary" ;
>
> *peut*, *peuvent*, *peux*, *pourra*, *pourrai*, *pourraient*, *pourrais*, *pourrait*, *pourras*, *pourrez*, *pourriez*, *pourrions*, *pourrons*, *pourront*, *pouvaient*, *pouvais*, *pouvait*, *pouvant*, *pouvez*, *pouviez*, *pouvions*, *pouvoir*, *pouvons*, *pu*, *puisse*, *puissent*, *puisses*, *puissiez*, *puissions*, *pûmes*, *purent*, *pus*, *pusse*, *pussent*, *pusses*, *pussiez*, *pussions*, *put*, *pût*, *pûtes* — "can".

Definition 6.12 (French stop words)**.** If a word belongs to the following list[73]:

> *a*, *à*, *afin*, *ai*, *aie*, *aient*, *aies*, *ainsi*, *ait*, *alors*, *après*, *arrière*, *as*, *assez*, *au*, *aucun*, *aucune*, *aucunes*, *aucuns*, *aura*, *aurai*, *auraient*, *aurais*, *aurait*, *auras*, *aurez*, *auriez*, *aurions*, *aurons*, *auront*, *aussi*, *autant*, *autour*, *autre*, *autrefois*, *autres*, *aux*, *avaient*, *avais*, *avait*, *avant*, *avec*, *avez*, *aviez*, *avions*, *avoir*, *avons*, *ayant*, *ayez*, *ayons*, *beaucoup*, *bien*, *bientôt*, *c*, *ça*, *çà*, *car*, *ce*, *ceci*, *cela*, *celà*, *celle*, *celles*, *celui*, *cependant*, *ces*, *cet*, *cette*, *ceux*, *chacun*, *chacune*, *chaque*, *chez*, *ci*, *combien*, *comme*, *comment*, *contre*, *d*, *dans*, *de*, *dedans*, *dehors*, *déjà*, *depuis*, *dernier*, *dernière*, *dernières*, *derniers*, *derrière*, *des*, *dès*, *dessous*, *dessus*, *devaient*, *devais*, *devait*, *devant*, *devez*, *deviez*, *devions*, *devoir*, *devons*, *devra*, *devrai*, *devraient*, *devrais*, *devrait*, *devras*, *devrez*, *devriez*, *devrions*, *devrons*, *devront*, *dois*, *doit*, *doive*, *doivent*, *doives*, *donc*, *dont*, *du*, *dû*, *dûmes*, *durant*, *durent*, *dus*, *dusse*, *dussent*, *dusses*, *dussiez*, *dussions*, *dut*, *dût*, *dûtes*, *elle*, *elles*, *en*, *encore*, *envers*, *es*, *est*, *et*, *étaient*, *étais*, *était*, *étant*, *été*, *êtes*, *étiez*, *étions*, *être*, *eu*, *eue*, *eues*, *eûmes*, *eurent*, *eus*, *eusse*, *eussent*, *eusses*, *eussiez*, *eussions*, *eut*, *eût*, *eûtes*, *eux*, *faire*, *fais*, *faisaient*, *faisais*, *faisait*, *faisant*, *faisiez*, *faisions*, *faisons*, *fait*, *faite*, *faites*, *faits*, *fallait*, *falloir*, *fallu*, *fallut*, *fasse*, *fassent*, *fasses*, *fassiez*, *fassions*, *faudra*, *faudrait*, *faut*, *fera*, *ferai*, *feraient*, *ferais*, *ferait*, *feras*, *ferez*, *feriez*, *ferions*, *ferons*, *feront*, *fîmes*, *firent*, *fis*, *fisse*, *fissent*, *fisses*, *fissiez*, *fissions*, *fit*, *fît*, *fîtes*, *fois*, *font*, *fûmes*, *furent*, *fus*, *fusse*, *fussent*, *fusses*, *fussiez*, *fussions*, *fut*, *fût*, *fûtes*, *hors*, *ici*, *il*, *ils*, *j*, *jamais*, *je*, *jusqu*, *jusque*, *juste*, *justement*, *l*, *la*, *là*, *laquelle*, *le*, *lequel*, *les*, *lesquelles*, *lesquels*, *leur*, *leurs*, *lors*, *lorsqu*, *lorsque*, *lui*, *m*, *ma*, *maintenant*, *mais*, *malgré*, *me*, *même*, *mêmes*, *mes*, *mien*, *mienne*, *miennes*, *miens*, *moi*, *moindre*, *moins*, *mon*, *n*, *ne*, *ni*, *non*, *nos*, *notre*, *nôtre*, *nôtres*, *nous*, *nul*, *nulle*, *nulles*, *nuls*, *on*, *ont*, *ou*, *où*, *oui*, *par*, *parce*, *parfois*, *parmi*, *partout*, *pas*, *pendant*, *personne*, *peu*, *peut*, *peuvent*, *peux*, *plupart*, *plus*, *plusieurs*, *plutôt*, *pour*, *pourquoi*, *pourra*, *pourrai*, *pourraient*, *pourrais*, *pourrait*, *pourras*, *pourrez*, *pourriez*, *pourrions*, *pourrons*, *pourront*, *pourtant*, *pouvaient*, *pouvais*, *pouvait*, *pouvant*, *pouvez*, *pouviez*, *pouvions*, *pouvoir*, *pouvons*, *presqu*, *presque*, *prochain*, *prochaine*, *prochaines*, *prochains*, *propos*, *pu*, *puis*, *puisque*, *puisse*, *puissent*, *puisses*, *puissiez*, *puissions*, *pûmes*, *purent*, *pus*, *pusse*, *pussent*, *pusses*, *pussiez*, *pussions*, *put*, *pût*, *pûtes*, *qu*, *quand*, *quant*, *que*, *quel*, *quelconque*, *quelle*, *quelles*, *quelqu*, *quelque*, *quelquefois*, *quelques*, *quels*, *qui*, *quiconque*, *quoi*, *quoique*, *rien*, *s*, *sa*, *sans*, *sauf*, *se*, *selon*, *sera*, *serai*, *seraient*, *serais*, *serait*, *seras*, *serez*, *seriez*, *serions*, *serons*, *seront*, *ses*, *si*, *sien*, *sienne*, *siennes*, *siens*, *sitôt*, *soi*, *soient*, *sois*, *soit*, *sommes*, *son*, *sont*, *sous*, *souvent*, *soyez*, *soyons*, *suis*, *sur*, *surtout*, *t*, *ta*, *tandis*, *tant*, *tantôt*, *te*, *tel*, *telle*, *tellement*, *telles*, *tels*, *tes*, *tien*, *tienne*, *tiennes*, *tiens*, *toi*, *ton*, *toujours*, *tous*, *tout*, *toute*, *toutefois*, *toutes*, *travers*, *très*, *trop*, *tu*, *un*, *une*, *unes*, *uns*, *vers*, *voici*, *voilà*, *vos*, *votre*, *vôtre*, *vôtres*, *vous*, *y*,

then we consider it a French stop word. All the French stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

[73] Our list of French stop words is based on `snowball.tartarus.org/algorithms/french/stop.txt`, with extensive modifications that roughly match their counterparts in English.

The French verbs for "to come" and "to go" also have highly irregular conjugations:[74]

> *aille*, *aillent*, *ailles*, *alla*, *allai*, *allaient*, *allais*, *allait*, *allâmes*, *allant*, *allas*, *allasse*, *allassent*, *allasses*, *allassiez*, *allassions*, *allât*, *allâtes*, *allé*, *allés*, *aller*, *allèrent*, *allez*, *alliez*, *allions*, *allons*, *ira*, *irai*, *iraient*, *irais*, *irait*, *iras*, *irez*, *iriez*, *irions*, *irons*, *iront*, *va*, *vais*, *vas*, *vont* — "go" ;
>
> *venaient*, *venais*, *venait*, *venant*, *venez*, *veniez*, *venions*, *venir*, *venons*, *venu*, *venus*, *viendra*, *viendrai*, *viendraient*, *viendrais*, *viendrait*, *viendras*, *viendrez*, *viendriez*, *viendrions*, *viendrons*, *viendront*, *vienne*, *viennent*, *viennes*, *viens*, *vient*, *vînmes*, *vinrent*, *vins*, *vinsse*, *vinssent*, *vinsses*, *vinssiez*, *vinssions*, *vint*, *vînt*, *vîntes* — "come".

We are not going to treat them as stop words in French, so as to be consistent with their counterparts in other languages. However, to prepare for clustering of content words, we will refer to these two groups of strings as **goFrench** and **comeFrench**, respectively.

6.2.1 Effective spelling and essential root

Definition 6.13 (French Vowel Extensions)**.** Hereafter in §6.2, the symbol $\mathbf{V}^*$ stands for any member from the list {*a*, *â*, *e*, *ê*, *i*, *î*, *o*, *u*, *û*}, the so-called French vowel extensions. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}^*_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of French vowel extensions.

Dual to the notations above, the symbol $\mathbf{C}^*$ stands for any character that does not belong to the list {*a*, *â*, *e*, *ê*, *i*, *î*, *o*, *u*, *û*}, and $\mathbf{C}^*_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list. □

Algorithm 6.14 (French effective spelling)**.** *For a French word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in six sequential steps:*

(1) *Replace*[75]

ab$\mathbf{C}^*$~	*amabilité~*	*chance~*	*charlotte*	*déf*$\mathbf{X}$~	*démi~*	*détect*$\mathbf{X}$~	*détei~*
bαa$\mathbf{C}^*$	*ami*	*βance*	*chaρλωτ*	*δϕ*	*dεmi*	*τδεκτ*	*τβtei*

détend~	*détest*$\mathbf{X}$~	*diab*$(\varnothing \mid o)$*l~*	*discour~*	*foyer~*	*frém~*	*guet~*	*guid~*	*imposs~*	*inct*$\mathbf{X}$	*lisièr~*
ρλaξ	*ηατε*	*διαβλ*	*δσκour*	*ϕoκaρe*	*ϕρm*	*γuεt*	*γuid*	*μροσ*	*ingue*	*liσερ*

list~	*litan~*	*litt~*	*livr*$\mathbf{X}$~	*louis~*	*message~*	*muet~*	*muf~*	*musi~*	*mutu~*	*ndité*$\mathbf{X}$	*nom~*	*ô~*	*partag~*
λiστ	*λiταν̃*	*liττ*	*λiβρ*	*λuiσ*	*μεßage*	*μuwτt*	*mf*	*μuσi*	*mτu*	*nd*	*noμ*	*wô*	*paρτag*

présen$(c \mid t)$*~*	*quotidien*$\mathbf{X}$	*récemment*	*régim~*	*région~*	*remerc~*	*rou*$(sse \mid x)$*~*	*salad~*	*sér~*
prseν̃τ	*jour*	*récent*	*ρεγμ*	*εργν̃*	*ρμερκ*	*rouge*	*saλδ*	*ρσεσ*

sommet~	*tableau*	*taill*	*témér~*	*term~*	*triste~*	*trois~*	*uni*$(cit \mid q)$	*volt~*	*volu~*
σoμμεt	*ταβλεau*	*tail*	*τεμερ*	*terμ*	*triστε*	*3trois*	*uν̃iκ*	*βoλτ*	*ωvol*

$\underline{(\varnothing \mid ma \mid mes)dame(\varnothing \mid s)}$	$\underline{(\varnothing \mid ma \mid mes)demoiselle(\varnothing \mid s)}$	$\underline{bessie}$	$\underline{hui}$	$\underline{joyau(\varnothing \mid x)}$	$\underline{mlle(\varnothing \mid s)}$	$\underline{mme(\varnothing \mid s)}$
lady	*miß*	*βεσθie*	*αujourδ*	*joaillerie*	*miß*	*lady*

$\underline{nu(\varnothing \mid e)(\varnothing \mid s)}$	$\underline{rite(\varnothing \mid s)}$	$\underline{vie(\varnothing \mid s)}$	$\sim^{\mathbf{X}\in}(\mathbf{V}^*\mathbf{C}^*_{m_0})ement(\varnothing \mid s)$	$\sim an(ce \mid t)$
ν̃uδε	*ριτ*	*vivre*	$\mathbf{X}u$	*ants*

(2) *Replace*

$(maman \mid matern \mid mère)\mathbf{X}\sim$	$(meilleur \mid mieux)\sim$	$\mathbf{C}^*y$
μαma	*μeilleur*	$\mathbf{C}^*\eta$

$^{\mathbf{X}\in}(aper \mid con \mid per)cept\mathbf{X}\sim$	$a^{\mathbf{X}\in}(cc \mid ss \mid tt)\sim$	*accrû*$\mathbf{X}$~	*accus~*	*ad~*	*aff~*	*aim~*	*air~*	*anv*	*avunculair*$\mathbf{X}$~
$\mathbf{X}cev$	$x\mathbf{X}^{[1]}$	*xkroissons*	*accuσ*	*zd*	*φ*	*amour*	*αir*	*atv*	*oncle*

$bon(\varnothing \mid ne)(\varnothing \mid s)\sim$	*cess~*	$choi(s \mid x)\sim$	$concl(u \mid û)\sim$	*courage~*	*ction*	*dame~*	*dat~*	*décept*$\mathbf{X}$~	*déte~*	*dét~*
μeilleur	*ξesσ*	*choiξ*	*qonclu*	*κourage*	*ctive*	*δaμe*	*daτ*	*decev*	*δτ*	*dεt*

[74] We have chosen not to include the feminine forms of past participles (*allée*, *allées*, *venue*, *venues*) in the lists below, because their functions as nouns deviate from the original verbs. Similarly, when the present participles *allant* and *venant* are used as adjectives, their meanings no longer align perfectly with the respective verbs. Therefore, we do not include the feminine and plural forms of these adjectives (such as *allante* and *allants*) in our lists of verb forms.

[75] To avoid confusion of Greek nu with Latin vee, we write the former as ν̃ in substitution rules.

différend~	*difficulté~*	*distant~*	*dort~*	*éclor*	*eill*	*excl(u|û)~*	*faço~*	*fratern***X~**	*front~*	*gén~*
different	*difficile*	*distance*	*dorm*	*eclos*	*eil*	*qwexclu*	*φaço*	*frère*	*frontτ*	*gzan*

hiv~	*hommage~*	*incl(u|û)~*	*ivité*	*k***C***	*matin~*	*matrimon***X~**	*mer~*	*mesu~*
hib	*ηommage*	*winclu*	*ive*	*κ***C***	*ματιῖ̃*	*marier*	*mɛr*	*meσw*

meur(ent|s)	*misér~*	*mois*	*moment~*	*mu***V***~	*muet~*	*mul~*	*mur~*	*occ~*	*op~*	*parent~*	*patern***X~**
mort	*μiσer*	*μoiσ*	*μomeῖ̃τ*	*μwu*	*muɛτ*	*μuλ*	*myρ*	*cc*	*zp*	*pareῖ̃τ*	*père*

petit~	*peur~*	*pré*	*prés~*	*qua(d|t)r*	*querell~*	*quest~*	*rd*	*rdin*	*re***C***~
peτiτ	*πeuρ*	*prae*	*rπeσ*	*4*	*qreλ*	*qqeστ*	*ρδ*	*ρρδiῖ̃*	*r***C***

réceptX·	*rey~*	*rn*	*rst*	*saisi~*	*salu~*	*sc*	*sit~*	*sororX·*	*spoir*	*squ*	*tôt*	*trava~*	*vers~*	*virt~*	*vite~*	*vitr~*
rcefer	*ρey*	*rῖ̃*	*ρστ*	*saiσi*	*σalu*	*σc*	*siτ*	*sœur*	*sper*	*σque*	*τoτ*	*τrava*	*βerσ*	*zyrt*	*φviτe*	*viτρ*

voiture	*voyag*	*vulg~*	*(messieurs|monsieur|mr)*	*(mistriss|mme|mrs)*
voiτure	*βoηaγ*	*βulγ*	*umur*	*μadame*

X$^{\in}$*(∅|con|de|re)***comeFrench**	*ans*	*bal(∅|s)*	*courent*	*dors*	*écri(***V***|è|é)**X***	*filleul(∅|s)*	*gens*
X*fcomeφ****X****i*	*an*	*βal*	*courez*	*dorm*	*fjeλ*	*fils*	*γenσ*

goFrench	*heure(∅|s)*	*mari(∅|s)*	*maria*	*mot(∅|s)*	*né(∅|e)(∅|s)*	*nez*	*papa(∅|s)*	*parc(∅|s)*	*vieux*
fgoφ	*zheure*	*μari*	*μaria*	*μoτ*	*naître*	*ṽɛζ*	*père*	*pparc*	*vieil*

voit	*yeux*	*~x̂∉(e)aux*	*~(îmes|îtes)*
verra	*œil*	*x̂al*	*i*

*~***X**$^{\in}$*(***V****C*$^{*}_{m_0}$*)ab(ilité|le)(∅|s*	*~eau(∅|té|tés|x)*	*~ial(∅|e|ité)(∅|s)*	*~iaux*	*~k*	*~m(î|i)t*	*~ss*	*~trice(∅|s)*	*~urent*	*~y*
X*u*	*eler*	*e*	*e*	*κ*	*mis*	*β*	*teur*	*u*	*η*

(3) *Do (é\|è\|ë) → e, (i\|y) → iî, (k\|q)~ → κ,* **X**$^{\in}$*(a\|o\|u)r →* **X***ρ, ç → ce, dis***C***~ *→ dss***C***, *gn → nad, jourṽée~ → jouρ, naîtr***X**~ *→ naquit, ô → os, s*$^{\mathbf{X}\in}$*(in\|ui) →β***X**, *sûr →βur, ü → u, û → uû, myρent → mouvez, ~au(∅\|x) → al, ~c → k, ~eux → eu, i*$^{\mathbf{X}\in}$*(me\|~t) → i***X**, *~in(s\|t) → in;*

(4) *Replace*

n(d|t)r	*x̂v*	*ρt*	**C****nee*	*c*$^{\mathbf{X}\in}$*(a|o|u|***C****)*	*dsss~*	*entier~*	*entrev~*	*g*$^{\mathbf{X}\in}$*(e|i)*	*qu*	*ui*	*vieil***X**	*~(ez|nt)*	*~euρ(∅|s)*
n	*x̂f*	*ρτ*	**C***	*k***X**	*diss*	*εṽtier*	*eṽτρev*	*j***X**	*k*	*uy*	*oλδ*	*i*	*euse*

(5) *Do ell → el, enn → en, ett → et,* **V**$^{\in}$*(a\|e\|o)i →* **V***y,* **V**$^{\in}$*(a\|e\|o)u →* **V***w.*

(6) *Do* **X**$^{\in}$*(m\|p\|v)ew~ →* **X***ow, kow~ → kkow, kru~ → kroy, mu~ → mowf, v(iî\|irey\|it\|u)~ → voyre, bow(s\|t) → bowyllons, vawt → vawdrons.*

(7) *Do* **X**$^{\in}$*(τrafa\|konse)yl(l\|s)~ →* **X***yl, mowf***X**~ *→ mowf,* **X**$^{\in}$*(p\|s)u(rey\|s\|t) →* **X***uss, mow(s\|t) → mowf, ~ewse(∅\|s) → e.*

Definition 6.15 (French protected range)**.** Let $\hat{\sigma}$ be a text string derived from a French word, its protected range ProtRg($\hat{\sigma}$) is an integer determined as follows:

- Try to find the string pattern (∅|*a*|*d*(*a*|*e*|*ia*|*if*|*ε*)|*e*|*fi*|*i*|*ko*|*mo*|*ni*|*ob*|*par*|*pr*(*ae*|*o*)|*r*(*a*|*e*)|*siî*|*sou*|*su*(*b*|*p*|*r*)|*ter*|*tran*|*x*(∅|*k*)) **C**$^{*}_{m_0}$**V*****C**$^{*}_{m_0}$~ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines ProtRg($\hat{\sigma}$); otherwise, set ProtRg($\hat{\sigma}$) = 0. □

Algorithm 6.16 (French essential root)**.** *Let $\hat{\sigma}$ be the effective spelling of a French word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Break down $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ into the concatenation of two strings $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ (see the notation in Definition 3.1) and $\hat{\sigma}_2$, where the length of the first string $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ is equal to the protected range of $\hat{\sigma}$.*

(2) *On $\hat{\sigma}_2$, perform the following substitutions in a sequel:*

(3.1) *Do ~(on\|s\|t) → ∅.*

(3.2) *Do (b(∅\|i)(∅\|l)\|nt\|~(â\|î\|r\|û)***X**\|st\|~n) → ∅.

(3.3) *Do* $(ion|ke|log|s|tif|y) \to \varnothing$.[76]

(3.4) *Do* $\sim\mathbf{V}_m^* \to \varnothing$.

The result after these four steps of operations is called $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(4) *Perform the following replacements on* $\hat{\sigma}_1\hat{\sigma}'_2$*:*

$mo(\rho\tau\|w\rho)\mathbf{X}\sim$	$mow(dr\|l)\mathbf{X}\sim$	$\underline{dekh(err\|u(\varnothing\|s\|ss\|t))}$		$\underline{di(\varnothing\|r\|s\|ss\|t)}$		$\underline{ekrif}$	$\underline{kkows}$
mowρ	*mowd*	*dekhoyr*		*dict*		*ekris*	*kkowd*
$\underline{lu(\varnothing\|r\|s\|ss\|t)}$	$\underline{p(ow(\varnothing\|f\|rr\|t)\|u(\varnothing\|yss))}$		$\underline{ri}$	$\underline{s(a(f\|kh\|w\rho\|ys\|yt)\|u)}$		$\underline{v(is(\varnothing\|s))(\varnothing\|err)}$	
lis	*puss*		*rir*	*suss*		*voyr*	
$\underline{va(wdr\|yll)}$	$\underline{vow(\varnothing\|dr\|t\|yll)}$			$\underline{xs(e(\varnothing\|y)\|i(\varnothing\|r\|s\|ss\|t)\|oy(\varnothing\|r\|s\|t))}$			
val	*vowl*			*xsied*			
$\sim ce(of\|w(\varnothing\|r\|s\|ss\|t))$	$\sim enferr$	$\sim na(k\|y(\varnothing\|s\|t))$	$\sim pr(i(\varnothing\|r\|s\|ss\|t)\|en)$	$\sim sol(\varnothing\|f)$	$\sim ti(\varnothing\|en)$	$\sim v(en\|i\|ien)$	$\sim vek$
cef	*enfoy*	*nayss*	*prendr*	*sowd*	*tendr*	*vendr*	*vif*

6.2.2 Admissible mutation and approximate clustering

Unlike the three representative Germanic languages and Spanish, we do not need vowel blotting for French, because vowel alternation is not found systematically in French verb conjugations.

In what follows, we will construct a bivariate Boolean-valued function HrdTest($\hat{\alpha},\hat{\beta}$) on a "simple heredity test function" in Algorithm 6.17, and a set of "admissible suffix mismatch" rules in Algorithm 6.18.

Algorithm 6.17 (Simple heredity test)**.** *Let* $\hat{\alpha}'$ *be the result from doing* $\sim(\varnothing|e)s \to \varnothing$ *on* $\hat{\alpha}$*, and define* $\hat{\beta}'$ *similarly. The Boolean-valued function* SimpHrdTest($\hat{\alpha},\hat{\beta}$) *returns* **TRUE** *if the lowercase form of* $\hat{\alpha}$ *contains at least one instance of* $\mathbf{V}^*$ *(Definition 6.13)* **AND** *at least one of the following three conditions holds:*[77]

(i) $\hat{\alpha}' = \hat{\beta}'$*;*

(ii) *After doing* $\sim(d|f|fr|t) \to \varnothing$ *on* $\hat{\beta}$*, we obtain* $\hat{\alpha}$*;*

(iii) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha} = \hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha}')+1\}} = \mathbf{V}^*$*. (See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

In what follows, we define SuffixNW($\hat{\alpha},\hat{\beta}$), RootNW($\hat{\alpha},\hat{\beta}$), NW*($\hat{\alpha},\hat{\beta}$) and SuffixSW($\hat{\alpha},\hat{\beta}$), SuffixSW($\hat{\alpha},\hat{\beta}$), SW*($\hat{\alpha},\hat{\beta}$) as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5.

Algorithm 6.18 (Admissible suffix mismatch)**.** *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function*

$$\text{AdmSM}(\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{SuffixNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if*

$$\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [(\varnothing|\mathbf{V}^*\mathbf{X}r|(m|r|s|t|y)\mathbf{X}|tif\mathbf{X}), (\varnothing|\mathbf{V}^*\mathbf{X}r|(m|r|s|t|y)\mathbf{X}|tif\mathbf{X})]$$

AND NW*($\hat{\alpha},\hat{\beta}$) = $\varnothing$ **AND** RootNW($\hat{\alpha},\hat{\beta}$) *contains at least one instance of* $\mathbf{V}^*$.

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmSM}(\text{RootSW}(\hat{\alpha},\hat{\beta}), \text{SuffixSW}(\hat{\alpha},\hat{\beta}), \text{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 6.19 (Heredity test function)**.** *The structure of the French heredity test function* HrdTest($\hat{\alpha},\hat{\beta}$) *is identical to the German version (Algorithm 8.1.2), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW**,* RootSW*,* SuffixSW*,* SW* *must follow the French rules stated above.*

Algorithm 6.20 (Approximate clustering of French words)**.** *The algorithm is essentially the same as Algorithm 5.10, except that French rules (instead of Danish rules) apply to all the tags (effective spelling, essential root etc.), and vowel blotting is not used.*

Concretely speaking, the approximate clustering of a list of French words $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

[76] We note that longer matches take priority over shorter matches: if *log* is found, then delete these three letters altogether, instead of just a single letter *g*.

[77] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

(1) *We sort the list* $\{(\hat{\alpha}_1, \text{EffSpell}(\hat{\alpha}_1)), \ldots, (\hat{\alpha}_N, \text{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the second component. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \text{EffSpell}(\hat{\alpha}_{(1)})), \ldots, (\hat{\alpha}_{(N)}, \text{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfy* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n)}), \text{EffSpell}(\hat{\alpha}_{(n+1)})) = \textbf{FALSE}$*, where* $n \in \mathbb{Z} \cap [1, N)$*, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \text{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)}, *)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, *), \ldots, (\hat{\alpha}_{(M,n_M)}, *)\}\}$,[78] *where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, *), \ldots, (\hat{\alpha}_{(m,n_m)}, *)\}$ *where* $m \in \mathbb{Z} \cap [1, M]$*, we augment it into a tagged entry* $G_m = (g_m, \text{EffSpell}(\hat{\alpha}_{(m,1)}), \text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with higher priority) and* $\text{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lower priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)})\}$ *satisfy*

$$\text{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$*, then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots, vowel blotted forms) from each* Γ_k*, where* $k \in \mathbb{Z} \cap [1, K]$.

Example 6.20.1. French verbs are divided into three conjugation groups. The first two groups are "regular", while the verbs in the third group are all irregular. Unlike Spanish regular verbs, the root of a verb in the first two conjugations in French sometimes may undergo cosmetic changes for phonological reasons — their counterparts in Spanish would be classified as "irregular", but we adhere to the traditional French grammar (as in `https://en.wiktionary.org/wiki/Appendix:French_verbs`) below, and include them as exceptional cases of the first two "regular" conjugations in French.

We test our algorithm on the following regular French verbs belonging to the first and second conjugation groups:

aima, *aimai*, *aimaient*, *aimais*, *aimait*, *aimâmes*, *aimant*, *aimas*, *aimasse*, *aimassent*, *aimasses*, *aimassiez*, *aimassions*, *aimât*, *aimâtes*, *aime*, *aimé*, *aiment*, *aimer*, *aimera*, *aimerai*, *aimeraient*, *aimerais*, *aimerait*, *aimeras*, *aimèrent*, *aimerez*, *aimeriez*, *aimerions*, *aimerons*, *aimeront*, *aimes*, *aimez*, *aimiez*, *aimions*, *aimons* — "love" ;

appela, *appelai*, *appelaient*, *appelais*, *appelait*, *appelâmes*, *appelant*, *appelas*, *appelasse*, *appelassent*, *appelasses*, *appelassiez*, *appelassions*, *appelât*, *appelâtes*, *appelé*, *appeler*, *appelèrent*, *appelez*, *appeliez*, *appelions*, *appelle*, *appellent*, *appellera*, *appellerai*, *appelleraient*, *appellerais*, *appellerait*, *appelleras*, *appellerez*, *appelleriez*, *appellerions*, *appellerons*, *appelleront*, *appelles*, *appelons* — "call" ;

créa, *créai*, *créaient*, *créais*, *créait*, *créâmes*, *créant*, *créas*, *créasse*, *créassent*, *créasses*, *créassiez*, *créassions*, *créât*, *créâtes*, *crée*, *créé*, *créent*, *créer*, *créera*, *créerai*, *créeraient*, *créerais*, *créerait*, *créeras*, *créèrent*, *créerez*, *créeriez*, *créerions*, *créerons*, *créeront*, *crées*, *créez*, *créiez*, *créions*, *créons* — "create" ;

fini, *finîmes*, *finir*, *finira*, *finirai*, *finiraient*, *finirais*, *finirait*, *finiras*, *finirent*, *finirez*, *finiriez*, *finirions*, *finirons*, *finiront*, *finis*, *finissaient*, *finissais*, *finissait*, *finissant*, *finisse*, *finissent*, *finisses*, *finissez*, *finissiez*, *finissions*, *finissons*, *finit*, *finît*, *finîtes* — "finish" ;

haï, *haïmes*, *haïr*, *haïra*, *haïrai*, *haïraient*, *haïrais*, *haïrait*, *haïras*, *haïrent*, *haïrez*, *haïriez*, *haïrions*, *haïrons*, *haïront*, *hais*, *haïs*, *haïssaient*, *haïssais*, *haïssait*, *haïssant*, *haïsse*, *haïssent*, *haïsses*, *haïssez*, *haïssiez*, *haïssions*, *haïssons*, *hait*, *haït*, *haïtes* — "hate" ;

mange, *mangé*, *mangea*, *mangeai*, *mangeaient*, *mangeais*, *mangeait*, *mangeâmes*, *mangeant*, *mangeas*, *mangeasse*, *mangeassent*, *mangeasses*, *mangeassiez*, *mangeassions*, *mangeât*, *mangeâtes*, *mangent*, *mangeons*, *manger*, *mangera*, *mangerai*, *mangeraient*, *mangerais*, *mangerait*, *mangeras*, *mangèrent*, *mangerez*, *mangeriez*, *mangerions*, *mangerons*, *mangeront*, *manges*, *mangez*, *mangiez*, *mangions* — "eat" ;

noie, *noient*, *noiera*, *noierai*, *noieraient*, *noierais*, *noierait*, *noieras*, *noierez*, *noieriez*, *noierions*, *noierons*, *noieront*, *noies*, *noya*, *noyai*, *noyaient*, *noyais*, *noyait*, *noyâmes*, *noyant*, *noyas*, *noyasse*, *noyassent*, *noyasses*, *noyassiez*, *noyassions*, *noyât*, *noyâtes*, *noyé*, *noyer*, *noyèrent*, *noyez*, *noyiez*, *noyions*, *noyons* — "drown" ;

[78] Here, we use an asterisk (∗) to abbreviate components that are clear from context.

paie, *paient*, *paiera*, *paierai*, *paieraient*, *paierais*, *paierait*, *paieras*, *paierez*, *paieriez*, *paierions*, *paierons*, *paieront*, *paies*, *paya*, *payai*, *payaient*, *payais*, *payait*, *payâmes*, *payant*, *payas*, *payasse*, *payassent*, *payasses*, *payassiez*, *payassions*, *payât*, *payâtes*, *paye*, *payé*, *payent*, *payer*, *payera*, *payerai*, *payeraient*, *payerais*, *payerait*, *payeras*, *payèrent*, *payerez*, *payeriez*, *payerions*, *payerons*, *payeront*, *payes*, *payez*, *payiez*, *payions*, *payons* — "pay" ;

plaça, *plaçai*, *plaçaient*, *plaçais*, *plaçait*, *plaçâmes*, *plaçant*, *plaças*, *plaçasse*, *plaçassent*, *plaçasses*, *plaçassiez*, *plaçassions*, *plaçât*, *plaçâtes*, *place*, *placé*, *placent*, *placer*, *placera*, *placerai*, *placeraient*, *placerais*, *placerait*, *placeras*, *placèrent*, *placerez*, *placeriez*, *placerions*, *placerons*, *placeront*, *places*, *placez*, *placiez*, *placions*, *plaçons* — "place".

and obtain the following results:

{*aima*, *aimai*, *aimaient*, *aimais*, *aimait*, *aimâmes*, *aimant*, *aimas*, *aimasse*, *aimassent*, *aimasses*, *aimassiez*, *aimassions*, *aimât*, *aimâtes*, *aime*, *aimé*, *aiment*, *aimer*, *aimera*, *aimerai*, *aimeraient*, *aimerais*, *aimerait*, *aimeras*, *aimèrent*, *aimerez*, *aimeriez*, *aimerions*, *aimerons*, *aimeront*, *aimes*, *aimez*, *aimiez*, *aimions*, *aimons*},

{*appela*, *appelai*, *appelaient*, *appelais*, *appelait*, *appelâmes*, *appelant*, *appelas*, *appelasse*, *appelassent*, *appelasses*, *appelassiez*, *appelassions*, *appelât*, *appelâtes*, *appelé*, *appeler*, *appelèrent*, *appelez*, *appeliez*, *appelions*, *appelle*, *appellent*, *appellera*, *appellerai*, *appelleraient*, *appellerais*, *appellerait*, *appelleras*, *appellerez*, *appelleriez*, *appellerions*, *appellerons*, *appelleront*, *appelles*, *appelons*},

{*créa*, *créai*, *créaient*, *créais*, *créait*, *créâmes*, *créant*, *créas*, *créasse*, *créassent*, *créasses*, *créassiez*, *créassions*, *créât*, *créâtes*, *crée*, *créé*, *créent*, *créer*, *créera*, *créerai*, *créeraient*, *créerais*, *créerait*, *créeras*, *créèrent*, *créerez*, *créeriez*, *créerions*, *créerons*, *créeront*, *crées*, *créez*, *créiez*, *créions*, *créons*},

{*fini*, *finîmes*, *finir*, *finira*, *finirai*, *finiraient*, *finirais*, *finirait*, *finiras*, *finirent*, *finirez*, *finiriez*, *finirions*, *finirons*, *finiront*, *finis*, *finissaient*, *finissais*, *finissait*, *finissant*, *finisse*, *finissent*, *finisses*, *finissez*, *finissiez*, *finissions*, *finissons*, *finit*, *finît*, *finîtes*},

{*haï*, *haïmes*, *haïr*, *haïra*, *haïrai*, *haïraient*, *haïrais*, *haïrait*, *haïras*, *haïrent*, *haïrez*, *haïriez*, *haïrions*, *haïrons*, *haïront*, *hais*, *haïs*, *haïssaient*, *haïssais*, *haïssait*, *haïssant*, *haïsse*, *haïssent*, *haïsses*, *haïssez*, *haïssiez*, *haïssions*, *haïssons*, *hait*, *haït*, *haïtes*},

{*mange*, *mangé*, *mangea*, *mangeai*, *mangeaient*, *mangeais*, *mangeait*, *mangeâmes*, *mangeant*, *mangeas*, *mangeasse*, *mangeassent*, *mangeasses*, *mangeassiez*, *mangeassions*, *mangeât*, *mangeâtes*, *mangent*, *mangeons*, *manger*, *mangera*, *mangerai*, *mangeraient*, *mangerais*, *mangerait*, *mangeras*, *mangèrent*, *mangerez*, *mangeriez*, *mangerions*, *mangerons*, *mangeront*, *manges*, *mangez*, *mangiez*, *mangions*},

{*noie*, *noient*, *noiera*, *noierai*, *noieraient*, *noierais*, *noierait*, *noieras*, *noierez*, *noieriez*, *noierions*, *noierons*, *noieront*, *noies*, *noya*, *noyai*, *noyaient*, *noyais*, *noyait*, *noyâmes*, *noyant*, *noyas*, *noyasse*, *noyassent*, *noyasses*, *noyassiez*, *noyassions*, *noyât*, *noyâtes*, *noyé*, *noyer*, *noyèrent*, *noyez*, *noyiez*, *noyions*, *noyons*},

{*paie*, *paient*, *paiera*, *paierai*, *paieraient*, *paierais*, *paierait*, *paieras*, *paierez*, *paieriez*, *paierions*, *paierons*, *paieront*, *paies*, *paya*, *payai*, *payaient*, *payais*, *payait*, *payâmes*, *payant*, *payas*, *payasse*, *payassent*, *payasses*, *payassiez*, *payassions*, *payât*, *payâtes*, *paye*, *payé*, *payent*, *payer*, *payera*, *payerai*, *payeraient*, *payerais*, *payerait*, *payeras*, *payèrent*, *payerez*, *payeriez*, *payerions*, *payerons*, *payeront*, *payes*, *payez*, *payiez*, *payions*, *payons*},

{*plaça*, *plaçai*, *plaçaient*, *plaçais*, *plaçait*, *plaçâmes*, *plaçant*, *plaças*, *plaçasse*, *plaçassent*, *plaçasses*, *plaçassiez*, *plaçassions*, *plaçât*, *plaçâtes*, *place*, *placé*, *placent*, *placer*, *placera*, *placerai*, *placeraient*, *placerais*, *placerait*, *placeras*, *placèrent*, *placerez*, *placeriez*, *placerions*, *placerons*, *placeront*, *places*, *placez*, *placiez*, *placions*, *plaçons*}.

Example 6.20.2. Unlike Spanish, French verb conjugations do not exhibit systematic vowel alternations. In fact, the vowels in almost all French verb stems are invariant. There are a few dozen commonly used French verbs that are irregular in their own unique ways. They are treated by dedicated exceptions to rules in our clustering algorithm. Similar to the sampling of Spanish verbs, we ignore cases where the French verb root is different by at most a diacritic mark on the letter *e* (as both *é* and *è* will be converted to *e* in the effective spelling) in conjugated forms.

We use the following representative irregular verbs in French as our input:

accrois, *accroissaient*, *accroissais*, *accroissait*, *accroissant*, *accroisse*, *accroissent*, *accroisses*, *accroissez*, *accroissiez*, *accroissions*, *accroissons*, *accroît*, *accroîtra*, *accroîtrai*, *accroîtraient*, *accroîtrais*, *accroîtrait*, *accroîtras*, *accroître*, *accroîtrez*, *accroîtriez*, *accroîtrions*, *accroîtrons*, *accroîtront*, *accrû*, *accrûmes*, *accrûrent*, *accrûs*, *accrûsse*, *accrûssent*, *accrûsses*, *accrûssiez*, *accrûssions*, *accrût*, *accrûtes* — "increase" ;

acquéraient, *acquérais*, *acquérait*, *acquérant*, *acquérez*, *acquériez*, *acquérions*, *acquérir*, *acquérons*, *acquerra*, *acquerrai*, *acquerraient*, *acquerrais*, *acquerrait*, *acquerras*, *acquerrez*, *acquerriez*, *acquerrions*, *acquerrons*, *acquerront*, *acquière*, *acquièrent*, *acquières*, *acquiers*, *acquiert*, *acquîmes*, *acquirent*, *acquis*, *acquisse*, *acquissent*, *acquisses*, *acquissiez*, *acquissions*, *acquit*, *acquît*, *acquîtes* — "acquire" ;

assaillaient, *assaillais*, *assaillait*, *assaillant*, *assaille*, *assaillé*, *assaillent*, *assailles*, *assaillez*, *assailliez*, *assaillîmes*, *assaillions*, *assaillir*, *assaillira*, *assaillirai*, *assailliraient*, *assaillirais*, *assaillirait*, *assailliras*, *assaillirent*, *assaillirez*, *assailliriez*, *assaillirions*, *assaillirons*, *assailliront*, *assaillis*, *assaillisse*, *assaillissent*, *assaillisses*, *assaillissiez*, *assaillissions*, *assaillit*, *assaillît*, *assaillîtes*, *assaillons* — "assail" ;

asseoir, *asseyaient*, *asseyais*, *asseyait*, *asseyant*, *asseye*, *asseyent*, *asseyes*, *asseyez*, *asseyiez*, *asseyions*, *asseyons*, *assied*, *assieds*, *assiéra*, *assiérai*, *assiéraient*, *assiérais*, *assiérait*, *assiéras*, *assiérez*, *assiériez*, *assiérions*, *assiérons*, *assiéront*, *assîmes*, *assirent*, *assis*, *assisse*, *assissent*, *assisses*, *assissiez*, *assissions*, *assit*, *assît*, *assîtes*, *assoie*, *assoient*, *assoies*, *assoira*, *assoirai*, *assoiraient*, *assoirais*, *assoirait*, *assoiras*, *assoirez*, *assoiriez*, *assoirions*, *assoirons*, *assoiront*, *assois*, *assoit*, *assoyaient*, *assoyais*, *assoyait*, *assoyant*, *assoyez*, *assoyiez*, *assoyions*, *assoyons* — "sit" ;

bat, *bats*, *battaient*, *battais*, *battait*, *battant*, *batte*, *battent*, *battes*, *battez*, *battiez*, *battîmes*, *battions*, *battirent*, *battis*, *battisse*, *battissent*, *battisses*, *battissiez*, *battissions*, *battit*, *battît*, *battîtes*, *battons*, *battra*, *battrai*, *battraient*, *battrais*, *battrait*, *battras*, *battre*, *battrez*, *battriez*, *battrions*, *battrons*, *battront*, *battu* — "beat" ;

bouillaient, *bouillais*, *bouillait*, *bouillant*, *bouille*, *bouillent*, *bouilles*, *bouillez*, *bouilli*, *bouilliez*, *bouillîmes*, *bouillions*, *bouillira*, *bouillirai*, *bouilliraient*, *bouillirais*, *bouillirait*, *bouilliras*, *bouillirent*, *bouillirez*, *bouilliriez*, *bouillirions*, *bouillirons*, *bouilliront*, *bouillis*, *bouillisse*, *bouillissent*, *bouillisses*, *bouillissiez*, *bouillissions*, *bouillit*, *bouillît*, *bouillîtes*, *bouillons*, *bous*, *bout* — "boil" ;

braie, *braient*, *braies*, *braira*, *brairai*, *brairaient*, *brairais*, *brairait*, *brairas*, *braire*, *brairez*, *brairiez*, *brairions*, *brairons*, *brairont*, *brais*, *brait*, *braya*, *brayai*, *brayaient*, *brayais*, *brayait*, *brayâmes*, *brayant*, *brayas*, *brayasse*, *brayassent*, *brayasses*, *brayassiez*, *brayassions*, *brayât*, *brayâtes*, *brayèrent*, *brayez*, *brayiez*, *brayions*, *brayons* — "bray" ;

conclu, *concluaient*, *concluais*, *concluait*, *concluant*, *conclue*, *concluent*, *conclues*, *concluez*, *concluiez*, *concluions*, *conclûmes*, *concluons*, *conclura*, *conclurai*, *concluraient*, *conclurais*, *conclurait*, *concluras*, *conclure*, *conclurent*, *conclurez*, *concluriez*, *conclurions*, *conclurons*, *concluront*, *conclus*, *conclusse*, *conclussent*, *conclusses*, *conclussiez*, *conclussions*, *conclut*, *conclût*, *conclûtes* — "conclude" ;

coud, *coudra*, *coudrai*, *coudraient*, *coudrais*, *coudrait*, *coudras*, *coudre*, *coudrez*, *coudriez*, *coudrions*, *coudrons*, *coudront*, *couds*, *cousaient*, *cousais*, *cousait*, *cousant*, *couse*, *cousent*, *couses*, *cousez*, *cousiez*, *cousîmes*, *cousions*, *cousirent*, *cousis*, *cousisse*, *cousissent*, *cousisses*, *cousissiez*, *cousissions*, *cousit*, *cousît*, *cousîtes*, *cousons*, *cousu* — "sew" ;

couraient, *courais*, *courait*, *courant*, *coure*, *courent*, *coures*, *courez*, *couriez*, *courions*, *courir*, *courons*, *courra*, *courrai*, *courraient*, *courrais*, *courrait*, *courras*, *courrez*, *courriez*, *courrions*, *courrons*, *courront*, *cours*, *court*, *couru*, *courûmes*, *coururent*, *courus*, *courusse*, *courussent*, *courusses*, *courussiez*, *courussions*, *courut*, *courût*, *courûtes* — "run" ;

craignaient, *craignais*, *craignait*, *craignant*, *craigne*, *craignent*, *craignes*, *craignez*, *craigniez*, *craignîmes*, *craignions*, *craignirent*, *craignis*, *craignisse*, *craignissent*, *craignisses*, *craignissiez*, *craignissions*, *craignit*, *craignît*, *craignîtes*, *craignons*, *craindra*, *craindrai*, *craindraient*, *craindrais*, *craindrait*, *craindras*, *craindre*, *craindrez*, *craindriez*, *craindrions*, *craindrons*, *craindront*, *crains*, *craint* — "fear" ;

croie, *croient*, *croies*, *croira*, *croirai*, *croiraient*, *croirais*, *croirait*, *croiras*, *croire*, *croirez*, *croiriez*, *croirions*, *croirons*, *croiront*, *crois*, *croit*, *croyaient*, *croyais*, *croyait*, *croyant*, *croyez*, *croyiez*, *croyions*, *croyons*, *cru*, *crûmes*, *crurent*, *crus*, *crusse*, *crussent*, *crusses*, *crussiez*, *crussions*, *crut*, *crût*, *crûtes* — "believe" ;

cueillaient, *cueillais*, *cueillait*, *cueillant*, *cueille*, *cueillent*, *cueillera*, *cueillerai*, *cueilleraient*, *cueillerais*, *cueillerait*, *cueilleras*, *cueillerez*, *cueilleriez*, *cueillerions*, *cueillerons*, *cueilleront*, *cueilles*, *cueillez*, *cueilli*, *cueilliez*, *cueillîmes*, *cueillions*, *cueillir*, *cueillirent*, *cueillis*, *cueillisse*, *cueillissent*, *cueillisses*, *cueillissiez*, *cueillissions*, *cueillit*, *cueillît*, *cueillîtes*, *cueillons* — "gather" ;

décherra, *décherrai*, *décherraient*, *décherrais*, *décherrait*, *décherras*, *décherrez*, *décherriez*, *décherrions*, *décherrons*, *décherront*, *déchet*, *déchoie*, *déchoient*, *déchoies*, *déchoir*, *déchoira*, *déchoirai*, *déchoiraient*, *déchoirais*, *déchoirait*, *déchoiras*, *déchoirez*, *déchoiriez*, *déchoirions*, *déchoirons*, *déchoiront*, *déchois*, *déchoit*, *déchoyant*, *déchoyez*, *déchoyiez*, *déchoyions*, *déchoyons*, *déchu*, *déchûmes*, *déchurent*, *déchus*, *déchusse*, *déchussent*, *déchusses*, *déchussiez*, *déchussions*, *déchut*, *déchût*, *déchûtes* — “wane” ;

dîmes, *dira*, *dirai*, *diraient*, *dirais*, *dirait*, *diras*, *dire*, *dirent*, *direz*, *diriez*, *dirions*, *dirons*, *diront*, *dis*, *disaient*, *disais*, *disait*, *disant*, *dise*, *disent*, *dises*, *disiez*, *disions*, *disons*, *disse*, *dissent*, *disses*, *dissiez*, *dissions*, *dit*, *dît*, *dites*, *dîtes* — “say” ;

dissolûmes, *dissolurent*, *dissolus*, *dissolut*, *dissolûtes*, *dissolvaient*, *dissolvais*, *dissolvait*, *dissolvant*, *dissolve*, *dissolvent*, *dissolves*, *dissolvez*, *dissolviez*, *dissolvions*, *dissolvons*, *dissoudra*, *dissoudrai*, *dissoudraient*, *dissoudrais*, *dissoudrait*, *dissoudras*, *dissoudre*, *dissoudrez*, *dissoudriez*, *dissoudrions*, *dissoudrons*, *dissoudront*, *dissous*, *dissout* — “dissolve” ;

dormaient, *dormais*, *dormait*, *dormant*, *dorme*, *dorment*, *dormes*, *dormez*, *dormi*, *dormiez*, *dormîmes*, *dormions*, *dormir*, *dormira*, *dormirai*, *dormiraient*, *dormirais*, *dormirait*, *dormiras*, *dormirent*, *dormirez*, *dormiriez*, *dormirions*, *dormirons*, *dormiront*, *dormis*, *dormisse*, *dormissent*, *dormisses*, *dormissiez*, *dormissions*, *dormit*, *dormît*, *dormîtes*, *dormons*, *dors*, *dort* — “sleep” ;

éclora, *éclorai*, *écloraient*, *éclorais*, *éclorait*, *écloras*, *éclore*, *éclorez*, *écloriez*, *éclorions*, *éclorons*, *écloront*, *éclos*, *éclosant*, *éclose*, *éclosent*, *écloses*, *éclosez*, *éclosiez*, *éclosions*, *éclosons*, *éclôt* — “hatch” ;

écrira, *écrirai*, *écriraient*, *écrirais*, *écrirait*, *écriras*, *écrire*, *écrirez*, *écririez*, *écririons*, *écrirons*, *écriront*, *écris*, *écrit*, *écrivaient*, *écrivais*, *écrivait*, *écrivant*, *écrive*, *écrivent*, *écrives*, *écrivez*, *écriviez*, *écrivîmes*, *écrivions*, *écrivirent*, *écrivis*, *écrivisse*, *écrivissent*, *écrivisses*, *écrivissiez*, *écrivissions*, *écrivit*, *écrivît*, *écrivîtes*, *écrivons* — “write” ;

enverra, *enverrai*, *enverraient*, *enverrais*, *enverrait*, *enverras*, *enverrez*, *enverriez*, *enverrions*, *enverrons*, *enverront*, *envoie*, *envoient*, *envoies*, *envoya*, *envoyai*, *envoyaient*, *envoyais*, *envoyait*, *envoyâmes*, *envoyant*, *envoyas*, *envoyasse*, *envoyassent*, *envoyasses*, *envoyassiez*, *envoyassions*, *envoyât*, *envoyâtes*, *envoyé*, *envoyer*, *envoyèrent*, *envoyez*, *envoyiez*, *envoyions*, *envoyons* — “send” ;

fui, *fuie*, *fuient*, *fuies*, *fuîmes*, *fuir*, *fuira*, *fuirai*, *fuiraient*, *fuirais*, *fuirait*, *fuiras*, *fuirent*, *fuirez*, *fuiriez*, *fuirions*, *fuirons*, *fuiront*, *fuis*, *fuisse*, *fuissent*, *fuisses*, *fuissiez*, *fuissions*, *fuit*, *fuît*, *fuîtes*, *fuyaient*, *fuyais*, *fuyait*, *fuyant*, *fuyez*, *fuyiez*, *fuyions*, *fuyons* — “escape” ;

incluaient, *incluais*, *incluait*, *incluant*, *inclue*, *incluent*, *inclues*, *incluez*, *incluiez*, *incluions*, *inclûmes*, *incluons*, *inclura*, *inclurai*, *incluraient*, *inclurais*, *inclurait*, *incluras*, *inclure*, *inclurent*, *inclurez*, *incluriez*, *inclurions*, *inclurons*, *incluront*, *inclus*, *inclusse*, *inclussent*, *inclusses*, *inclussiez*, *inclussions*, *inclut*, *inclût*, *inclûtes* — “include” ;

joignaient, *joignais*, *joignait*, *joignant*, *joigne*, *joignent*, *joignes*, *joignez*, *joigniez*, *joignîmes*, *joignions*, *joignirent*, *joignis*, *joignisse*, *joignissent*, *joignisses*, *joignissiez*, *joignissions*, *joignit*, *joignît*, *joignîtes*, *joignons*, *joindra*, *joindrai*, *joindraient*, *joindrais*, *joindrait*, *joindras*, *joindre*, *joindrez*, *joindriez*, *joindrions*, *joindrons*, *joindront*, *joins*, *joint* — “join” ;

lira, *lirai*, *liraient*, *lirais*, *lirait*, *liras*, *lire*, *lirez*, *liriez*, *lirions*, *lirons*, *liront*, *lis*, *lisaient*, *lisais*, *lisait*, *lisant*, *lise*, *lisent*, *lises*, *lisez*, *lisiez*, *lisions*, *lisons*, *lit*, *lu*, *lûmes*, *lurent*, *lus*, *lusse*, *lussent*, *lusses*, *lussiez*, *lussions*, *lut*, *lût*, *lûtes* — “read” ;

meure, *meurent*, *meures*, *meurs*, *meurt*, *mort*, *mouraient*, *mourais*, *mourait*, *mourant*, *mourez*, *mouriez*, *mourions*, *mourir*, *mourons*, *mourra*, *mourrai*, *mourraient*, *mourrais*, *mourrait*, *mourras*, *mourrez*, *mourriez*, *mourrions*, *mourrons*, *mourront*, *mourûmes*, *moururent*, *mourus*, *mourusse*, *mourussent*, *mourusses*, *mourussiez*, *mourussions*, *mourut*, *mourût*, *mourûtes* — “die” ;

meus, *meut*, *meuve*, *meuvent*, *meuves*, *mouvaient*, *mouvais*, *mouvait*, *mouvant*, *mouvez*, *mouviez*, *mouvions*, *mouvoir*, *mouvons*, *mouvra*, *mouvrai*, *mouvraient*, *mouvrais*, *mouvrait*, *mouvras*, *mouvrez*, *mouvriez*, *mouvrions*, *mouvrons*, *mouvront*, *mû*, *mûmes*, *murent*, *mus*, *musse*, *mussent*, *musses*, *mussiez*, *mussions*, *mut*, *mût*, *mûtes* — “move” ;

moud, *moudra*, *moudrai*, *moudraient*, *moudrais*, *moudrait*, *moudras*, *moudre*, *moudrez*, *moudriez*, *moudrions*, *moudrons*, *moudront*, *mouds*, *moulaient*, *moulais*, *moulait*, *moulant*, *moule*, *moulent*, *moules*, *moulez*, *mouliez*, *moulions*, *moulons*, *moulu*, *moulûmes*, *moulurent*, *moulus*, *moulusse*, *moulussent*, *moulusses*, *moulussiez*, *moulussions*, *moulut*, *moulût*, *moulûtes* — "grind" ;

nais, *naissaient*, *naissais*, *naissait*, *naissant*, *naisse*, *naissent*, *naisses*, *naissez*, *naissiez*, *naissions*, *naissons*, *naît*, *naîtra*, *naîtrai*, *naîtraient*, *naîtrais*, *naîtrait*, *naîtras*, *naître*, *naîtrez*, *naîtriez*, *naîtrions*, *naîtrons*, *naîtront*, *naquîmes*, *naquirent*, *naquis*, *naquisse*, *naquissent*, *naquisses*, *naquissiez*, *naquissions*, *naquit*, *naquît*, *naquîtes*, *né* — "be born" ;

nui, *nuira*, *nuirai*, *nuiraient*, *nuirais*, *nuirait*, *nuiras*, *nuire*, *nuirez*, *nuiriez*, *nuirions*, *nuirons*, *nuiront*, *nuis*, *nuisaient*, *nuisais*, *nuisait*, *nuisant*, *nuise*, *nuisent*, *nuises*, *nuisez*, *nuisiez*, *nuisîmes*, *nuisions*, *nuisirent*, *nuisis*, *nuisisse*, *nuisissent*, *nuisisses*, *nuisissiez*, *nuisissions*, *nuisit*, *nuisît*, *nuisîtes*, *nuisons*, *nuit* — "spoil" ;

offert, *offraient*, *offrais*, *offrait*, *offrant*, *offre*, *offrent*, *offres*, *offrez*, *offriez*, *offrîmes*, *offrions*, *offrir*, *offrira*, *offrirai*, *offriraient*, *offrirais*, *offrirait*, *offriras*, *offrirent*, *offrirez*, *offririez*, *offririons*, *offrirons*, *offriront*, *offris*, *offrisse*, *offrissent*, *offrisses*, *offrissiez*, *offrissions*, *offrit*, *offrît*, *offrîtes*, *offrons* — "offer" ;

peignaient, *peignais*, *peignait*, *peignant*, *peigne*, *peignent*, *peignes*, *peignez*, *peigniez*, *peignîmes*, *peignions*, *peignirent*, *peignis*, *peignisse*, *peignissent*, *peignisses*, *peignissiez*, *peignissions*, *peignit*, *peignît*, *peignîtes*, *peignons*, *peindra*, *peindrai*, *peindraient*, *peindrais*, *peindrait*, *peindras*, *peindre*, *peindrez*, *peindriez*, *peindrions*, *peindrons*, *peindront*, *peins*, *peint* — "paint" ;

plaira, *plairai*, *plairaient*, *plairais*, *plairait*, *plairas*, *plaire*, *plairez*, *plairiez*, *plairions*, *plairons*, *plairont*, *plais*, *plaisaient*, *plaisais*, *plaisait*, *plaisant*, *plaise*, *plaisent*, *plaises*, *plaisez*, *plaisiez*, *plaisions*, *plaisons*, *plait*, *plaît*, *plu*, *plûmes*, *plurent*, *plus*, *plusse*, *plussent*, *plusses*, *plussiez*, *plussions*, *plut*, *plût*, *plûtes* — "please" ;

prenaient, *prenais*, *prenait*, *prenant*, *prend*, *prendra*, *prendrai*, *prendraient*, *prendrais*, *prendrait*, *prendras*, *prendre*, *prendrez*, *prendriez*, *prendrions*, *prendrons*, *prendront*, *prends*, *prenez*, *preniez*, *prenions*, *prenne*, *prennent*, *prennes*, *prenons*, *prîmes*, *prirent*, *pris*, *prisse*, *prissent*, *prisses*, *prissiez*, *prissions*, *prit*, *prît*, *prîtes* — "take" ;

recevaient, *recevais*, *recevait*, *recevant*, *recevez*, *receviez*, *recevions*, *recevoir*, *recevons*, *recevra*, *recevrai*, *recevraient*, *recevrais*, *recevrait*, *recevras*, *recevrez*, *recevriez*, *recevrions*, *recevrons*, *recevront*, *reçois*, *reçoit*, *reçoive*, *reçoivent*, *reçoives*, *reçu*, *reçûmes*, *reçurent*, *reçus*, *reçusse*, *reçussent*, *reçusses*, *reçussiez*, *reçussions*, *reçut*, *reçût*, *reçûtes* — "receive" ;

riaient, *riais*, *riait*, *riant*, *rie*, *rient*, *ries*, *riez*, *riiez*, *riions*, *rîmes*, *rions*, *rira*, *rirai*, *riraient*, *rirais*, *rirait*, *riras*, *rire*, *rirent*, *rirez*, *ririez*, *ririons*, *rirons*, *riront*, *ris*, *risse*, *rissent*, *risses*, *rissiez*, *rissions*, *rit*, *rît*, *rîtes* — "laugh" ;

sachant, *sache*, *sachent*, *saches*, *sachez*, *sachiez*, *sachions*, *sachons*, *sais*, *sait*, *saura*, *saurai*, *sauraient*, *saurais*, *saurait*, *sauras*, *saurez*, *sauriez*, *saurions*, *saurons*, *sauront*, *savaient*, *savais*, *savait*, *savent*, *savez*, *saviez*, *savions*, *savoir*, *savons*, *su*, *sûmes*, *surent*, *sus*, *susse*, *sussent*, *susses*, *sussiez*, *sussions*, *sut*, *sût*, *sûtes* — "know" ;

suffi, *suffimes*, *suffira*, *suffirai*, *suffiraient*, *suffirais*, *suffirait*, *suffiras*, *suffire*, *suffirent*, *suffirez*, *suffiriez*, *suffirions*, *suffirons*, *suffiront*, *suffis*, *suffisaient*, *suffisais*, *suffisait*, *suffisant*, *suffise*, *suffisent*, *suffises*, *suffisez*, *suffisiez*, *suffisions*, *suffisons*, *suffisse*, *suffissent*, *suffisses*, *suffissiez*, *suffissions*, *suffit*, *suffît*, *suffîtes* — "suffice" ;

traie, *traient*, *traies*, *traira*, *trairai*, *trairaient*, *trairais*, *trairait*, *trairas*, *traire*, *trairez*, *trairiez*, *trairions*, *trairons*, *trairont*, *trais*, *trait*, *traya*, *trayai*, *trayaient*, *trayais*, *trayait*, *trayâmes*, *trayant*, *trayas*, *trayasse*, *trayassent*, *trayasses*, *trayassiez*, *trayassions*, *trayât*, *trayâtes*, *trayèrent*, *trayez*, *trayiez*, *trayions*, *trayons* — "milk" ;

vaille, *vaillent*, *vailles*, *valaient*, *valais*, *valait*, *valant*, *valent*, *valez*, *valiez*, *valions*, *valoir*, *valons*, *valu*, *valûmes*, *valurent*, *valus*, *valusse*, *valussent*, *valusses*, *valussiez*, *valussions*, *valut*, *valût*, *valûtes*, *vaudra*, *vaudrai*, *vaudraient*, *vaudrais*, *vaudrait*, *vaudras*, *vaudrez*, *vaudriez*, *vaudrions*, *vaudrons*, *vaudront*, *vaut*, *vaux* — "earn" ;

vainc, *vaincra*, *vaincrai*, *vaincraient*, *vaincrais*, *vaincrait*, *vaincras*, *vaincre*, *vaincrez*, *vaincriez*, *vaincrions*, *vaincrons*, *vaincront*, *vaincs*, *vaincu*, *vainquaient*, *vainquais*, *vainquait*, *vainquant*, *vainque*, *vainquent*, *vainques*, *vainquez*, *vainquiez*, *vainquîmes*, *vainquions*, *vainquirent*, *vainquis*, *vainquisse*, *vainquissent*, *vainquisses*, *vainquissiez*, *vainquissions*, *vainquit*, *vainquît*, *vainquîtes*, *vainquons* — "win" ;

vécu, *vécûmes*, *vécurent*, *vécus*, *vécusse*, *vécussent*, *vécusses*, *vécussiez*, *vécussions*, *vécut*, *vécût*, *vécûtes*, *vis*, *vit*, *vivaient*, *vivais*, *vivait*, *vivant*, *vive*, *vivent*, *vives*, *vivez*, *viviez*, *vivions*, *vivons*, *vivra*, *vivrai*, *vivraient*, *vivrais*, *vivrait*, *vivras*, *vivre*, *vivrez*, *vivriez*, *vivrions*, *vivrons*, *vivront* — "live" ;

vêt, *vêtaient*, *vêtais*, *vêtait*, *vêtant*, *vête*, *vêtent*, *vêtes*, *vêtez*, *vêtiez*, *vêtîmes*, *vêtions*, *vêtir*, *vêtira*, *vêtirai*, *vêtiraient*, *vêtirais*, *vêtirait*, *vêtiras*, *vêtirent*, *vêtirez*, *vêtiriez*, *vêtirions*, *vêtirons*, *vêtiront*, *vêtis*, *vêtisse*, *vêtissent*, *vêtisses*, *vêtissiez*, *vêtissions*, *vêtit*, *vêtît*, *vêtîtes*, *vêtons*, *vêts*, *vêtu* — "dress" ;

veuille, *veuillent*, *veuilles*, *veulent*, *veut*, *veux*, *voudra*, *voudrai*, *voudraient*, *voudrais*, *voudrait*, *voudras*, *voudrez*, *voudriez*, *voudrions*, *voudrons*, *voudront*, *voulaient*, *voulais*, *voulait*, *voulant*, *voulez*, *vouliez*, *voulions*, *vouloir*, *voulons*, *voulu*, *voulûmes*, *voulurent*, *voulus*, *voulusse*, *voulussent*, *voulusses*, *voulussiez*, *voulussions*, *voulut*, *voulût*, *voulûtes* — "wish".

and obtain the following output:

{*accrois*, *accroissaient*, *accroissais*, *accroissait*, *accroissant*, *accroisse*, *accroissent*, *accroisses*, *accroissez*, *accroissiez*, *accroissions*, *accroissons*, *accroît*, *accroîtra*, *accroîtrai*, *accroîtraient*, *accroîtrais*, *accroîtrait*, *accroîtras*, *accroître*, *accroîtrez*, *accroîtriez*, *accroîtrions*, *accroîtrons*, *accroîtront*, *accrû*, *accrûmes*, *accrûrent*, *accrûs*, *accrûsse*, *accrûssent*, *accrûsses*, *accrûssiez*, *accrûssions*, *accrût*, *accrûtes*},

{*acquéraient*, *acquérais*, *acquérait*, *acquérant*, *acquérez*, *acquériez*, *acquérions*, *acquérir*, *acquérons*, *acquerra*, *acquerrai*, *acquerraient*, *acquerrais*, *acquerrait*, *acquerras*, *acquerrez*, *acquerriez*, *acquerrions*, *acquerrons*, *acquerront*, *acquière*, *acquièrent*, *acquières*, *acquiers*, *acquiert*, *acquîmes*, *acquirent*, *acquis*, *acquisse*, *acquissent*, *acquisses*, *acquissiez*, *acquissions*, *acquit*, *acquît*, *acquîtes*},

{*assaillaient*, *assaillais*, *assaillait*, *assaillant*, *assaille*, *assaillé*, *assaillent*, *assailles*, *assaillez*, *assailliez*, *assaillîmes*, *assaillions*, *assaillir*, *assaillira*, *assaillirai*, *assailliraient*, *assaillirais*, *assaillirait*, *assailliras*, *assaillirent*, *assaillirez*, *assailliriez*, *assaillirions*, *assaillirons*, *assailliront*, *assaillis*, *assaillisse*, *assaillissent*, *assaillisses*, *assaillissiez*, *assaillissions*, *assaillit*, *assaillît*, *assaillîtes*, *assaillons*},

{*asseoir*, *asseyaient*, *asseyais*, *asseyait*, *asseyant*, *asseye*, *asseyent*, *asseyes*, *asseyez*, *asseyiez*, *asseyions*, *asseyons*, *assied*, *assieds*, *assiéra*, *assiérai*, *assiéraient*, *assiérais*, *assiérait*, *assiéras*, *assiérez*, *assiériez*, *assiérions*, *assiérons*, *assiéront*, *assîmes*, *assirent*, *assis*, *assisse*, *assissent*, *assisses*, *assissiez*, *assissions*, *assit*, *assît*, *assîtes*, *assoie*, *assoient*, *assoies*, *assoira*, *assoirai*, *assoiraient*, *assoirais*, *assoirait*, *assoiras*, *assoirez*, *assoiriez*, *assoirions*, *assoirons*, *assoiront*, *assois*, *assoit*, *assoyaient*, *assoyais*, *assoyait*, *assoyant*, *assoyez*, *assoyiez*, *assoyions*, *assoyons*},

{*bat*, *bats*, *battaient*, *battais*, *battait*, *battant*, *batte*, *battent*, *battes*, *battez*, *battiez*, *battîmes*, *battions*, *battirent*, *battis*, *battisse*, *battissent*, *battisses*, *battissiez*, *battissions*, *battit*, *battît*, *battîtes*, *battons*, *battra*, *battrai*, *battraient*, *battrais*, *battrait*, *battras*, *battre*, *battrez*, *battriez*, *battrions*, *battrons*, *battront*, *battu*},

{*bouillaient*, *bouillais*, *bouillait*, *bouillant*, *bouille*, *bouillent*, *bouilles*, *bouillez*, *bouilli*, *bouilliez*, *bouillîmes*, *bouillions*, *bouillira*, *bouillirai*, *bouilliraient*, *bouillirais*, *bouillirait*, *bouilliras*, *bouillirent*, *bouillirez*, *bouilliriez*, *bouillirions*, *bouillirons*, *bouilliront*, *bouillis*, *bouillisse*, *bouillissent*, *bouillisses*, *bouillissiez*, *bouillissions*, *bouillit*, *bouillît*, *bouillîtes*, *bouillons*, *bous*, *bout*},

{*braie*, *braient*, *braies*, *braira*, *brairai*, *brairaient*, *brairais*, *brairait*, *brairas*, *braire*, *brairez*, *brairiez*, *brairions*, *brairons*, *brairont*, *brais*, *brait*, *braya*, *brayai*, *brayaient*, *brayais*, *brayait*, *brayâmes*, *brayant*, *brayas*, *brayasse*, *brayassent*, *brayasses*, *brayassiez*, *brayassions*, *brayât*, *brayâtes*, *brayèrent*, *brayez*, *brayiez*, *brayions*, *brayons*},

{*conclu*, *concluaient*, *concluais*, *concluait*, *concluant*, *conclue*, *concluent*, *conclues*, *concluez*, *concluiez*, *concluions*, *conclûmes*, *concluons*, *conclura*, *conclurai*, *concluraient*, *conclurais*, *conclurait*, *concluras*, *conclure*, *conclurent*, *conclurez*, *concluriez*, *conclurions*, *conclurons*, *concluront*, *conclus*, *conclusse*, *conclussent*, *conclusses*, *conclussiez*, *conclussions*, *conclut*, *conclût*, *conclûtes*},

{*coud*, *coudra*, *coudrai*, *coudraient*, *coudrais*, *coudrait*, *coudras*, *coudre*, *coudrez*, *coudriez*, *coudrions*, *coudrons*, *coudront*, *couds*, *cousaient*, *cousais*, *cousait*, *cousant*, *couse*, *cousent*, *couses*, *cousez*, *cousiez*, *cousîmes*, *cousions*, *cousirent*, *cousis*, *cousisse*, *cousissent*, *cousisses*, *cousissiez*, *cousissions*, *cousit*, *cousît*, *cousîtes*, *cousons*, *cousu*},

{*couraient*, *courais*, *courait*, *courant*, *coure*, *courent*, *coures*, *courez*, *couriez*, *courions*, *courir*, *courons*, *courra*, *courrai*, *courraient*, *courrais*, *courrait*, *courras*, *courrez*, *courriez*, *courrions*, *courrons*, *courront*, *cours*, *couru*, *courûmes*, *coururent*, *courus*, *courusse*, *courussent*, *courusses*, *courussiez*, *courussions*, *courut*, *courût*, *courûtes*},

{*court*},

{*craignaient*, *craignais*, *craignait*, *craignant*, *craigne*, *craignent*, *craignes*, *craignez*, *craigniez*, *craignîmes*, *craignions*, *craignirent*, *craignis*, *craignisse*, *craignissent*, *craignisses*, *craignissiez*, *craignissions*, *craignit*, *craignît*, *craignîtes*, *craignons*, *craindra*, *craindrai*, *craindraient*, *craindrais*, *craindrait*, *craindras*, *craindre*, *craindrez*, *craindriez*, *craindrions*, *craindrons*, *craindront*, *crains*, *craint*},

{*croie*, *croient*, *croies*, *croira*, *croirai*, *croiraient*, *croirais*, *croirait*, *croiras*, *croire*, *croirez*, *croiriez*, *croirions*, *croirons*, *croiront*, *crois*, *croit*, *croyaient*, *croyais*, *croyait*, *croyant*, *croyez*, *croyiez*, *croyions*, *croyons*, *cru*, *crûmes*, *crurent*, *crus*, *crusse*, *crussent*, *crusses*, *crussiez*, *crussions*, *crut*, *crût*, *crûtes*},

{*cueillaient*, *cueillais*, *cueillait*, *cueillant*, *cueille*, *cueillent*, *cueillera*, *cueillerai*, *cueilleraient*, *cueillerais*, *cueillerait*, *cueilleras*, *cueillerez*, *cueilleriez*, *cueillerions*, *cueillerons*, *cueilleront*, *cueilles*, *cueillez*, *cueilli*, *cueilliez*, *cueillîmes*, *cueillions*, *cueillir*, *cueillirent*, *cueillis*, *cueillisse*, *cueillissent*, *cueillisses*, *cueillissiez*, *cueillissions*, *cueillit*, *cueillît*, *cueillîtes*, *cueillons*},

{*décherra*, *décherrai*, *décherraient*, *décherrais*, *décherrait*, *décherras*, *décherrez*, *décherriez*, *décherrions*, *décherrons*, *décherront*, *déchet*, *déchoie*, *déchoient*, *déchoies*, *déchoir*, *déchoira*, *déchoirai*, *déchoiraient*, *déchoirais*, *déchoirait*, *déchoiras*, *déchoirez*, *déchoiriez*, *déchoirions*, *déchoirons*, *déchoiront*, *déchois*, *déchoit*, *déchoyant*, *déchoyez*, *déchoyiez*, *déchoyions*, *déchoyons*, *déchu*, *déchûmes*, *déchurent*, *déchus*, *déchusse*, *déchussent*, *déchusses*, *déchussiez*, *déchussions*, *déchut*, *déchût*, *déchûtes*},

{*dîmes*, *dira*, *dirai*, *diraient*, *dirais*, *dirait*, *diras*, *dire*, *dirent*, *direz*, *diriez*, *dirions*, *dirons*, *diront*, *dis*, *disaient*, *disais*, *disait*, *disant*, *dise*, *disent*, *dises*, *disiez*, *disions*, *disons*, *disse*, *dissent*, *disses*, *dissiez*, *dissions*, *dit*, *dît*, *dites*, *dîtes*},

{*dissolûmes*, *dissolurent*, *dissolus*, *dissolut*, *dissolûtes*, *dissolvaient*, *dissolvais*, *dissolvait*, *dissolvant*, *dissolve*, *dissolvent*, *dissolves*, *dissolvez*, *dissolviez*, *dissolvions*, *dissolvons*, *dissoudra*, *dissoudrai*, *dissoudraient*, *dissoudrais*, *dissoudrait*, *dissoudras*, *dissoudre*, *dissoudrez*, *dissoudriez*, *dissoudrions*, *dissoudrons*, *dissoudront*, *dissous*, *dissout*},

{*dormaient*, *dormais*, *dormait*, *dormant*, *dorme*, *dorment*, *dormes*, *dormez*, *dormi*, *dormiez*, *dormîmes*, *dormions*, *dormir*, *dormira*, *dormirai*, *dormiraient*, *dormirais*, *dormirait*, *dormiras*, *dormirent*, *dormirez*, *dormiriez*, *dormirions*, *dormirons*, *dormiront*, *dormis*, *dormisse*, *dormissent*, *dormisses*, *dormissiez*, *dormissions*, *dormit*, *dormît*, *dormîtes*, *dormons*, *dors*, *dort*},

{*éclora*, *éclorai*, *écloraient*, *éclorais*, *éclorait*, *écloras*, *éclore*, *éclorez*, *écloriez*, *éclorions*, *éclorons*, *écloront*, *éclos*, *éclosant*, *éclose*, *éclosent*, *écloses*, *éclosez*, *éclosiez*, *éclosions*, *éclosons*, *éclôt*},

{*écrira*, *écrirai*, *écriraient*, *écrirais*, *écrirait*, *écriras*, *écrire*, *écrirez*, *écririez*, *écririons*, *écrirons*, *écriront*, *écris*, *écrit*, *écrivaient*, *écrivais*, *écrivait*, *écrivant*, *écrive*, *écrivent*, *écrives*, *écrivez*, *écriviez*, *écrivîmes*, *écrivions*, *écrivirent*, *écrivis*, *écrivisse*, *écrivissent*, *écrivisses*, *écrivissiez*, *écrivissions*, *écrivit*, *écrivît*, *écrivîtes*, *écrivons*},

{*enverra*, *enverrai*, *enverraient*, *enverrais*, *enverrait*, *enverras*, *enverrez*, *enverriez*, *enverrions*, *enverrons*, *enverront*, *envoie*, *envoient*, *envoies*, *envoya*, *envoyai*, *envoyaient*, *envoyais*, *envoyait*, *envoyâmes*, *envoyant*, *envoyas*, *envoyasse*, *envoyassent*, *envoyasses*, *envoyassiez*, *envoyassions*, *envoyât*, *envoyâtes*, *envoyé*, *envoyer*, *envoyèrent*, *envoyez*, *envoyiez*, *envoyions*, *envoyons*},

{*fui*, *fuie*, *fuient*, *fuies*, *fuîmes*, *fuir*, *fuira*, *fuirai*, *fuiraient*, *fuirais*, *fuirait*, *fuiras*, *fuirent*, *fuirez*, *fuiriez*, *fuirions*, *fuirons*, *fuiront*, *fuis*, *fuisse*, *fuissent*, *fuisses*, *fuissiez*, *fuissions*, *fuit*, *fuît*, *fuîtes*, *fuyaient*, *fuyais*, *fuyait*, *fuyant*, *fuyez*, *fuyiez*, *fuyions*, *fuyons*},

{*incluaient*, *incluais*, *incluait*, *incluant*, *inclue*, *incluent*, *inclues*, *incluez*, *incluiez*, *incluions*, *inclûmes*, *incluons*, *inclura*, *inclurai*, *incluraient*, *inclurais*, *inclurait*, *incluras*, *inclure*, *inclurent*, *inclurez*, *incluriez*, *inclurions*, *inclurons*, *incluront*, *inclus*, *inclusse*, *inclussent*, *inclusses*, *inclussiez*, *inclussions*, *inclut*, *inclût*, *inclûtes*},

{*joignaient*, *joignais*, *joignait*, *joignant*, *joigne*, *joignent*, *joignes*, *joignez*, *joigniez*, *joignîmes*, *joignions*, *joignirent*, *joignis*, *joignisse*, *joignissent*, *joignisses*, *joignissiez*, *joignissions*, *joignit*, *joignît*, *joignîtes*, *joignons*, *joindra*, *joindrai*, *joindraient*, *joindrais*, *joindrait*, *joindras*, *joindre*, *joindrez*, *joindriez*, *joindrions*, *joindrons*, *joindront*, *joins*, *joint*},

{*lira*, *lirai*, *liraient*, *lirais*, *lirait*, *liras*, *lire*, *lirez*, *liriez*, *lirions*, *lirons*, *liront*, *lis*, *lisaient*, *lisais*, *lisait*, *lisant*, *lise*, *lisent*, *lises*, *lisez*, *lisiez*, *lisions*, *lisons*, *lit*, *lu*, *lûmes*, *lurent*, *lus*, *lusse*, *lussent*, *lusses*, *lussiez*, *lussions*, *lut*, *lût*, *lûtes*},

{*meure*, *meurent*, *meures*, *meurs*, *meurt*, *mort*, *mouraient*, *mourais*, *mourait*, *mourant*, *mourez*, *mouriez*, *mourions*, *mourir*, *mourons*, *mourra*, *mourrai*, *mourraient*, *mourrais*, *mourrait*, *mourras*, *mourrez*, *mourriez*, *mourrions*, *mourrons*, *mourront*, *mourûmes*, *moururent*, *mourus*, *mourusse*, *mourussent*, *mourusses*, *mourussiez*, *mourussions*, *mourut*, *mourût*, *mourûtes*},

{*meus*, *meut*, *meuve*, *meuvent*, *meuves*, *mouvaient*, *mouvais*, *mouvait*, *mouvant*, *mouvez*, *mouviez*, *mouvions*, *mouvoir*, *mouvons*, *mouvra*, *mouvrai*, *mouvraient*, *mouvrais*, *mouvrait*, *mouvras*, *mouvrez*, *mouvriez*, *mouvrions*, *mouvrons*, *mouvront*, *mû*, *mûmes*, *murent*, *mus*, *musse*, *mussent*, *musses*, *mussiez*, *mussions*, *mut*, *mût*, *mûtes*},

{*moud*, *moudra*, *moudrai*, *moudraient*, *moudrais*, *moudrait*, *moudras*, *moudre*, *moudrez*, *moudriez*, *moudrions*, *moudrons*, *moudront*, *mouds*, *moulaient*, *moulais*, *moulait*, *moulant*, *moule*, *moulent*, *moules*, *moulez*, *mouliez*, *moulions*, *moulons*, *moulu*, *moulûmes*, *moulurent*, *moulus*, *moulusse*, *moulussent*, *moulusses*, *moulussiez*, *moulussions*, *moulut*, *moulût*, *moulûtes*},

{*nais*, *naissaient*, *naissais*, *naissait*, *naissant*, *naisse*, *naissent*, *naisses*, *naissez*, *naissiez*, *naissions*, *naissons*, *naît*, *naîtra*, *naîtrai*, *naîtraient*, *naîtrais*, *naîtrait*, *naîtras*, *naître*, *naîtrez*, *naîtriez*, *naîtrions*, *naîtrons*, *naîtront*, *naquîmes*, *naquirent*, *naquis*, *naquisse*, *naquissent*, *naquisses*, *naquissiez*, *naquissions*, *naquit*, *naquît*, *naquîtes*, *né*},

{*nui*, *nuira*, *nuirai*, *nuiraient*, *nuirais*, *nuirait*, *nuiras*, *nuire*, *nuirez*, *nuiriez*, *nuirions*, *nuirons*, *nuiront*, *nuis*, *nuisaient*, *nuisais*, *nuisait*, *nuisant*, *nuise*, *nuisent*, *nuises*, *nuisez*, *nuisiez*, *nuisîmes*, *nuisions*, *nuisirent*, *nuisis*, *nuisisse*, *nuisissent*, *nuisisses*, *nuisissiez*, *nuisissions*, *nuisit*, *nuisît*, *nuisîtes*, *nuisons*, *nuit*},

{*offert*, *offraient*, *offrais*, *offrait*, *offrant*, *offre*, *offrent*, *offres*, *offrez*, *offriez*, *offrîmes*, *offrions*, *offrir*, *offrira*, *offrirai*, *offriraient*, *offrirais*, *offrirait*, *offriras*, *offrirent*, *offrirez*, *offririez*, *offririons*, *offrirons*, *offriront*, *offris*, *offrisse*, *offrissent*, *offrisses*, *offrissiez*, *offrissions*, *offrit*, *offrît*, *offrîtes*, *offrons*},

{*peignaient*, *peignais*, *peignait*, *peignant*, *peigne*, *peignent*, *peignes*, *peignez*, *peigniez*, *peignîmes*, *peignions*, *peignirent*, *peignis*, *peignisse*, *peignissent*, *peignisses*, *peignissiez*, *peignissions*, *peignit*, *peignît*, *peignîtes*, *peignons*, *peindra*, *peindrai*, *peindraient*, *peindrais*, *peindrait*, *peindras*, *peindre*, *peindrez*, *peindriez*, *peindrions*, *peindrons*, *peindront*, *peins*, *peint*},

{*plaira*, *plairai*, *plairaient*, *plairais*, *plairait*, *plairas*, *plaire*, *plairez*, *plairiez*, *plairions*, *plairons*, *plairont*, *plais*, *plaisaient*, *plaisais*, *plaisait*, *plaisant*, *plaise*, *plaisent*, *plaises*, *plaisez*, *plaisiez*, *plaisions*, *plaisons*, *plait*, *plaît*},

{*plu*, *plûmes*, *plurent*, *plus*, *plusse*, *plussent*, *plusses*, *plussiez*, *plussions*, *plut*, *plût*, *plûtes*},

{*prenaient*, *prenais*, *prenait*, *prenant*, *prend*, *prendra*, *prendrai*, *prendraient*, *prendrais*, *prendrait*, *prendras*, *prendre*, *prendrez*, *prendriez*, *prendrions*, *prendrons*, *prendront*, *prends*, *prenez*, *preniez*, *prenions*, *prenne*, *prennent*, *prennes*, *prenons*, *prîmes*, *prirent*, *pris*, *prisse*, *prissent*, *prisses*, *prissiez*, *prissions*, *prit*, *prît*, *prîtes*},

{*recevaient*, *recevais*, *recevait*, *recevant*, *recevez*, *receviez*, *recevions*, *recevoir*, *recevons*, *recevra*, *recevrai*, *recevraient*, *recevrais*, *recevrait*, *recevras*, *recevrez*, *recevriez*, *recevrions*, *recevrons*, *recevront*, *reçois*, *reçoit*, *reçoive*, *reçoivent*, *reçoives*, *reçu*, *reçûmes*, *reçurent*, *reçus*, *reçusse*, *reçussent*, *reçusses*, *reçussiez*, *reçussions*, *reçut*, *reçût*, *reçûtes*},

{*riaient*, *riais*, *riait*, *riant*, *rie*, *rient*, *ries*, *riez*, *riiez*, *riions*, *rîmes*, *rions*, *rira*, *rirai*, *riraient*, *rirais*, *rirait*, *riras*, *rire*, *rirent*, *rirez*, *ririez*, *ririons*, *rirons*, *riront*, *ris*, *risse*, *rissent*, *risses*, *rissiez*, *rissions*, *rit*, *rît*, *rîtes*},

{*sachant*, *sache*, *sachent*, *saches*, *sachez*, *sachiez*, *sachions*, *sachons*, *sais*, *sait*, *saura*, *saurai*, *sauraient*, *saurais*, *saurait*, *sauras*, *saurez*, *sauriez*, *saurions*, *saurons*, *sauront*, *savaient*, *savais*, *savait*, *savent*, *savez*, *saviez*, *savions*, *savoir*, *savons*, *su*, *sûmes*, *surent*, *sus*, *susse*, *sussent*, *susses*, *sussiez*, *sussions*, *sut*, *sût*, *sûtes*},

{*suffi*, *suffîmes*, *suffira*, *suffirai*, *suffiraient*, *suffirais*, *suffirait*, *suffiras*, *suffire*, *suffirent*, *suffirez*, *suffiriez*, *suffirions*, *suffirons*, *suffiront*, *suffis*, *suffisaient*, *suffisais*, *suffisait*, *suffisant*, *suffise*, *suffisent*, *suffises*, *suffisez*, *suffisiez*, *suffisions*, *suffisons*, *suffisse*, *suffissent*, *suffisses*, *suffissiez*, *suffissions*, *suffit*, *suffit*, *suffîtes*},

{*traie*, *traient*, *traies*, *traira*, *trairai*, *trairaient*, *trairais*, *trairait*, *trairas*, *traire*, *trairez*, *trairiez*, *trairions*, *trairons*, *trairont*, *trais*, *trait*, *traya*, *trayai*, *trayaient*, *trayais*, *trayait*, *trayâmes*, *trayant*, *trayas*, *trayasse*, *trayassent*, *trayasses*, *trayassiez*, *trayassions*, *trayât*, *trayâtes*, *trayèrent*, *trayez*, *trayiez*, *trayions*, *trayons*},

{*vaille*, *vaillent*, *vailles*, *valaient*, *valais*, *valait*, *valant*, *valent*, *valez*, *valiez*, *valions*, *valoir*, *valons*, *valu*, *valûmes*, *valurent*, *valus*, *valusse*, *valussent*, *valusses*, *valussiez*, *valussions*, *valut*, *valût*, *valûtes*, *vaudra*, *vaudrai*, *vaudraient*, *vaudrais*, *vaudrait*, *vaudras*, *vaudrez*, *vaudriez*, *vaudrions*, *vaudrons*, *vaudront*, *vaut*, *vaux*},

{*vainc*, *vaincra*, *vaincrai*, *vaincraient*, *vaincrais*, *vaincrait*, *vaincras*, *vaincre*, *vaincrez*, *vaincriez*, *vaincrions*, *vaincrons*, *vaincront*, *vaincs*, *vaincu*, *vainquaient*, *vainquais*, *vainquait*, *vainquant*, *vainque*, *vainquent*, *vainques*, *vainquez*, *vainquiez*, *vainquîmes*, *vainquions*, *vainquirent*, *vainquis*, *vainquisse*, *vainquissent*, *vainquisses*, *vainquissiez*, *vainquissions*, *vainquit*, *vainquît*, *vainquîtes*, *vainquons*},

{*vécu*, *vécûmes*, *vécurent*, *vécus*, *vécusse*, *vécussent*, *vécusses*, *vécussiez*, *vécussions*, *vécut*, *vécût*, *vécûtes*, *vivaient*, *vivais*, *vivait*, *vivant*, *vive*, *vivent*, *vives*, *vivez*, *viviez*, *vivions*, *vivons*, *vivra*, *vivrai*, *vivraient*, *vivrais*, *vivrait*, *vivras*, *vivre*, *vivrez*, *vivriez*, *vivrions*, *vivrons*, *vivront*},

{*vêt*, *vêtaient*, *vêtais*, *vêtait*, *vêtant*, *vête*, *vêtent*, *vêtes*, *vêtez*, *vêtiez*, *vêtîmes*, *vêtions*, *vêtir*, *vêtira*, *vêtirai*, *vêtiraient*, *vêtirais*, *vêtirait*, *vêtiras*, *vêtirent*, *vêtirez*, *vêtiriez*, *vêtirions*, *vêtirons*, *vêtiront*, *vêtis*, *vêtisse*, *vêtissent*, *vêtisses*, *vêtissiez*, *vêtissions*, *vêtit*, *vêtît*, *vêtîtes*, *vêtons*, *vêts*, *vêtu*},

{*veuille*, *veuillent*, *veuilles*, *veulent*, *veut*, *veux*, *voudra*, *voudrai*, *voudraient*, *voudrais*, *voudrait*, *voudras*, *voudrez*, *voudriez*, *voudrions*, *voudrons*, *voudront*, *voulaient*, *voulais*, *voulait*, *voulant*, *voulez*, *vouliez*, *voulions*, *vouloir*, *voulons*, *voulu*, *voulûmes*, *voulurent*, *voulus*, *voulusse*, *voulussent*, *voulusses*, *voulussiez*, *voulussions*, *voulut*, *voulût*, *voulûtes*},

{*vis*, *vit*},

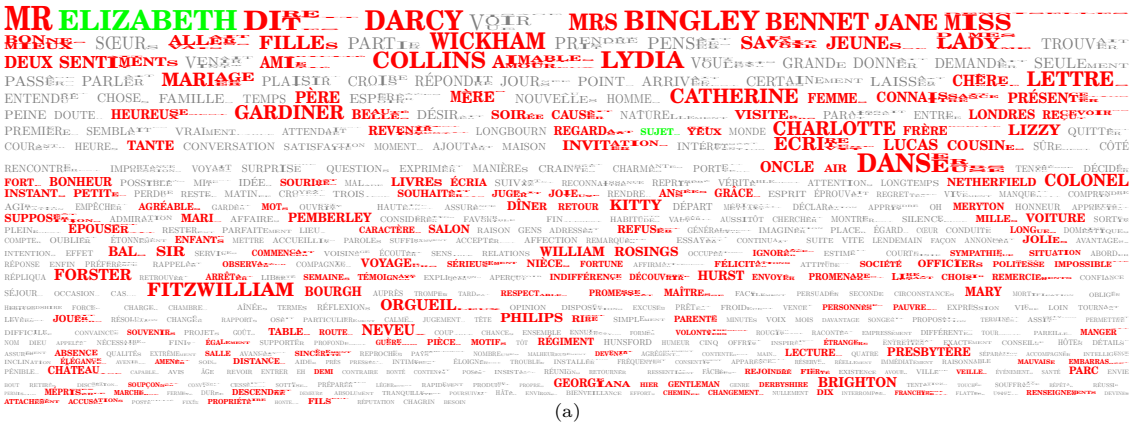
Here, some verb conjugations are not clustered as expected, because we want to accommodate to inherent ambiguities of certain French verb forms. For example, both *vis* and *vit* may be conjugated forms of *vivre* “live” or *voir* “see”, and the infinitive form of *plu* may be either *plaire* “please” or *pleuvoir* “rain”.

Example 6.20.3. In [1, Figs. 3c, 6] and Fig. S8, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text sources).

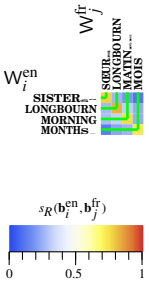
It should be noted that the performance of machine translation is affected by polysemy of certain commonly occurring French words. For example, *livre* “book, pound” and *livres* “books, pounds” can also be conjugated forms of the verb *livrer* “deliver”. Furthermore, a word in English may correspond to a phrase in French, such as *jeune fille* “girl” (in contrast with *fille* “daughter”). This explains why our algorithm does not identify French translations for such common words like “girl” and “daughter”.

Topicality (non-Poissonian behavior) plays an essential rôle in our algorithms. If a word pattern generates a significantly non-Poissonian trajectory, then it will have distinguishing features in its time structure ([1, Figs. 3c, 6] and Fig. S8c), which allow better semantic resolution than nearly Poissonian trajectories (Fig. S8b and Fig. S8d). The recurrence kinetics of non-topics are very close to single exponential decays, so the vector embedding of recurrence eigenvalues ([1, Figs. 3c, 6] and Fig. S8c) will not help here.

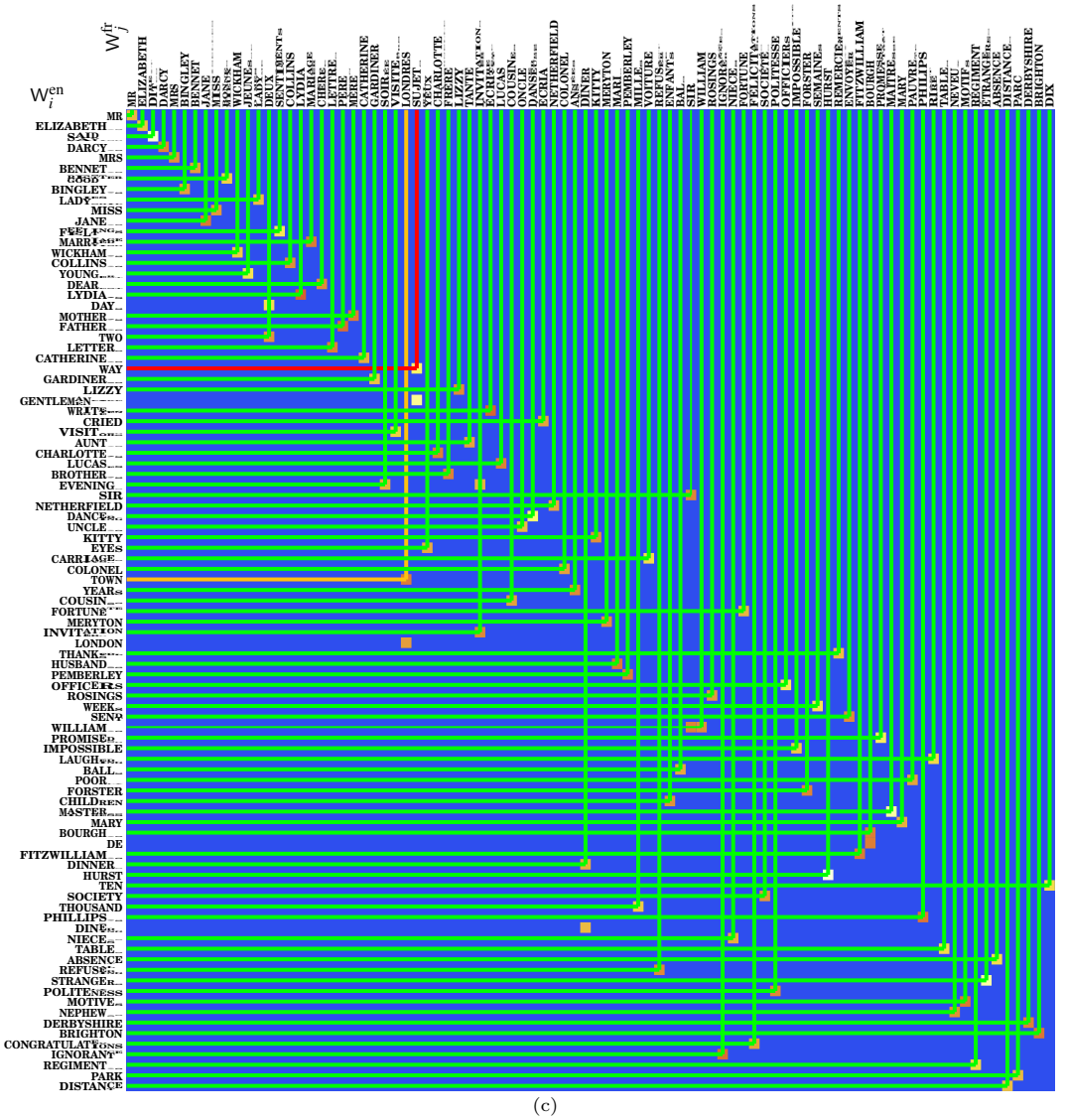
In French, “bat” (a kind of flying mammal) is called *chauve-souris* (literally “bald mouse”), so we consider *bat* an exact match to *chauve* “bald” in Fig. S8b″. (Also note that the row for *bat* in Fig. S8b″ indeed contains two hot spots, one for *chauve* and one for *souris*.) In Darwin’s *Origin of Species*, the adjective *fresh* is mostly used in the context of *fresh water*, which is *eau douce* in French. Therefore, we consider *fresh* an exact match to *douce* (meaning “sweet” in general) in Fig. S8b″.



(a)

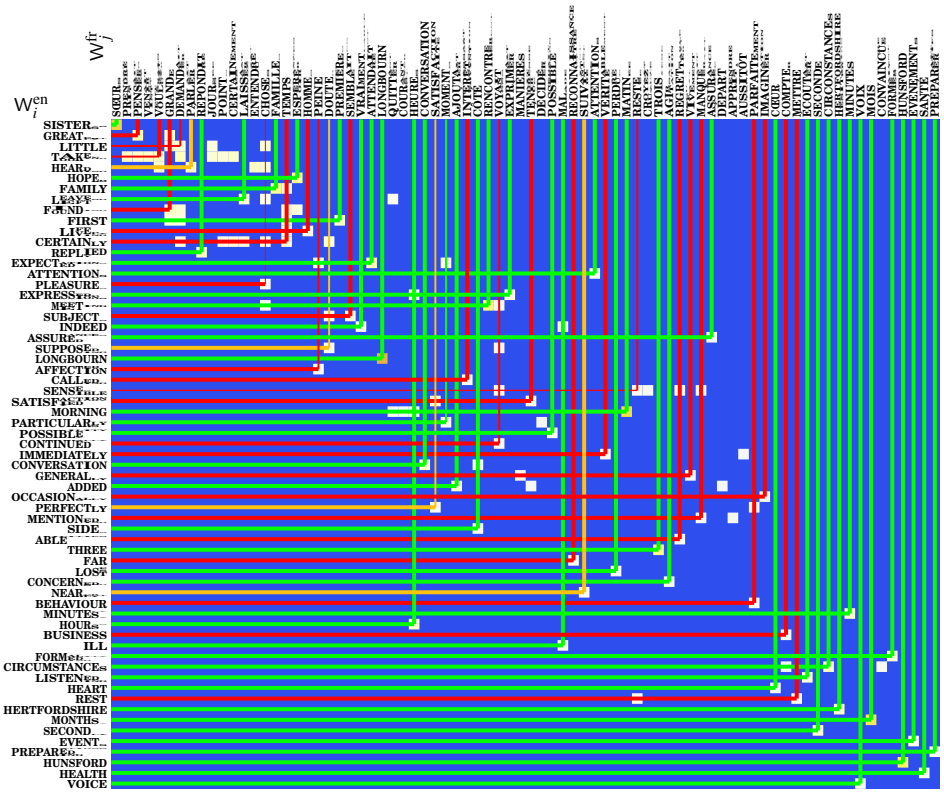


(b)



(c)

Fig. S8. Text mining in French. (a) Statistically identified topics ($n_{ii} \geq 20$) in a French version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) An experiment parallel to [1, Fig. 3c], carried out on non-topics instead of topics. Rows and columns with maximal $s_R(\mathbf{b}_i^{\text{en}}, \mathbf{b}_j^{\text{fr}})$ less than 0.7 are not shown. Correct matchings are indicated by green cross-hairs. (c) A control experiment for [1, Fig. 6]. The same ballpark screening on bulk vectors (based on Poisson noise model) is used. The original censorship $s_R(\mathbf{v}_i^{\text{en}}, \mathbf{v}_j^{\text{fr}}) \geq 0.7$ is no longer imposed here.



(d)

Fig. S8. Text mining in French. (*Continued*) (d) An experiment parallel to [1, Fig. 6]. Non-topics first undergo the same ballpark screening as in [1, Fig. 6], then their $s_R(\mathbf{b}_i^{en}, \mathbf{b}_j^{fr})$ scores are sent to the Hungarian algorithm for bipartite matching.

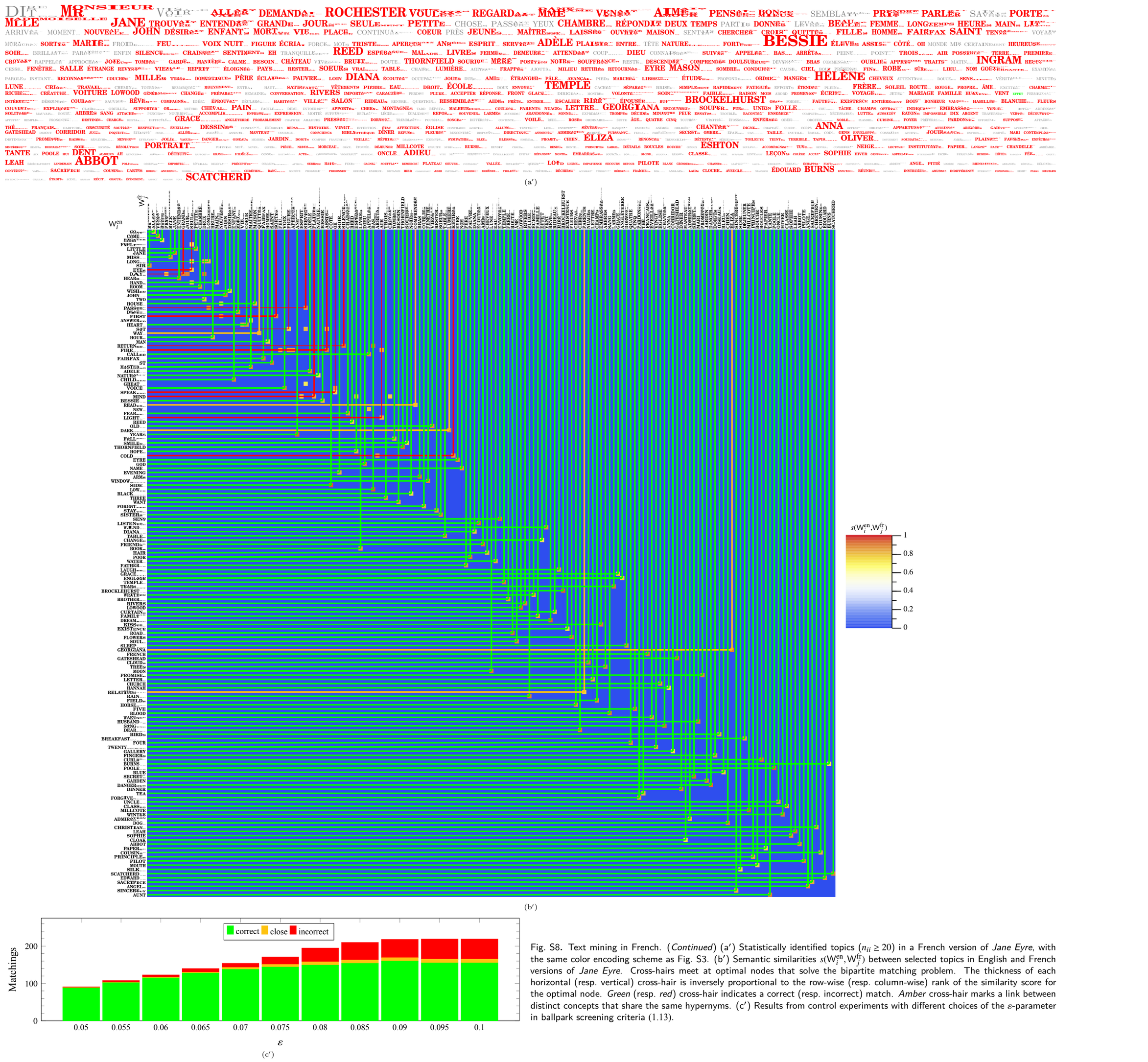
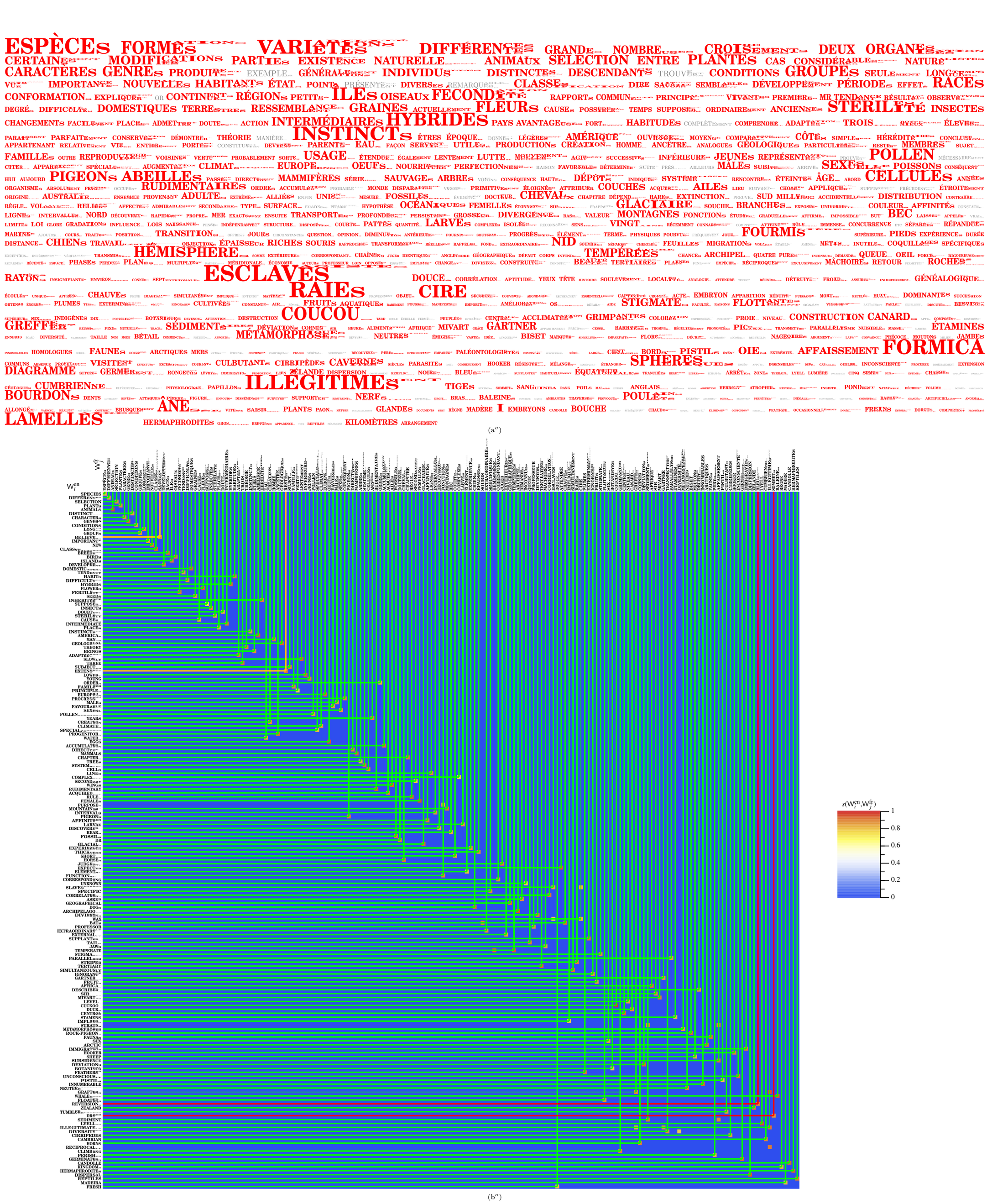


Fig. S8. Text mining in French. (*Continued*) (a′) Statistically identified topics ($n_{ii} \geq 20$) in a French version of *Jane Eyre*, with the same color encoding scheme as Fig. S3. (b′) Semantic similarities $s(W_i^{en}, W_j^{fr})$ between selected topics in English and French versions of *Jane Eyre*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c′) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).



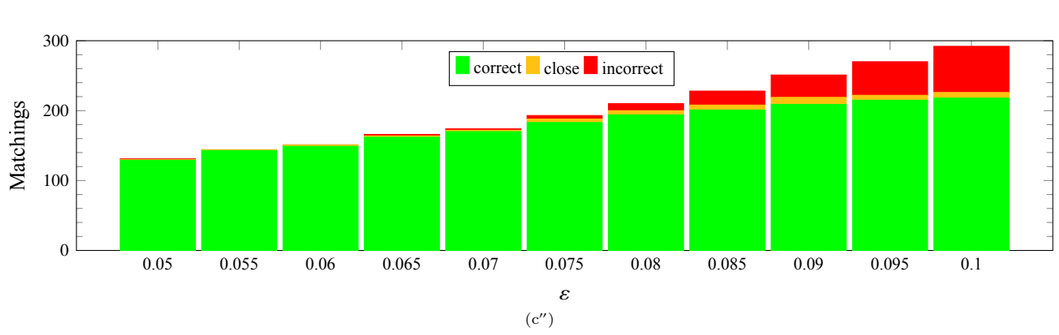


Fig. S8. Text mining in French. (*Continued*) (a″) Statistically identified topics ($n_{ii} \geq 20$) in a French version of *Origin of Species*, with the same color encoding scheme as Fig. S3. (b″) Semantic similarities $s(W_i^{en}, W_j^{fr})$ between selected topics in English and French versions of *Origin of Species*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c″) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

6.3 Modified Porter stemming algorithm for Latin

Definition 6.21 (Latin stop words)**.** If a word belongs to the following list[79]:

a, *ab*, *abhinc*, *absque*, *abusque*, *ac*, *acta*, *actu*, *actum*, *acturus*, *actus*, *ad*, *adaeque*, *adhic*, *adhuc*, *adusque*, *adversum*, *adversus*, *agam*, *agamini*, *agamur*, *agamus*, *agant*, *agantur*, *agar*, *agare*, *agaris*, *agas*, *agat*, *agatis*, *agatur*, *age*, *agebam*, *agebamini*, *agebamur*, *agebamus*, *agebant*, *agebantur*, *agebar*, *agebare*, *agebaris*, *agebas*, *agebat*, *agebatis*, *agebatur*, *agemini*, *agemur*, *agemus*, *agendi*, *agendo*, *agendum*, *agendus*, *agens*, *agent*, *agentur*, *agere*, *agerem*, *ageremini*, *ageremur*, *ageremus*, *agerent*, *agerentur*, *agerer*, *agerere*, *agereris*, *ageres*, *ageret*, *ageretis*, *ageretur*, *ageris*, *ages*, *aget*, *agetis*, *agetur*, *agi*, *agimini*, *agimur*, *agimus*, *agis*, *agit*, *agite*, *agitis*, *agito*, *agitor*, *agitote*, *agitur*, *ago*, *agor*, *agunt*, *agunto*, *aguntor*, *aguntur*, *alia*, *aliae*, *aliam*, *aliarum*, *alias*, *alicui*, *alicuius*, *alicujus*, *alie*, *alii*, *aliis*, *alio*, *aliorum*, *alios*, *aliquae*, *aliquando*, *aliquantulum*, *aliquarum*, *aliquas*, *aliquem*, *aliqui*, *aliquibus*, *aliquid*, *aliquis*, *aliquo*, *aliquorum*, *aliquos*, *aliquot*, *aliquotiens*, *aliter*, *aliud*, *alium*, *alius*, *alter*, *altera*, *alterae*, *alteram*, *alterarum*, *alteras*, *alteri*, *alteris*, *alterius*, *altero*, *alterorum*, *alteros*, *alterum*, *an*, *ante*, *antea*, *apsque*, *apud*, *at*, *atque*, *attorque*, *aut*, *autem*, *caetera*, *cætera*, *caeterae*, *cæteræ*, *caeteram*, *cæteram*, *caeterarum*, *cæterarum*, *caeteras*, *cæteras*, *caetere*, *cætere*, *caeteri*, *cæteri*, *caeteris*, *cæteris*, *caetero*, *cætero*, *caeterorum*, *cæterorum*, *caeteros*, *cæteros*, *caeterum*, *cæterum*, *caeterus*, *cæterus*, *causa*, *cetera*, *ceterae*, *ceteræ*, *ceteram*, *ceterarum*, *ceteras*, *cetere*, *ceteri*, *ceteris*, *cetero*, *ceterorum*, *ceteros*, *ceterum*, *ceterus*, *circa*, *circum*, *cis*, *citra*, *com*, *concoque*, *contorque*, *contra*, *coque*, *coram*, *cui*, *cuidam*, *cuiquam*, *cuique*, *cuius*, *cuiusdam*, *cuiusquam*, *cuiusque*, *cuiusvis*, *cuivis*, *cujus*, *cujusdam*, *cujusquam*, *cujusque*, *cujusvis*, *cum*, *cur*, *de*, *decoque*, *dein*, *deinde*, *denique*, *denuo*, *deque*, *detorque*, *dum*, *e*, *ea*, *eadem*, *eae*, *eaedem*, *eam*, *eandem*, *earum*, *earundem*, *easdem*, *ecquando*, *egeram*, *egeramus*, *egerant*, *egeras*, *egerat*, *egeratis*, *egere*, *egerim*, *egerimus*, *egerint*, *egeris*, *egerit*, *egeritis*, *egero*, *egerunt*, *egi*, *egimus*, *egisse*, *egissem*, *egissemus*, *egissent*, *egisses*, *egisset*, *egissetis*, *egisti*, *egistis*, *egit*, *ego*, *ei*, *eidem*, *eis*, *eisdem*, *eius*, *eiusdem*, *ejus*, *ejusdem*, *en*, *enim*, *eo*, *eodem*, *eorum*, *eorundem*, *eos*, *eosdem*, *eram*, *eramus*, *erant*, *eras*, *erat*, *eratis*, *ere*, *erga*, *ergo*, *erimus*, *eris*, *erit*, *eritis*, *ero*, *erunt*, *es*, *esse*, *essem*, *essemus*, *essent*, *esses*, *esset*, *essetis*, *est*, *este*, *estis*, *esto*, *estote*, *et*, *etiam*, *etiamsi*, *etsi*, *eum*, *eundem*, *ex*, *excoque*, *extorque*, *extra*, *fac*, *face*, *facere*, *facerem*, *faceremus*, *facerent*, *faceres*, *faceret*, *faceretis*, *faciam*, *faciamus*, *faciant*, *facias*, *faciat*, *faciatis*, *faciebam*, *faciebamus*, *faciebant*, *faciebas*, *faciebat*, *faciebatis*, *faciemus*, *faciendi*, *faciendo*, *faciendum*, *faciendus*, *faciens*, *facient*, *facies*, *faciet*, *facietis*, *facimus*, *facio*, *facis*, *facit*, *facite*, *facitis*, *facito*, *facitote*, *faciunt*, *faciunto*, *facta*, *factae*, *factam*, *factarum*, *factas*, *facte*, *facti*, *factis*, *facto*, *factorum*, *factos*, *factu*, *factum*, *facturus*, *factus*, *feceram*, *feceramus*, *fecerant*, *feceras*, *fecerat*, *feceratis*, *fecere*, *fecerim*, *fecerimus*, *fecerint*, *feceris*, *fecerit*, *feceritis*, *fecero*, *fecerunt*, *feci*, *fecimus*, *fecisse*, *fecissem*, *fecissemus*, *fecissent*, *fecisses*, *fecisset*, *fecissetis*, *fecisti*, *fecistis*, *fecit*, *fi*, *fiam*, *fiamus*, *fiant*, *fias*, *fiat*, *fiatis*, *fiebam*, *fiebamus*, *fiebant*, *fiebas*, *fiebat*, *fiebatis*, *fiemus*, *fiendi*, *fiendo*, *fiendum*, *fient*, *fierem*, *fieremus*, *fierent*, *fieres*, *fieret*, *fieretis*, *fieri*, *fies*, *fiet*, *fietis*, *fimus*, *fio*, *fis*, *fit*, *fite*, *fitis*, *fito*, *fitote*, *fiunt*, *fiunto*, *fore*, *forem*, *foremus*, *forent*, *fores*, *foret*, *foretis*, *forms*, *forsan*, *forsit*, *forsitan*, *fortasse*, *fortassis*, *frequenter*, *frequentissime*, *frequentius*, *fueram*, *fueramus*, *fuerant*, *fueras*, *fuerat*, *fueratis*, *fuere*, *fuerim*, *fuerimus*, *fuerint*, *fueris*, *fuerit*, *fueritis*, *fuero*, *fuerunt*, *fui*, *fuimus*, *fuisse*, *fuissem*, *fuissemus*, *fuissent*, *fuisses*, *fuisset*, *fuissetis*, *fuisti*, *fuistis*, *fuit*, *futura*, *futurae*, *futuram*, *futurarum*, *futuras*, *future*, *futuri*, *futuris*, *futuro*, *futurorum*, *futuros*, *futurum*, *futurus*, *habe*, *habeam*, *habeamini*, *habeamur*, *habeamus*, *habeant*, *habeantur*, *habear*, *habeare*, *habearis*, *habeas*, *habeat*, *habeatis*, *habeatur*, *habebam*, *habebamini*, *habebamur*, *habebamus*, *habebant*, *habebantur*, *habebar*, *habebare*, *habebaris*, *habebas*, *habebat*, *habebatis*, *habebatur*, *habebere*, *habeberis*, *habebimini*, *habebimur*, *habebimus*, *habebis*, *habebit*, *habebitis*, *habebitur*, *habebo*, *habebor*, *habebunt*, *habebuntur*, *habemini*, *habemur*, *habemus*, *habenda*, *habendae*, *habendam*, *habendarum*, *habendas*, *habende*, *habendi*, *habendis*, *habendo*, *habendorum*, *habendos*, *habendum*, *habendus*, *habens*, *habent*, *habente*, *habentem*, *habentes*, *habenti*, *habentia*, *habentibus*, *habentis*, *habentium*, *habento*, *habentor*, *habentur*, *habeo*, *habeor*, *habere*, *haberem*, *haberemini*, *haberemur*, *haberemus*, *haberent*, *haberentur*, *haberer*, *haberere*, *habereris*, *haberes*, *haberet*, *haberetis*, *haberetur*, *haberi*, *haberier*, *haberis*, *habes*, *habet*, *habete*, *habetis*, *habeto*, *habetor*, *habetote*, *habetur*, *habita*, *habitae*, *habitam*, *habitarum*, *habitas*, *habite*, *habiti*, *habitis*, *habito*, *habitorum*, *habitos*, *habitu*, *habitum*, *habitura*, *habiturae*, *habituram*, *habiturarum*, *habituras*, *habiture*, *habituri*, *habituris*, *habituro*, *habiturorum*, *habituros*, *habiturum*, *habiturus*, *habitus*, *habueram*, *habueramus*, *habuerant*, *habueras*, *habuerat*, *habueratis*, *habuere*, *habuerim*, *habuerimus*, *habuerint*, *habueris*, *habuerit*, *habueritis*, *habuero*, *habuerunt*, *habui*, *habuimus*, *habuisse*, *habuissem*, *habuissemus*, *habuissent*, *habuisses*, *habuisset*, *habuissetis*, *habuisti*, *habuistis*, *habuit*, *hac*, *hae*, *haec*, *hanc*, *harum*, *has*, *haud*, *hi*, *hic*, *hinc*, *his*, *hoc*, *horum*, *hos*, *huc*, *huic*, *huius*, *hujus*, *hunc*, *iam*, *ibi*, *id*, *idem*, *igitur*, *iis*, *illa*, *illae*, *illam*, *illarum*, *illas*, *ille*, *illi*, *illis*, *illius*, *illo*, *illorum*, *illos*, *illud*, *illum*, *immo*, *in*, *incoque*, *inde*, *infra*, *infrequenter*,

[79] Our list of Latin stop words is based on `https://wiki.digitalclassicist.org/Stopwords_for_Greek_and_Latin` and `http://snowball.tartarus.org/otherapps/schinke/intro.html`, with extensive additions to roughly match their counterparts in English. In particular, we have included all the conjugated forms of *faciō* "make", *fīō* "become", *sum* "be" and *possum* "can". It is worth noting that the conjugation tables for *faciō* and *fīō* partially overlap.

infrequentissime, *infrequentius*, *inquam*, *inque*, *inquiat*, *inquiebat*, *inquiens*, *inquies*, *inquiet*, *inquii*, *inquimus*, *inquis*, *inquisti*, *inquit*, *inquitis*, *inquito*, *inquiunt*, *inter*, *interim*, *intorque*, *intra*, *ipsa*, *ipsae*, *ipsam*, *ipsarum*, *ipsas*, *ipse*, *ipsi*, *ipsis*, *ipsius*, *ipso*, *ipsorum*, *ipsos*, *ipsum*, *is*, *isdem*, *ista*, *istae*, *istam*, *istarum*, *istas*, *iste*, *isti*, *istis*, *istius*, *isto*, *istorum*, *istos*, *istud*, *istum*, *ita*, *itaque*, *item*, *iter*, *iterum*, *iuxta*, *jam*, *juxta*, *magis*, *me*, *mea*, *meae*, *meam*, *mearum*, *meas*, *mecum*, *mei*, *meis*, *meo*, *meorum*, *meos*, *meum*, *meus*, *mi*, *mihi*, *minime*, *minus*, *modo*, *mox*, *multa*, *multae*, *multam*, *multarum*, *multas*, *multe*, *multi*, *multis*, *multo*, *multorum*, *multos*, *multum*, *multus*, *nam*, *ne*, *nec*, *necque*, *nemine*, *neminem*, *nemini*, *neminis*, *nemo*, *neque*, *nequeam*, *nequeamus*, *nequeant*, *nequeas*, *nequeat*, *nequeatis*, *nequeo*, *nequeundi*, *nequeundo*, *nequeundum*, *nequeunt*, *nequeunto*, *nequi*, *nequibam*, *nequibamus*, *nequibant*, *nequibas*, *nequibat*, *nequibatis*, *nequibimus*, *nequibis*, *nequibit*, *nequibitis*, *nequibo*, *nequibunt*, *nequiens*, *nequieram*, *nequieramus*, *nequierant*, *nequieras*, *nequierat*, *nequieratis*, *nequiere*, *nequierim*, *nequierimus*, *nequierint*, *nequieris*, *nequierit*, *nequieritis*, *nequiero*, *nequierunt*, *nequii*, *nequiimus*, *nequiit*, *nequimus*, *nequire*, *nequirem*, *nequiremus*, *nequirent*, *nequires*, *nequiret*, *nequiretis*, *nequis*, *nequisse*, *nequissem*, *nequissemus*, *nequissent*, *nequisses*, *nequisset*, *nequissetis*, *nequisti*, *nequistis*, *nequit*, *nequite*, *nequitis*, *nequito*, *nequitote*, *nequitu*, *nequitum*, *nequiturus*, *nequivi*, *nequivisti*, *nequivit*, *nihil*, *nihila*, *nihili*, *nihilis*, *nihilo*, *nihilorum*, *nihilum*, *nil*, *nimie*, *nimis*, *nimium*, *nisi*, *nobis*, *nobiscum*, *nolam*, *nolebam*, *nolebamus*, *nolebant*, *nolebas*, *nolebat*, *nolebatis*, *nolemus*, *nolens*, *nolent*, *noles*, *nolet*, *noletis*, *noli*, *nolim*, *nolimus*, *nolint*, *nolis*, *nolit*, *nolite*, *nolitis*, *nolito*, *nolitote*, *nolle*, *nollem*, *nollemus*, *nollent*, *nolles*, *nollet*, *nolletis*, *nolo*, *nolueram*, *nolueramus*, *noluerant*, *nolueras*, *noluerat*, *nolueratis*, *noluere*, *noluerim*, *noluerimus*, *noluerint*, *nolueris*, *noluerit*, *nolueritis*, *noluero*, *noluerunt*, *nolui*, *noluimus*, *noluisse*, *noluissem*, *noluissemus*, *noluissent*, *noluisses*, *noluisset*, *noluissetis*, *noluisti*, *noluistis*, *noluit*, *nolumus*, *nolunt*, *nolunto*, *non*, *nondum*, *nonne*, *nonnulla*, *nonnullae*, *nonnullam*, *nonnullarum*, *nonnullas*, *nonnulle*, *nonnulli*, *nonnullis*, *nonnullius*, *nonnullo*, *nonnullorum*, *nonnullos*, *nonnullum*, *nonnullus*, *nonnumquam*, *nos*, *noster*, *nostra*, *nostrae*, *nostram*, *nostrarum*, *nostras*, *nostri*, *nostris*, *nostro*, *nostrorum*, *nostros*, *nostrum*, *nulla*, *nullae*, *nullam*, *nullarum*, *nullas*, *nulle*, *nulli*, *nullis*, *nullius*, *nullo*, *nullorum*, *nullos*, *nullum*, *nullus*, *num*, *numquam*, *nunc*, *nunquam*, *nusquam*, *o*, *ob*, *oblique*, *obtorque*, *olim*, *omne*, *omnem*, *omnes*, *omni*, *omnia*, *omnibus*, *omnino*, *omnis*, *omnium*, *optorque*, *paene*, *parum*, *pauca*, *paucae*, *paucam*, *paucarum*, *paucas*, *pauce*, *pauci*, *paucior*, *pauciora*, *pauciore*, *pauciorem*, *pauciores*, *pauciori*, *paucioribus*, *paucioris*, *pauciorum*, *paucis*, *paucissima*, *paucissimae*, *paucissimam*, *paucissimarum*, *paucissimas*, *paucissime*, *paucissimi*, *paucissimis*, *paucissimo*, *paucissimorum*, *paucissimos*, *paucissimum*, *paucissimus*, *paucius*, *pauco*, *paucorum*, *paucos*, *paucum*, *paucus*, *penes*, *penitus*, *per*, *peraeque*, *plenisque*, *plerumque*, *plura*, *plure*, *plures*, *pluribus*, *plurima*, *plurimae*, *plurimam*, *plurimarum*, *plurimas*, *plurime*, *plurimi*, *plurimis*, *plurimo*, *plurimorum*, *plurimos*, *plurimum*, *plurimus*, *pluris*, *plurium*, *plus*, *pone*, *posse*, *possem*, *possemus*, *possent*, *posses*, *posset*, *possetis*, *possim*, *possimus*, *possint*, *possis*, *possit*, *possitis*, *possum*, *possumus*, *possunt*, *post*, *postea*, *potens*, *potente*, *potentem*, *potentes*, *potenti*, *potentia*, *potentibus*, *potentior*, *potentiora*, *potentiore*, *potentiorem*, *potentiores*, *potentiori*, *potentioribus*, *potentioris*, *potentiorum*, *potentis*, *potentissima*, *potentissimae*, *potentissimam*, *potentissimarum*, *potentissimas*, *potentissime*, *potentissimi*, *potentissimis*, *potentissimo*, *potentissimorum*, *potentissimos*, *potentissimum*, *potentissimus*, *potentium*, *potentius*, *poteram*, *poteramus*, *poterant*, *poteras*, *poterat*, *poteratis*, *potere*, *poterimus*, *poteris*, *poterit*, *poteritis*, *potero*, *poterunt*, *potes*, *potest*, *potestis*, *potueram*, *potueramus*, *potuerant*, *potueras*, *potuerat*, *potueratis*, *potuere*, *potuerim*, *potuerimus*, *potuerint*, *potueris*, *potuerit*, *potueritis*, *potuero*, *potuerunt*, *potui*, *potuimus*, *potuisse*, *potuissem*, *potuissemus*, *potuissent*, *potuisses*, *potuisset*, *potuissetis*, *potuisti*, *potuistis*, *potuit*, *prae*, *præ*, *praeter*, *præter*, *praetorque*, *prius*, *pro*, *prope*, *propter*, *qua*, *quacum*, *quadam*, *quae*, *quaedam*, *quaelibet*, *quaequam*, *quaeque*, *quaevis*, *quale*, *qualem*, *quales*, *quali*, *qualia*, *qualibus*, *qualis*, *qualium*, *quam*, *quamquam*, *quamque*, *quandam*, *quando*, *quandoque*, *quanta*, *quantae*, *quantam*, *quantarum*, *quantas*, *quante*, *quanti*, *quantis*, *quanto*, *quantorum*, *quantos*, *quantum*, *quantus*, *quanvis*, *quaque*, *quare*, *quarum*, *quarumquam*, *quarumque*, *quarundam*, *quarunvis*, *quas*, *quasdam*, *quasi*, *quasquam*, *quasque*, *quasvis*, *quavis*, *queam*, *queamus*, *queant*, *queas*, *queat*, *queatis*, *queatur*, *quem*, *quemquam*, *quemque*, *quendam*, *quenvis*, *queo*, *queundi*, *queundo*, *queundum*, *queundus*, *queunt*, *queunto*, *qui*, *quia*, *quibam*, *quibamus*, *quibant*, *quibas*, *quibat*, *quibatis*, *quibatur*, *quibimus*, *quibis*, *quibit*, *quibitis*, *quibitur*, *quibo*, *quibunt*, *quibus*, *quibuscum*, *quibusdam*, *quibusquam*, *quibusque*, *quibusvis*, *quicum*, *quicumque*, *quid*, *quidam*, *quiddam*, *quidem*, *quidlibet*, *quidquam*, *quidque*, *quidvis*, *quiens*, *quieram*, *quieramus*, *quierant*, *quieras*, *quierat*, *quieratis*, *quiere*, *quierim*, *quierimus*, *quierint*, *quieris*, *quierit*, *quieritis*, *quiero*, *quierunt*, *quii*, *quiimus*, *quiit*, *quilibet*, *quimus*, *quin*, *quiquam*, *quique*, *quire*, *quirem*, *quiremus*, *quirent*, *quires*, *quiret*, *quiretis*, *quiretur*, *quiri*, *quis*, *quisnam*, *quisquam*, *quisque*, *quisquis*, *quisse*, *quissem*, *quissemus*, *quissent*, *quisses*, *quisset*, *quissetis*, *quisti*, *quistis*, *quit*, *quite*, *quitis*, *quito*, *quitote*, *quitu*, *quitum*, *quitur*, *quiturus*, *quivi*, *quivis*, *quivisti*, *quivit*, *quo*, *quoad*, *quocum*, *quod*, *quodam*, *quodlibet*, *quomadmodum*, *quominus*, *quomodo*, *quonam*, *quondam*, *quoniam*, *quoquam*, *quoque*, *quorsum*, *quorum*, *quorumquam*, *quorumque*, *quorundam*, *quorunvis*, *quos*, *quosdam*, *quosquam*, *quosque*, *quosvis*, *quot*, *quotusquisque*, *quousque*, *quovis*, *recoque*, *retorque*, *rursus*, *saepe*, *saepissime*, *saepius*, *saltem*, *sane*, *se*, *secum*, *secundum*, *sed*, *semper*, *sese*, *si*, *sibi*, *sic*, *sicut*, *sim*, *simul*, *simus*, *sine*, *sint*, *sis*, *sit*, *sitis*, *sive*, *solum*, *statim*, *sua*, *suae*, *suam*, *suarum*, *suas*, *sub*, *subter*, *sue*, *sui*, *suis*, *sum*, *sumus*, *sunt*, *sunto*, *suo*, *suorum*, *suos*,

super, *supra*, *susque*, *suum*, *suus*, *tale*, *talem*, *tales*, *tali*, *talia*, *talibus*, *talis*, *talium*, *tam*, *tamen*, *tametsi*, *tandem*, *tanta*, *tantae*, *tantam*, *tantarum*, *tantas*, *tante*, *tanti*, *tantis*, *tanto*, *tantorum*, *tantos*, *tantum*, *tantus*, *te*, *tecum*, *tenus*, *tibi*, *torque*, *tot*, *tota*, *totae*, *totam*, *totarum*, *totas*, *tote*, *toti*, *totis*, *totius*, *toto*, *totorum*, *totos*, *totum*, *totus*, *trans*, *tu*, *tua*, *tuae*, *tuam*, *tuarum*, *tuas*, *tue*, *tui*, *tuis*, *tum*, *tuo*, *tuorum*, *tuos*, *tuum*, *tuus*, *ubi*, *ubicumque*, *ubinam*, *ubique*, *ubiubi*, *uel*, *uero*, *ultra*, *umquam*, *una*, *unae*, *unam*, *unarum*, *unas*, *unde*, *undique*, *une*, *uni*, *unis*, *unius*, *uno*, *unorum*, *unos*, *unquam*, *unum*, *unus*, *usquam*, *usque*, *ut*, *uter*, *uterque*, *utique*, *utra*, *utrae*, *utram*, *utrarum*, *utras*, *utri*, *utribique*, *utris*, *utrius*, *utro*, *utroque*, *utrorum*, *utros*, *utrum*, *valde*, *vel*, *versum*, *versus*, *vester*, *vestra*, *vestrae*, *vestram*, *vestrarum*, *vestras*, *vestri*, *vestris*, *vestro*, *vestrorum*, *vestros*, *vestrum*, *vix*, *vobis*, *vobiscum*, *vos*, *voster*,

or is spelt as a word from the list above, followed by suffix *que*, then we consider it a Latin stop word. All the Latin stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

Similar to what we did in the algorithm for French, we need to define string patterns **goLatin**, **goodLatin**, **betterLatin**, **bestLatin** to cover some irregular inflections:

eam, *eamus*, *eant*, *eas*, *eat*, *eatis*, *eatur*, *eo*, *eundi*, *eundo*, *eundum*, *eundus*, *eunt*, *eunto*, *i*, *ibam*, *ibamus*, *ibant*, *ibas*, *ibat*, *ibatis*, *ibatur*, *ibimus*, *ibis*, *ibit*, *ibitis*, *ibitur*, *ibo*, *ibunt*, *iens*, *ieram*, *ieramus*, *ierant*, *ieras*, *ierat*, *ieratis*, *iere*, *ierim*, *ierimus*, *ierint*, *ieris*, *ierit*, *ieritis*, *iero*, *ierunt*, *ii*, *iimus*, *iit*, *imus*, *ire*, *irem*, *iremus*, *irent*, *ires*, *iret*, *iretis*, *iretur*, *iri*, *is*, *isse*, *issem*, *issemus*, *issent*, *isses*, *isset*, *issetis*, *isti*, *istis*, *it*, *ite*, *itis*, *ito*, *itote*, *itu*, *itum*, *itur*, *iturus*, *itus*, *ivi*, *ivisti*, *ivit* — "go";

bona, *bonae*, *bonam*, *bonarum*, *bonas*, *bone*, *boni*, *bonis*, *bono*, *bonorum*, *bonos*, *bonum*, *bonus* — "good";

melior, *meliora*, *meliore*, *meliorem*, *meliores*, *meliori*, *melioribus*, *melioris*, *meliorum*, *melius* — "better";

optima, *optimae*, *optimam*, *optimarum*, *optimas*, *optime*, *optimi*, *optimis*, *optimo*, *optimorum*, *optimos*, *optimum*, *optimus*, *optuma*, *optumae*, *optumam*, *optumarum*, *optumas*, *optume*, *optumi*, *optumis*, *optumo*, *optumorum*, *optumos*, *optumum*, *optumus* — "best".

However, we are not going to treat these strings as stop words[80], so as to be consistent with their counterparts in other languages.

6.3.1 Effective spelling and essential root

Definition 6.22 (Latin Vowel Extensions)**.** Hereafter in §6.3, the symbol $\mathbf{V}^*$ stands for any member from the list {*a*, *e*, *i*, *o*, *u*, *y*}, the so-called Latin vowel extensions. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}^*_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Latin vowel extensions.

Dual to the notations above, the symbol $\mathbf{C}^*$ stands for any character that does not belong to the list {*a*, *e*, *i*, *o*, *u*, *y*}, and $\mathbf{C}^*_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list. □

Definition 6.23 (Latin protected range)**.** Let $\hat{\sigma}$ be a text string derived from a Latin word, its protected range ProtRg($\hat{\sigma}$) is an integer determined as follows:

- Try to find the string pattern $(\varnothing|a(b|d|mb)|circum|co(m|\mu|n|rr)|de(f|l|t)|i|jnter|o(b|f)|per|pr(ae|e|o)|re|su|tran)\mathbf{C}^*_{m_0}\mathbf{V}^*\mathbf{C}^*_{m_0}\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines ProtRg($\hat{\sigma}$); otherwise, set ProtRg($\hat{\sigma}$) = 0. □

Algorithm 6.24 (Latin effective spelling)**.** *For a Latin word $\hat{\sigma}$, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in sequential steps:*

(1) *Do*[81] $\bar{a}\rightarrow a$, $æ\rightarrow ae$, $\bar{e}\rightarrow e$, $\bar{\imath}\rightarrow i$, $\bar{o}\rightarrow o$, $œ\rightarrow oe$, $\bar{u}\rightarrow u$, $\bar{y}\rightarrow y$.

(2) *Do carolin*$\sim\rightarrow$ *καρoliν*, *emi*$(s|ss|t|tt)\rightarrow$ *εμισ*, *impr*$\sim\rightarrow$ *mpρ*, *pass*$\rightarrow$ *pand*, *pater*$\sim\rightarrow$ *patre*, *separ*$\sim\rightarrow$ *σεπαρ*, *vac*$\rightarrow$ *βεac*.

[80] Nevertheless, we have included *eam*, *eo*, *is*, *iste* and *istis* in our list of Latin stop words, because they not only function as conjugated forms of the Latin verb for "to go", but also as pronouns.

[81] Note that the italic forms of æ (*æ*) and œ (*œ*) are very similar.

(3) *Replace*

(∅|*con*)*sobrin***X**	(∅|*cu*)*cur*$^{\mathbf{X}\in}$(*r*|*s*)	(*amit*|*materter*)$^{\mathbf{X}\in}$(*a*|*e*|*i*|*o*|*u*|*y*)~	(*m*|*p*)*atruel***X**
κuσν͂	*κur***X**	*αuν͂τ***X**	*κuσν͂*

aestim~	*ami*(*s*|*tt*)(∅|*s*)~	*amiciss***X**~	*asp*~	*camer*~	*car*(*a*|*i*|*o*)~	*collin*~	*color*~	*delect*~	*dign*~
αστiμ	*αμiσ*	*amica*	*αspa*	*ρuμ*	*δεαρi*	*κολλiν͂*	*φαρβr*	*δλiγτ*	*δiγn*

divid~	*divin*~	*er*(*ect*|*ex*|*ig*)~	*fem*(*ell*|*in*)~	*haer*~	*here*(*d*|*s*)~	*hestern*~	*honest*~	*horta*~
diviδ	*divν͂*	*ερεξt*	*φεμλn*	*ηaer*	*ηeiρ*	*ηεστ*	*ηoneστ*	*ηoρta*

hr~	*ingenu*~	*laet*~	*lisul*~	*locut*	*loq*	*lucas*~	*marit*~	*materi*~	*matron*~	*matur*~	*matut*~
μiστer	*inγeν͂u*	*χαππt*	*λiζζ*	*lλoquor*	*lλoq*	*λuξσ*	*μαρτ*	*ματρi*	*λαδη*	*μτuρ*	*ματiν͂*

memin~	*met*~	*mra*~	*nupt*~	*opin*~	*pact*~	*pag*~	*pet*~	*pig*~	*rheda*~	*salt***X**~	*salut*~	*san*~	*sapien*~
μν͂in	*φεαρ*	*μρσa*	*ν͂uπt*	*ωpin*	*pang*	*paγ*	*πετ*	*πiγ*	*ωoγν͂a*	*δaν͂σ*	*σαλut*	*σaν͂*	*ωiσn*

scri(*b*|*ps*|*pt*)~	*senior*~	*sign*~	*soror*~	*tem*~	*van*~	*ven*(*a*|*o*)~	*vene*(*m*|*t*)~	*vidu*~
ωριτr	*σγν͂ορ*	*siγn*	*σορορ*	*τme*	*βaν͂*	*veν͂a*	*veν͂at*	*ωiδo*

d(*e*(*e*|*i*|*is*|*o*|*orum*|*os*|*um*|*us*))|*i*(∅|*i*|*is*|*s*	*amic*(*i*|*u*)(∅|*m*|*s*)	*car*(*e*|*um*|*us*)	*die*	*heri*	*vener*(∅|*e*)
γoδ	*amica*	*δεαρi*	*diebus*	*ηεστ*	*veν͂atur*

(4) *Replace*

δεαρil	*χ̂tri*(*c*|*x*)	*a*(*dfer*|*llat*|*ttul*)~	*ab*(*lat*|*tul*)~	*amic*(*i*(*a*|*e*|*m*|*r*|*te*|*to*)|*u*|*t*)~				
καrλ	*χ̂tor*	*affer*	*aufer*	*aμβicu*				

bib~	*co*(*llat*|*ntul*)~	*complement*	*concors*	*creat*~	*dea*~	*di*(*lat*|*stul*)~	*diurn*~
biβ	*confer*	*compler*	*concord*	*crear*	*deαu*	*differ*	*di*

domin($\overline{a}$)~	*domina*~	*e*(*lat*|*xtul*)~	*expos*~	*filia*~	*hort*~	*ira*(*sc*|*t*)~	*latrin*~	*matrimon*~
doμiν͂o	*domiν͂αa*	*effer*	*expon*	*filiαa*	*γaρδt*	*ir*	*lav*	*μatrimon*

memor~	*mens*$^{\mathbf{X}\in}$(*a*|*e*|*i*|*o*|*u*|*y*)~	*ob*(*lat*|*tul*)~	*oneros***X**~	*patrimon*~	*patru*$_{\hat{\chi}\notin}$(*m*)**X**~	*pin*(*g*|*x*)~	*reg*~
μem	*menσ***X**	*offer*	*oneris*	*πatrimon*	*avunculus*	*pict*	*ρeg*

sci(*am*|*ar*|*eb*|*em*|*en*|*es*|*et*|*m*|*o*|*re*|*ri*|*sc*|*t*|*un*|*v*)**X**	*str*	*su*(*blat*|*stul*)~	*terter*
sci	*starr*	*suffer*	*tertερ*

*truculent***X**~	*tutel*~	*us*(*i*|*u*)~	*via*~	*xt*	(**bestLatin**|**betterLatin**|**goodLatin**)
trucum	*tutar*	*uto*	*viαa*	*x*	*λgoodλ*

$^{\mathbf{X}\in}$(∅|*ab*|*ad*|*amb*|*circum*|*co*|*de*|*in*|*inter*|*ob*|*prod*|*red*|*suss*|*trans*)**goLatin**
X*λgoλ*

cor	*d*(*ee*|*ei*|*eis*|*eo*|*eorum*|*eos*|*eum*|*eus*|*ia*|*iae*|*iam*|*iarum*|*ias*|*ie*|*ii*|*iis*|*io*|*iorum*|*ios*|*is*|*ium*|*ius*)
cordis	*divus*

lux	*mens*	*mos*	*nix*	*ops*	*re*	*trux*	*vi*(∅|*m*)	~*que*	~*scis*
lucum	*mentis*	*moris*	*nivis*	*opum*	*rebus*	*trucum*	*viribus*	*'que*	*sci*

(5) *Do* $a^{\mathbf{X}\in}$(*cc*\|*dd*\|*ff*\|*gg*\|*rr*\|*ss*\|*tt*)~ → *α***X**, *x* → *č*.

(6) *Replace*

$_{\hat{\chi}\notin}$(*m*)*om*	$^{\mathbf{X}\in}$(*a*|*e*|*i*|*o*|*r*|*u*|*y*)*p*(*s*|*t*)	$^{\mathbf{X}\in}$($\hat{\chi}$($\overline{a\|e\|i\|o\|r\|u\|y}$))*r*	*cantor*~	*dele*~
χ̂oμ	**X***p*	**X***erque*	*canere*	*ddele*

mpt	*os*$\hat{\chi}$	*patr*~	*puls*	*tang*~	*tarrid*	*tr*$_{\hat{\chi}\notin}$(*a*)~	*trah*	*ult*	*ire*(*m*|*mus*|*s*)	*pater*	*vis*	~*emen*(*d***X**|*s*	~*ster*
m	*oσ*$\hat{\chi}$	*pαatr*	*pell*	*tact*	*tarriz*	*ter*$\hat{\chi}$	*tract*	*ol*	*eo*	*pαater*	*vol*	*em*	*starr*

(7) *Do* ~*rs* → *rt*, ~$^{\hat{\chi}\in\overline{(a|e|i|o|r|t|u)}}$*er* → $\hat{\chi}$*erquei*, *inter*~ → *jnter*, *super*~ → *suzp*.

(8) *Remove apostrophe, and call the string obtained so far as* $\hat{\sigma}_\sharp$.

(9) *Break down* $\hat{\sigma}_\sharp = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}_\sharp^{[\mathrm{ProtRg}(\hat{\sigma}_\sharp)]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma}_\sharp)$ *is equal to the protected range of* $\hat{\sigma}_\sharp$*.*

(10) *Do act* $\to$ *ag, erb* $\to$ *erv, eg*$\sim\to$*ag,* $^{\mathbf{X}\in}(i|o|u)v \to \mathbf{X}t$*, ub* $\to$ *uss on* $\hat{\sigma}_1$ *and call the result* $\hat{\sigma}'_1$*.*

(11) *Do* $\sim$*que*$\to$ *'que on* $\hat{\sigma}_2$ *and call the result* $\hat{\sigma}'_2$*.*

(12) *If the pattern* $\sim\mathbf{C}^*\mathbf{V}^*\mathbf{C}^*$ *is found in* $\hat{\sigma}'_1$ *and the two occurrences of* $\mathbf{C}^*$ *represent the same letter, then remove the last letter from* $\hat{\sigma}'_1$*, and do* $\mathbf{V}^*\sim\to\varnothing$ *on* $\hat{\sigma}'_2$*.*[82] *Call the results after these operations* $\hat{\sigma}''_1$ *and* $\hat{\sigma}''_2$*, respectively.*

(13) *On* $\hat{\sigma}''_2$*, perform the following substitutions in a sequel:*

(10.1) *Do* $_{\hat{\chi}\notin}(c|t)ul\to\hat{\chi}$*,* $(cul|lent|mon)\to\varnothing$*,* $(errim|issim)\to\varnothing$*,* $men(\varnothing|t)\to\varnothing$*.*

(10.2) *Do* $\sim(m|ni|nt(\varnothing|\mathbf{V}^*)(\varnothing|r)|r\mathbf{V}^*|s)\to\varnothing$*.*

(10.3) *Do* $\sim\hat{\chi}(mi|mu|mur|ntur|sti|ri)\to\hat{\chi}$*,* $_{\hat{\chi}\notin}(r)t\to\hat{\chi}$*.*

The result after these three steps of operations is called $\hat{\sigma}'''_2$*.*

(14) *Concatenate* $\hat{\sigma}''_1$ *and* $\hat{\sigma}'''_2$*.*

(15) *Do die*$(bu|m|ru|s)\to$*dieque, ign*$\to$*en,* $\underline{re(bu|m|rum|s)}\to$*reque, ct*$\to$*č, pro*$\sigma\sim\to$*pros,* $\mathbf{V}^*men(\varnothing|t)\to\mathbf{V}^*$*,* $'\to\varnothing$*.*

Algorithm 6.25 (Latin essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Latin word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$*.*

(2) *Do* $\mathbf{C}^*(a|ae)\mathbf{C}^*\to\mathbf{C}^*e\mathbf{C}^*$ *on* $\hat{\sigma}_1$*,*[83] *and call the result* $\hat{\sigma}'_1$*.*

(3) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(3.1) *Do* $\sim$*que*$\to\varnothing$*.*

(3.2) *Do* $(b|id|nd|t|v)\mathbf{V}^*(\varnothing|m|n|r|(s)_m|t)\to\varnothing$*.*

(3.3) *Do* $\sim(a|e|i|in|o|r|ss|u)_m\to\varnothing$*.*

The result after these three steps of operations is called $\hat{\sigma}'_2$*.*

(4) *Concatenate* $\hat{\sigma}'_1$ *and* $\hat{\sigma}'_2$*.*

(5) *Replace*[84]

sens$\sim$	$\underline{^{\mathbf{X}\in}(\varnothing\|a\|circum\|con\|di\|ec\check{}\|in\|jnter\|ob\|per\|prae\|re\|suzp)st(a\|ae\|ans\|ant\|e\|eb\|em\|end\|ens\|ent\|er\|es\|ess\|est\|et\|o)}$						
sent	$\mathbf{X}stand$						
$\underline{^{\mathbf{X}\in}(\varnothing\|ab\|ad\|con\|in\|per\|prae\|pro\|red\|suss)d(a\|and\|ans\|ant\|e\|eb\|em\|end\|ens\|ent\|er\|erq\|es\|ess\|est\|et\|etis\|etur\|o\|or)}$							
$\mathbf{X}datt$							
$\underline{^{\mathbf{X}\in}(\varnothing\|ad\|af\|au\|circum\|con\|de\|dif\|in\|jnter(\varnothing\|queo)\|of\|prae\|pro\|re\|suf\|trans)fer(q\|r\|t)}$							
$\mathbf{X}fer$							
$\underline{^{\mathbf{X}\in}(\varnothing\|ad\|circum\|com\|de\|di\|e\|im\|jnter(\varnothing\|queo)\|o\|per\|prae\|pro\|re\|suss\|suzp\|trans)mis(\varnothing\|s)}$						$\underline{^{\mathbf{X}\in}(\varnothing\|co\mu\|prae\|pro)pass}$	
$\mathbf{X}mitt$						$\mathbf{XX}et$	
$\underline{ci(e\|eam\|eant\|ear\|eas\|eat\|eb\|em\|end\|ens\|ent\|eo\|eor\|er\|es\|et\|t}$			$\underline{es(\varnothing\|s\|t)}$	$\underline{i(\check{c}\|c)\mathbf{V}^*\mathbf{C}^*_{m_\phi}}$	$\underline{m(all\|el\|ev)}$	$\underline{noll}$	$\underline{v(el\|ell\|olt)}$
cit			*ed*	*ic*	*mevol*	*nol*	*vol*
$\sim(lev\|lit)$	$\sim aperien$	$\sim cur(r\|ren\|s)$	$\sim fund(\varnothing\|en)$	$\sim ges(s\|t)$	$\sim mess$	$\sim no\sigma c(\varnothing\|en)$	$\sim or(ien\|t)$
lin	*ap*	*cur*	*fus*	*ger*	*met*	*not*	*or*
$\sim p(ast\|ev)$	$\sim pel$	$\sim portan$	$\sim se(t\|v)$	$\sim starr(\varnothing\|d)$	$\sim vent$	$\sim vers$	$\sim vinc(\varnothing\|en)$
pasc	*pell*	*por*	*ser*	*stern*	*ven*	*vert*	*vic*
$\underline{^{\mathbf{X}\in}(\varnothing\|ab\|abs\|al\|at\|circum\|con\|de\|dis\|in\|jnter(\varnothing\|queo)\|ob\|prae\|pro\|re\|sub\|trans)(let\|tul)}$							
$\mathbf{X}'fer$							

[82] This peculiar step is tailored for reduplications in the perfect forms of certain Latin verbs. Such reduplications are not attested in most modern Indo-European languages, with Greek being a notable exception.

[83] Here, the two occurrences of $\mathbf{C}^*$ may or may not represent the same letter.

[84] All the patches below are devoted to some common verbs that exhibit highly irregular conjugations.

where in the last step, one constructs $\mathbf{X}'$ *from* $\mathbf{X}$ *by doing* $\sim ab(\varnothing|s) \to au$, $\sim(\check{c}|l|t) \to f$, *dis* $\to$ *dif, ob* $\to$ *of, suss* $\to$ *suf.*

(6) *Set*

$$\mathbf{X}_1 = (eam|eant|eas|eat|eo|eund|eunt|i|ib|ibant|ibem|ibes|ibet|ibim|ibis|ibit|ibo|ibunt|iens|ieq|ier|ii|iim|iit|im|imus|ir|ire|irent|ireq|iret|iri|is|iss|isse|issem|issent|isser|ist|isti|istis|it|ite|iti|itis|itist|ito|itu|itum|itur|itus),$$

and

$$\mathbf{X}_2 = (er|es|ess|est|for|fuer|fui|fuim|fuiss|fuist|fuit|fut|si|sim|sint|sit|sum|sumus|sunt|sunto).$$

Replace[85]

$$\frac{\mathbf{X}^{\in}(\varnothing|ab|ad|amb|circum|co|de|in|jnter(\varnothing|queo)|ob|prod|red|suss|trans)\mathbf{X}_1}{\mathbf{X}eo}$$

$$\frac{\mathbf{X}^{\in}(ab|ad|de|in|jnter|ob|prae|pro|suss|suzp)\mathbf{X}_2}{\mathbf{X}sum}$$

6.3.2 Admissible mutation and approximate clustering

Like the case of French, vowel blotting is unnecessary for Latin. The only significant pattern of vowel alternation in Latin verb conjugations has already been taken care of by Algorithm 6.25(2).

In what follows, we will construct a bivariate Boolean-valued function HrdTest($\hat{\alpha},\hat{\beta}$) on a "simple heredity test function" in Algorithm 6.26, and a set of "admissible suffix mismatch" rules in Algorithm 6.27.

Algorithm 6.26 (Simple heredity test)**.** *Let* $\hat{\alpha}'$ *be the result from doing* $\sim\mathbf{V}^*_{m_0}(\varnothing|erquei|que|s|st) \to \varnothing$ *on* $\hat{\alpha}$*, and define* $\hat{\beta}'$ *similarly. The Boolean-valued function* SimpHrdTest($\hat{\alpha},\hat{\beta}$) *returns* **TRUE** *if the lowercase form of* $\hat{\alpha}$ *contains at least one instance of* $\mathbf{V}^*$ *(Definition 6.22)* **AND** *at least one of the following three conditions holds:*[86]

(i) $\hat{\alpha}' = \hat{\beta}'$*;*

(ii) $\hat{\beta} = \hat{\alpha}t$*;*

(iii) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\alpha} = \hat{\beta}^{[\ell(\hat{\alpha})]}$ **AND** $\hat{\beta}^{\{\ell(\hat{\alpha}')+1\}} = \mathbf{V}^*$*. (See Definition 3.1 for the notations* $\hat{\beta}^{[n]}$ *and* $\hat{\beta}^{\{n\}}$*.)*

In what follows, we define SuffixNW($\hat{\alpha},\hat{\beta}$), RootNW($\hat{\alpha},\hat{\beta}$), NW*($\hat{\alpha},\hat{\beta}$) and SuffixSW($\hat{\alpha},\hat{\beta}$), SuffixSW($\hat{\alpha},\hat{\beta}$), SW*($\hat{\alpha},\hat{\beta}$) as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5.

Algorithm 6.27 (Admissible suffix mismatch)**.** *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function*

$$\text{AdmSM}(\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{SuffixNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if* NW*($\hat{\alpha},\hat{\beta}$) $= \varnothing$ **AND** RootNW($\hat{\alpha},\hat{\beta}$) *contains at least one instance of* $\mathbf{V}^*$ **AND** *at least one of the following four conditions holds:*

(i) SuffixNW($\hat{\alpha},\hat{\beta}$) $= [(\varnothing|(a|e|i|o|u)_m(\varnothing|n)), (\varnothing|(a|e|i|o|u)_m(\varnothing|n))]$*;*

(ii) SuffixNW($\hat{\alpha},\hat{\beta}$) $= [c,\check{c}]|[\check{c},g]|[c,q]|[d,s]|[den,s]$*;*

(iii) SuffixNW($\hat{\alpha},\hat{\beta}$) $= [d,t]$ **AND** Ω(RootNW($\hat{\alpha},\hat{\beta}$)) $= r$*;*

(iv) SuffixNW($\hat{\alpha},\hat{\beta}$) $= [\check{c}, c\mathbf{V}^*\mathbf{X}]$.

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmSM}(\text{RootSW}(\hat{\alpha},\hat{\beta}), \text{SuffixSW}(\hat{\alpha},\hat{\beta}), \text{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 6.28 (Heredity test function)**.** *The structure of the Latin heredity test function* HrdTest($\hat{\alpha},\hat{\beta}$) *is identical to the German version (Algorithm 8.1.2), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW**,* RootSW*,* SuffixSW*,* SW* *must follow the Latin rules stated above.*

[85] All the patches below are devoted to suppletive verbs derived from *eō* "go" and *sum* "be".

[86] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 6.29 (Approximate clustering of Latin words)**.** *The algorithm is essentially the same as Algorithm 6.20, except that Latin rules (instead of French rules) apply to all the tags (effective spelling, essential root etc.).*

Example 6.29.1. There are five declensions in Latin. The first three declensions cover both nouns and adjectives, while the last two are reserved for nouns only. We extract sample nouns and adjectives from the following Wiktionary and Wikipedia pages:

`https://en.wiktionary.org/wiki/Appendix:Latin_first_declension`

comētae, *comētārum*, *comētās*, *comētē*, *comētem*, *comētēn*, *comētēs*, *comētīs* — "comet";

dynastae, *dynastārum*, *dynastās*, *dynastē*, *dynastēn*, *dynastēs*, *dynastīs* — "ruler";

nauta, *nautā*, *nautae*, *nautam*, *nautārum*, *nautās*, *nautīs* — "sailor";

nymphae, *nymphārum*, *nymphās*, *nymphē*, *nymphēn*, *nymphēs*, *nymphīs* — "bride";

Rōma, *Rōmā*, *Rōmae*, *Rōmam* — "Rome";

rosa, *rosā*, *rosae*, *rosam*, *rosārum*, *rosās*, *rosīs* — "rose";

stēlla, *stēllā*, *stēllae*, *stēllam*, *stēllārum*, *stēllās*, *stēllīs* — "star";

xiphiā, *xiphiae*, *xiphiān*, *xiphiārum*, *xiphiās*, *xiphiīs* — "swordfish".

`https://en.wiktionary.org/wiki/Appendix:Latin_second_declension`

ager, *agrī*, *agrīs*, *agrō*, *agrōrum*, *agrōs*, *agrum* — "field";

ampele, *ampelī*, *ampelīs*, *ampelō*, *ampelon*, *ampelōrum*, *ampelos*, *ampelōs*, *ampelum* — "vine";

atome, *atomī*, *atomīs*, *atomō*, *atomōrum*, *atomōs*, *atomum*, *atomus* — "atom";

bella, *bellī*, *bellīs*, *bellō*, *bellōrum*, *bellum* — "war";

fīlī, *fīliī*, *fīliīs*, *fīliō*, *fīliōrum*, *fīliōs*, *fīlium*, *fīlius* — "son";

magister, *magistrī*, *magistrīs*, *magistrō*, *magistrōrum*, *magistrōs*, *magistrum* — "teacher";

mūre, *mūrī*, *mūrīs*, *mūrō*, *mūrōrum*, *mūrōs*, *mūrum*, *mūrus* — "wall";

mȳthe, *mȳthī*, *mȳthīs*, *mȳthō*, *mȳthon*, *mȳthōrum*, *mȳthos*, *mȳthōs*, *mȳthum* — "myth";

phaenomena, *phaenomenī*, *phaenomenīs*, *phaenomenō*, *phaenomenon*, *phaenomenōrum* — "phenomenon";

puer, *puerī*, *puerīs*, *puerō*, *puerōrum*, *puerōs*, *puerum* — "child";

templa, *templī*, *templīs*, *templō*, *templōrum*, *templum* — "temple".

`https://en.wiktionary.org/wiki/Appendix:Latin_third_declension`

āēr, *āera*, *āere*, *āerem*, *āerēs*, *āerī*, *āeribus*, *āeris*, *āeros*, *āerum* — "air";

animal, *animālī*, *animālia*, *animālibus*, *animālis*, *animālium* — "animal";

base, *basem*, *basēs*, *basī*, *basibus*, *basim*, *basis*, *basīs*, *basium* — "pedestal";

haerese, *haeresem*, *haeresēs*, *haeresī*, *haeresibus*, *haeresim*, *haeresis*, *haeresīs*, *haeresium* — "sect";

homine, *hominem*, *hominēs*, *hominī*, *hominibus*, *hominis*, *hominum*, *homō* — "human";

nocte, *noctem*, *noctēs*, *noctī*, *noctibus*, *noctis*, *noctium*, *nox* — "night";

nōmen, *nōmina*, *nōmine*, *nōminī*, *nōminibus*, *nōminis*, *nōminum* — "name";

tigre, *tigrem*, *tigrēs*, *tigrī*, *tigribus*, *tigride*, *tigridem*, *tigridēs*, *tigridī*, *tigridibus*, *tigridis*, *tigridum*, *tigrim*, *tigris*, *tigrīs*, *tīgris*, *tigrium* — "tiger";

turre, *turrem*, *turrēs*, *turrī*, *turribus*, *turrim*, *turris*, *turrīs*, *turrium* — "tower".

`https://en.wiktionary.org/wiki/Appendix:Latin_fourth_declension`

cornibus, *cornū*, *cornua*, *cornūs*, *cornuum* — "horn";

Dīdō, *Dīdōne*, *Dīdōnem*, *Dīdōnēs*, *Dīdōnī*, *Dīdōnibus*, *Dīdōnis*, *Dīdōnum* — "Dido";

ēchibus, *ēcho*, *ēchū*, *ēchuī*, *ēchum*, *ēchūs*, *ēchuum* — "echo";

manibus, *manū*, *manuī*, *manum*, *manus*, *manūs*, *manuum* — "hand".

`https://en.wiktionary.org/wiki/Appendix:Latin_fifth_declension`

diē, *diēbus*, *diēī*, *diem*, *diērum*, *diēs* — "day";

fidē, *fidēbus*, *fideī*, *fidem*, *fidērum*, *fidēs* — "faith";

rē, *rēbus*, *reī*, *rem*, *rērum*, *rēs* — "thing";

speciē, *speciēbus*, *speciēī*, *speciem*, *speciērum*, *speciēs* — "view".

`https://en.wikipedia.org/wiki/Latin_declension#Adjectives`

Here are some Latin adjectives whose positive forms belong to the first and second declensions. (We also incorporate, in our list below, comparative and superlative forms of certain sample adjectives, which may or may not share the same declension pattern with the original adjectives in positive forms.)

alta, *altā*, *altae*, *altam*, *altārum*, *altās*, *alte*, *altī*, *altior*, *altiōra*, *altiōre*, *altiōrem*, *altiōrēs*, *altiōrī*, *altiōribus*, *altiōris*, *altiōrum*, *altīs*, *altissima*, *altissimā*, *altissimae*, *altissimam*, *altissimārum*, *altissimās*, *altissime*, *altissimī*, *altissimīs*, *altissimō*, *altissimōrum*, *altissimōs*, *altissimum*, *altissimus*, *altius*, *altō*, *altōrum*, *altōs*, *altum*, *altus* — "tall";

atoma, *atomā*, *atomae*, *atomam*, *atomārum*, *atomās*, *atome*, *atomī*, *atomīs*, *atomō*, *atomōrum*, *atomōs*, *atomum*, *atomus* — "indivisible";

miser, *misera*, *miserā*, *miserae*, *miseram*, *miserārum*, *miserās*, *miserī*, *miserior*, *miseriōra*, *miseriōre*, *miseriōrem*, *miseriōrēs*, *miseriōrī*, *miseriōribus*, *miseriōris*, *miseriōrum*, *miserīs*, *miserius*, *miserō*, *miserōrum*, *miserōs*, *miserrima*, *miserrimā*, *miserrimae*, *miserrimam*, *miserrimārum*, *miserrimās*, *miserrime*, *miserrimī*, *miserrimīs*, *miserrimō*, *miserrimōrum*, *miserrimōs*, *miserrimum*, *miserrimus*, *miserum* — "poor";

sacer, *sacra*, *sacrā*, *sacrae*, *sacram*, *sacrārum*, *sacrās*, *sacrī*, *sacrīs*, *sacrō*, *sacrōrum*, *sacrōs*, *sacrum* — "sacred";

ūlla, *ūllā*, *ūllae*, *ūllam*, *ūllārum*, *ūllās*, *ūlle*, *ūllī*, *ūllīs*, *ūllīus*, *ūllō*, *ūllōrum*, *ūllōs*, *ūllum*, *ūllus* — "any".

Here are some Latin adjectives whose positive forms belong to the third declension.

agile, *agilem*, *agilēs*, *agilī*, *agilia*, *agilibus*, *agilis*, *agilium* — "agile";

alacer, *alacre*, *alacrem*, *alacrēs*, *alacrī*, *alacria*, *alacribus*, *alacris*, *alacrium* — "lively";

atrōcem, *atrōcēs*, *atrōcī*, *atrōcia*, *atrōcibus*, *atrōcior*, *atrōciōra*, *atrōciōre*, *atrōciōrem*, *atrōciōrēs*, *atrōciōrī*, *atrōciōribus*, *atrōciōris*, *atrōciōrum*, *atrōcis*, *atrōcissima*, *atrōcissimā*, *atrōcissimae*, *atrōcissimam*, *atrōcissimārum*, *atrōcissimās*, *atrōcissime*, *atrōcissimī*, *atrōcissimīs*, *atrōcissimō*, *atrōcissimōrum*, *atrōcissimōs*, *atrōcissimum*, *atrōcissimus*, *atrōcium*, *atrōcius*, *atrōx* — "fierce";

celer, *celere*, *celerem*, *celerēs*, *celerī*, *celeria*, *celeribus*, *celerior*, *celeriōra*, *celeriōre*, *celeriōrem*, *celeriōrēs*, *celeriōrī*, *celeriōribus*, *celeriōris*, *celeriōrum*, *celeris*, *celerium*, *celerius*, *celerrima*, *celerrimā*, *celerrimae*, *celerrimam*, *celerrimārum*, *celerrimās*, *celerrime*, *celerrimī*, *celerrimīs*, *celerrimō*, *celerrimōrum*, *celerrimōs*, *celerrimum*, *celerrimus* — "fast";

melior, *meliōra*, *meliōre*, *meliōrem*, *meliōrēs*, *meliōrī*, *meliōribus*, *meliōris*, *meliōrum*, *melius* — “better”;

trīste, *trīstem*, *trīstēs*, *trīstī*, *trīstia*, *trīstibus*, *trīstior*, *trīstiōra*, *trīstiōre*, *trīstiōrem*, *trīstiōrēs*, *trīstiōrī*, *trīstiōribus*, *trīstiōris*, *trīstiōrum*, *trīstis*, *trīstissima*, *trīstissimā*, *trīstissimae*, *trīstissimam*, *trīstissimārum*, *trīstissimās*, *trīstissime*, *trīstissimī*, *trīstissimīs*, *trīstissimō*, *trīstissimōrum*, *trīstissimōs*, *trīstissimum*, *trīstissimus*, *trīstium*, *trīstius* — “unhappy”;

vetera, *vetere*, *veterem*, *veterēs*, *veterī*, *veteribus*, *veteris*, *veterrima*, *veterrimā*, *veterrimae*, *veterrimam*, *veterrimārum*, *veterrimās*, *veterrime*, *veterrimī*, *veterrimīs*, *veterrimō*, *veterrimōrum*, *veterrimōs*, *veterrimum*, *veterrimus*, *veterum*, *vetus*, *vetustior*, *vetustiōra*, *vetustiōre*, *vetustiōrem*, *vetustiōrēs*, *vetustiōrī*, *vetustiōribus*, *vetustiōris*, *vetustiōrum*, *vetustissima*, *vetustissimā*, *vetustissimae*, *vetustissimam*, *vetustissimārum*, *vetustissimās*, *vetustissime*, *vetustissimī*, *vetustissimīs*, *vetustissimō*, *vetustissimōrum*, *vetustissimōs*, *vetustissimum*, *vetustissimus*, *vetustius* — “old”.

Our clustering algorithm yields the following result:

{*āēr*, *āera*, *āere*, *āerem*, *āerēs*, *āerī*, *āeribus*, *āeris*, *āeros*, *āerum*},

{*ager*, *agrī*, *agrīs*, *agrō*, *agrōrum*, *agrōs*, *agrum*},

{*agile*, *agilem*, *agilēs*, *agilī*, *agilia*, *agilibus*, *agilis*, *agilium*},

{*alacer*, *alacre*, *alacrem*, *alacrēs*, *alacrī*, *alacria*, *alacribus*, *alacris*, *alacrium*},

{*alta*, *altā*, *altae*, *altam*, *altārum*, *altās*, *alte*, *altī*, *altior*, *altiōra*, *altiōre*, *altiōrem*, *altiōrēs*, *altiōrī*, *altiōribus*, *altiōris*, *altiōrum*, *altīs*, *altissima*, *altissimā*, *altissimae*, *altissimam*, *altissimārum*, *altissimās*, *altissime*, *altissimī*, *altissimīs*, *altissimō*, *altissimōrum*, *altissimōs*, *altissimum*, *altissimus*, *altius*, *altō*, *altōrum*, *altōs*, *altum*, *altus*},

{*ampele*, *ampelī*, *ampelīs*, *ampelō*, *ampelon*, *ampelōrum*, *ampelos*, *ampelōs*, *ampelum*},

{*animal*, *animālī*, *animālia*, *animālibus*, *animālis*, *animālium*},

{*atoma*, *atomā*, *atomae*, *atomam*, *atomārum*, *atomās*, *atome*, *atomī*, *atomīs*, *atomō*, *atomōrum*, *atomōs*, *atomum*, *atomus*},

{*atrōcem*, *atrōcēs*, *atrōcī*, *atrōcia*, *atrōcibus*, *atrōcior*, *atrōciōra*, *atrōciōre*, *atrōciōrem*, *atrōciōrēs*, *atrōciōrī*, *atrōciōribus*, *atrōciōris*, *atrōciōrum*, *atrōcis*, *atrōcissima*, *atrōcissimā*, *atrōcissimae*, *atrōcissimam*, *atrōcissimārum*, *atrōcissimās*, *atrōcissime*, *atrōcissimī*, *atrōcissimīs*, *atrōcissimō*, *atrōcissimōrum*, *atrōcissimōs*, *atrōcissimum*, *atrōcissimus*, *atrōcium*, *atrōcius*, *atrōx*},

{*base*, *basem*, *basēs*, *basī*, *basibus*, *basim*, *basis*, *basīs*, *basium*},

{*bella*, *bellī*, *bellīs*, *bellō*, *bellōrum*, *bellum*},

{*celer*, *celere*, *celerem*, *celerēs*, *celerī*, *celeria*, *celeribus*, *celerior*, *celeriōra*, *celeriōre*, *celeriōrem*, *celeriōrēs*, *celeriōrī*, *celeriōribus*, *celeriōris*, *celeriōrum*, *celeris*, *celerium*, *celerius*, *celerrima*, *celerrimā*, *celerrimae*, *celerrimam*, *celerrimārum*, *celerrimās*, *celerrime*, *celerrimī*, *celerrimīs*, *celerrimō*, *celerrimōrum*, *celerrimōs*, *celerrimum*, *celerrimus*},

{*comētae*, *comētārum*, *comētās*, *comētē*, *comētem*, *comētēn*, *comētēs*, *comētīs*},

{*cornibus*, *cornū*, *cornua*, *cornūs*, *cornuum*},

{*Dīdō*, *Dīdōne*, *Dīdōnem*, *Dīdōnēs*, *Dīdōnī*, *Dīdōnibus*, *Dīdōnis*, *Dīdōnum*, *diē*, *diēbus*, *diēī*, *diem*, *diērum*, *diēs*},

{*dynastae*, *dynastārum*, *dynastās*, *dynastē*, *dynastēn*, *dynastēs*, *dynastīs*},

{*ēchibus*, *ēcho*, *ēchū*, *ēchuī*, *ēchum*, *ēchūs*, *ēchuum*},

{*fidē*, *fidēbus*, *fideī*, *fidem*, *fidērum*, *fidēs*},

{*fīlī*, *fīliī*, *fīliīs*, *fīliō*, *fīliōrum*, *fīliōs*, *fīlium*, *fīlius*},

{*haerese*, *haeresem*, *haeresēs*, *haeresī*, *haeresibus*, *haeresim*, *haeresis*, *haeresīs*, *haeresium*},

{*homine*, *hominem*, *hominēs*, *hominī*, *hominibus*, *hominis*, *hominum*, *homō*},

{*magister*, *magistrī*, *magistrīs*, *magistrō*, *magistrōrum*, *magistrōs*, *magistrum*},

{*manibus*, *manū*, *manuī*, *manum*, *manus*, *manūs*, *manuum*},

{*melior*, *melius*},

{*meliōra*, *meliōre*, *meliōrem*, *meliōrēs*, *meliōrī*, *meliōribus*, *meliōris*, *meliōrum*},

{*miser*, *misera*, *miserā*, *miserae*, *miseram*, *miserārum*, *miserās*, *miserī*, *miserior*, *miseriōra*, *miseriōre*, *miseriōrem*, *miseriōrēs*, *miseriōrī*, *miseriōribus*, *miseriōris*, *miseriōrum*, *miserīs*, *miserius*, *miserō*, *miserōrum*, *miserōs*, *miserrima*, *miserrimā*, *miserrimae*, *miserrimam*, *miserrimārum*, *miserrimās*, *miserrime*, *miserrimī*, *miserrimīs*, *miserrimō*, *miserrimōrum*, *miserrimōs*, *miserrimum*, *miserrimus*, *miserum*},

{*mūre*, *mūrī*, *mūrīs*, *mūrō*, *mūrōrum*, *mūrōs*, *mūrum*, *mūrus*},

{*mȳthe*, *mȳthī*, *mȳthīs*, *mȳthō*, *mȳthon*, *mȳthōrum*, *mȳthos*, *mȳthōs*, *mȳthum*},

{*nauta*, *nautā*, *nautae*, *nautam*, *nautārum*, *nautās*, *nautīs*},

{*nocte*, *noctem*, *noctēs*, *noctī*, *noctibus*, *noctis*, *noctium*, *nox*},

{*nōmen*, *nōmina*, *nōmine*, *nōminī*, *nōminibus*, *nōminis*, *nōminum*},

{*nymphae*, *nymphārum*, *nymphās*, *nymphē*, *nymphēn*, *nymphēs*, *nymphīs*},

{*phaenomena*, *phaenomenī*, *phaenomenīs*, *phaenomenō*, *phaenomenon*, *phaenomenōrum*},

{*puer*, *puerī*, *puerīs*, *puerō*, *puerōrum*, *puerōs*, *puerum*},

{*rē*, *rēbus*, *reī*, *rem*, *rērum*, *rēs*},

{*Rōma*, *Rōmā*, *Rōmae*, *Rōmam*},

{*rosa*, *rosā*, *rosae*, *rosam*, *rosārum*, *rosās*, *rosīs*},

{*sacer*, *sacra*, *sacrā*, *sacrae*, *sacram*, *sacrārum*, *sacrās*, *sacrī*, *sacrīs*, *sacrō*, *sacrōrum*, *sacrōs*, *sacrum*},

{*speciē*, *speciēbus*, *speciēī*, *speciem*, *speciērum*, *speciēs*},

{*stēlla*, *stēllā*, *stēllae*, *stēllam*, *stēllārum*, *stēllās*, *stēllīs*},

{*templa*, *templī*, *templīs*, *templō*, *templōrum*, *templum*},

{*tigre*, *tigrem*, *tigrēs*, *tigrī*, *tigribus*, *tigride*, *tigridem*, *tigridēs*, *tigridī*, *tigridibus*, *tigridis*, *tigridum*, *tigrim*, *tigris*, *tigrīs*, *tīgris*, *tigrium*},

{*trīste*, *trīstem*, *trīstēs*, *trīstī*, *trīstia*, *trīstibus*, *trīstior*, *trīstiōra*, *trīstiōre*, *trīstiōrem*, *trīstiōrēs*, *trīstiōrī*, *trīstiōribus*, *trīstiōris*, *trīstiōrum*, *trīstis*, *trīstissima*, *trīstissimā*, *trīstissimae*, *trīstissimam*, *trīstissimārum*, *trīstissimās*, *trīstissime*, *trīstissimī*, *trīstissimīs*, *trīstissimō*, *trīstissimōrum*, *trīstissimōs*, *trīstissimum*, *trīstissimus*, *trīstium*, *trīstius*},

{*turre*, *turrem*, *turrēs*, *turrī*, *turribus*, *turrim*, *turris*, *turrīs*, *turrium*},

{*ūlla*, *ūllā*, *ūllae*, *ūllam*, *ūllārum*, *ūllās*, *ūlle*, *ūllī*, *ūllīs*, *ūllīus*, *ūllō*, *ūllōrum*, *ūllōs*, *ūllum*, *ūllus*},

{*vetera*, *vetere*, *veterem*, *veterēs*, *veterī*, *veteribus*, *veteris*, *veterrima*, *veterrimā*, *veterrimae*, *veterrimam*, *veterrimārum*, *veterrimās*, *veterrime*, *veterrimī*, *veterrimīs*, *veterrimō*, *veterrimōrum*, *veterrimōs*, *veterrimum*, *veterrimus*, *veterum*, *vetus*, *vetustior*, *vetustiōra*, *vetustiōre*, *vetustiōrem*, *vetustiōrēs*, *vetustiōrī*, *vetustiōribus*, *vetustiōris*, *vetustiōrum*, *vetustissima*, *vetustissimā*, *vetustissimae*, *vetustissimam*, *vetustissimārum*, *vetustissimās*, *vetustissime*, *vetustissimī*, *vetustissimīs*, *vetustissimō*, *vetustissimōrum*, *vetustissimōs*, *vetustissimum*, *vetustissimus*, *vetustius*},

{*xiphiā*, *xiphiae*, *xiphiān*, *xiphiārum*, *xiphiās*, *xiphiīs*}.

Example 6.29.2. Latin verbs are divided into four conjugation groups. We test our algorithm on the following selections of verbs, based primarily on the Wikipedia links below:

`https://en.wikipedia.org/wiki/Latin_conjugation#First_conjugation`

amā, *amābam*, *amābāminī*, *amābāmur*, *amābāmus*, *amābant*, *amābantur*, *amābar*, *amābāre*, *amābāris*, *amābās*, *amābat*, *amābātis*, *amābātur*, *amābere*, *amāberis*, *amābiminī*, *amābimur*, *amābimus*, *amābis*, *amābit*, *amābitis*, *amābitur*, *amābō*, *amābor*, *amābunt*, *amābuntur*, *amāminī*, *amāmur*, *amāmus*, *amandī*, *amandō*, *amandum*, *amandus*, *amāns*, *amant*, *amantō*, *amantor*, *amantur*, *amāre*, *amārem*, *amārēminī*, *amārēmur*, *amārēmus*, *amārent*, *amārentur*, *amārer*, *amārēre*, *amārēris*, *amārēs*, *amāret*, *amārētis*, *amārētur*, *amārī*, *amāris*, *amās*, *amat*, *amāta*, *amātā*, *amātae*, *amātam*, *amātārum*, *amātās*, *amāte*, *amātī*, *amātis*, *amātīs*, *amātō*, *amātor*, *amātōrum*, *amātōs*, *amātōte*, *amātū*, *amātum*, *amātur*, *amātūrus*, *amātus*, *amāveram*, *amāverāmus*, *amāverant*, *amāverās*, *amāverat*, *amāverātis*, *amāvēre*, *amāverim*, *amāverimus*, *amāverīmus*, *amāverint*, *amāveris*, *amāverīs*, *amāverit*, *amāveritis*, *amāverītis*, *amāverō*, *amāvērunt*, *amāvī*, *amāvimus*, *amāvisse*, *amāvissem*, *amāvissēmus*, *amāvissent*, *amāvissēs*, *amāvisset*, *amāvissētis*, *amāvistī*, *amāvistis*, *amāvit*, *amem*, *amēminī*, *amēmur*, *amēmus*, *ament*, *amentur*, *amer*, *amēre*, *amēris*, *amēs*, *amet*, *amētis*, *amētur*, *amō*, *amor* — "love";

dā, *dabam*, *dabāminī*, *dabāmur*, *dabāmus*, *dabant*, *dabantur*, *dabar*, *dabāre*, *dabāris*, *dabās*, *dabat*, *dabātis*, *dabātur*, *dabere*, *daberis*, *dabiminī*, *dabimur*, *dabimus*, *dabis*, *dabit*, *dabitis*, *dabitur*, *dabō*, *dabor*, *dabunt*, *dabuntur*, *daminī*, *damur*, *damus*, *dandī*, *dandō*, *dandum*, *dandus*, *dāns*, *dant*, *dantō*, *dantor*, *dantur*, *dare*, *darem*, *darēminī*, *darēmur*, *darēmus*, *darent*, *darentur*, *darer*, *darēre*, *darēris*, *darēs*, *daret*, *darētis*, *darētur*, *darī*, *daris*, *dās*, *dat*, *data*, *datā*, *datae*, *datam*, *datārum*, *datās*, *date*, *datī*, *datis*, *datīs*, *datō*, *dator*, *datōrum*, *datōs*, *datōte*, *datū*, *datum*, *datur*, *datūrus*, *datus*, *dederam*, *dederāmus*, *dederant*, *dederās*, *dederat*, *dederātis*, *dedēre*, *dederim*, *dederimus*, *dederīmus*, *dederint*, *dederis*, *dederīs*, *dederit*, *dederitis*, *dederītis*, *dederō*, *dedērunt*, *dedī*, *dedimus*, *dedisse*, *dedissem*, *dedissēmus*, *dedissent*, *dedissēs*, *dedisset*, *dedissētis*, *dedistī*, *dedistis*, *dedit*, *dem*, *dēminī*, *dēmur*, *dēmus*, *dent*, *dentur*, *der*, *dēre*, *dēris*, *dēs*, *det*, *dētis*, *dētur*, *dō*, *dor* — "give";

fricā, *fricābam*, *fricābāminī*, *fricābāmur*, *fricābāmus*, *fricābant*, *fricābantur*, *fricābar*, *fricābāre*, *fricābāris*, *fricābās*, *fricābat*, *fricābātis*, *fricābātur*, *fricābere*, *fricāberis*, *fricābiminī*, *fricābimur*, *fricābimus*, *fricābis*, *fricābit*, *fricābitis*, *fricābitur*, *fricābō*, *fricābor*, *fricābunt*, *fricābuntur*, *fricāminī*, *fricāmur*, *fricāmus*, *fricandī*, *fricandō*, *fricandum*, *fricandus*, *fricāns*, *fricant*, *fricantō*, *fricantor*, *fricantur*, *fricāre*, *fricārem*, *fricārēminī*, *fricārēmur*, *fricārēmus*, *fricārent*, *fricārentur*, *fricārer*, *fricārēre*, *fricārēris*, *fricārēs*, *fricāret*, *fricārētis*, *fricārētur*, *fricārī*, *fricāris*, *fricās*, *fricat*, *fricāte*, *fricātibus*, *fricātis*, *fricātō*, *fricātor*, *fricātōte*, *fricātū*, *fricātuī*, *fricātum*, *fricātur*, *fricātūrus*, *fricātus*, *fricātūs*, *fricātuum*, *fricem*, *fricēminī*, *fricēmur*, *fricēmus*, *fricent*, *fricentur*, *fricer*, *fricēre*, *fricēris*, *fricēs*, *fricet*, *fricētis*, *fricētur*, *fricō*, *fricor*, *fricueram*, *fricuerāmus*, *fricuerant*, *fricuerās*, *fricuerat*, *fricuerātis*, *fricuēre*, *fricuerim*, *fricuerimus*, *fricuerīmus*, *fricuerint*, *fricueris*, *fricuerīs*, *fricuerit*, *fricueritis*, *fricuerītis*, *fricuerō*, *fricuērunt*, *fricuī*, *fricuimus*, *fricuisse*, *fricuissem*, *fricuissēmus*, *fricuissent*, *fricuissēs*, *fricuisset*, *fricuissētis*, *fricuistī*, *fricuistis*, *fricuit* — "rub";

iūta, *iūtā*, *iūtae*, *iūtam*, *iūtārum*, *iūtās*, *iūte*, *iūtī*, *iūtīs*, *iūtō*, *iūtōrum*, *iūtōs*, *iūtū*, *iūtum*, *iūtūrus*, *iūtus*, *iuvā*, *iuvābam*, *iuvābāminī*, *iuvābāmur*, *iuvābāmus*, *iuvābant*, *iuvābantur*, *iuvābar*, *iuvābāre*, *iuvābāris*, *iuvābās*, *iuvābat*, *iuvābātis*, *iuvābātur*, *iuvābere*, *iuvāberis*, *iuvābiminī*, *iuvābimur*, *iuvābimus*, *iuvābis*, *iuvābit*, *iuvābitis*, *iuvābitur*, *iuvābō*, *iuvābor*, *iuvābunt*, *iuvābuntur*, *iuvāminī*, *iuvāmur*, *iuvāmus*, *iuvandī*, *iuvandō*, *iuvandum*, *iuvandus*, *iuvāns*, *iuvant*, *iuvantō*, *iuvantor*, *iuvantur*, *iuvāre*, *iuvārem*, *iuvārēminī*, *iuvārēmur*, *iuvārēmus*, *iuvārent*, *iuvārentur*, *iuvārer*, *iuvārēre*, *iuvārēris*, *iuvārēs*, *iuvāret*, *iuvārētis*, *iuvārētur*, *iuvārī*, *iuvāris*, *iuvās*, *iuvat*, *iuvāte*, *iuvātis*, *iuvātō*, *iuvātor*, *iuvātōte*, *iuvātur*, *iuvem*, *iuvēminī*, *iuvēmur*, *iuvēmus*, *iuvent*, *iuventur*, *iuver*, *iūveram*, *iūverāmus*, *iūverant*, *iūverās*, *iūverat*, *iūverātis*, *iuvēre*, *iūvēre*, *iūverim*, *iūverimus*, *iūverīmus*, *iūverint*, *iuvēris*, *iūveris*, *iūverīs*, *iūverit*, *iūveritis*, *iūverītis*, *iūverō*, *iūvērunt*, *iuvēs*, *iuvet*, *iuvētis*, *iuvētur*, *iūvī*, *iūvimus*, *iūvisse*, *iūvissem*, *iūvissēmus*, *iūvissent*, *iūvissēs*, *iūvisset*, *iūvissētis*, *iūvistī*, *iūvistis*, *iūvit*, *iuvō*, *iuvor* — "help";

lavā, *lavābam*, *lavābāminī*, *lavābāmur*, *lavābāmus*, *lavābant*, *lavābantur*, *lavābar*, *lavābāre*, *lavābāris*, *lavābās*, *lavābat*, *lavābātis*, *lavābātur*, *lavābere*, *lavāberis*, *lavābiminī*, *lavābimur*, *lavābimus*, *lavābis*, *lavābit*, *lavābitis*, *lavābitur*, *lavābō*, *lavābor*, *lavābunt*, *lavābuntur*, *lavāminī*, *lavāmur*, *lavāmus*, *lavandī*, *lavandō*, *lavandum*, *lavandus*, *lavāns*, *lavant*, *lavantō*, *lavantor*, *lavantur*, *lavāre*, *lavārem*, *lavārēminī*, *lavārēmur*, *lavārēmus*, *lavārent*, *lavārentur*, *lavārer*, *lavārēre*, *lavārēris*, *lavārēs*, *lavāret*, *lavārētis*, *lavārētur*, *lavārī*, *lavāris*, *lavās*, *lavat*, *lavāta*, *lavātā*, *lavātae*, *lavātam*, *lavātārum*, *lavātās*, *lavāte*, *lavātī*, *lavātis*, *lavātīs*, *lavātō*, *lavātor*, *lavātōrum*, *lavātōs*, *lavātōte*, *lavātū*, *lavātum*, *lavātur*, *lavātūrus*, *lavātus*, *lavem*, *lavēminī*, *lavēmur*, *lavēmus*, *lavent*, *laventur*, *laver*, *lavēre*, *lavēris*, *lavēs*, *lavet*, *lavētis*, *lavētur*, *lavō*, *lavor* — "wash";

mīrābāminī, *mīrābāmur*, *mīrābantur*, *mīrābar*, *mīrābāre*, *mīrābāris*, *mīrābātur*, *mīrābere*, *mīrāberis*, *mīrābiminī*, *mīrābimur*, *mīrābitur*, *mīrābor*, *mīrābuntur*, *mīrāminī*, *mīrāmur*, *mīrandī*, *mīrandō*, *mīrandum*, *mīrandus*, *mīrāns*, *mīrantor*, *mīrantur*, *mīrāre*, *mīrārēminī*, *mīrārēmur*, *mīrārentur*, *mīrārer*, *mīrārēre*, *mīrārēris*, *mīrārētur*, *mīrārī*, *mīrāris*, *mīrāta*, *mīrātā*, *mīrātae*, *mīrātam*, *mīrātārum*, *mīrātās*, *mīrāte*, *mīrātī*, *mīrātīs*, *mīrātō*, *mīrātor*, *mīrātōrum*, *mīrātōs*, *mīrātū*, *mīrātum*, *mīrātur*, *mīrātūrus*, *mīrātus*, *mīrēminī*, *mīrēmur*, *mīrentur*, *mīrer*, *mīrēre*, *mīrēris*, *mīrētur*, *mīror* — "admire";

portā, *portābam*, *portābāminī*, *portābāmur*, *portābāmus*, *portābant*, *portābantur*, *portābar*, *portābāre*, *portābāris*, *portābās*, *portābat*, *portābātis*, *portābātur*, *portābere*, *portāberis*, *portābiminī*, *portābimur*, *portābimus*, *portābis*, *portābit*, *portābitis*, *portābitur*, *portābō*, *portābor*, *portābunt*, *portābuntur*, *portāminī*, *portāmur*, *portāmus*, *portandī*, *portandō*, *portandum*, *portandus*, *portāns*, *portant*, *portantō*, *portantor*, *portantur*, *portāre*, *portārem*, *portārēminī*, *portārēmur*, *portārēmus*, *portārent*, *portārentur*, *portārer*, *portārēre*, *portārēris*, *portārēs*, *portāret*, *portārētis*, *portārētur*, *portārī*, *portāris*, *portās*, *portat*, *portāta*, *portātā*, *portātae*, *portātam*, *portātārum*, *portātās*, *portāte*, *portātī*, *portātis*, *portātīs*, *portātō*, *portātor*, *portātōrum*, *portātōs*, *portātōte*, *portātū*, *portātum*, *portātur*, *portātūrus*, *portātus*, *portāveram*, *portāverāmus*, *portāverant*, *portāverās*, *portāverat*, *portāverātis*, *portāvēre*, *portāverim*, *portāverimus*, *portāverīmus*, *portāverint*, *portāveris*, *portāverīs*, *portāverit*, *portāveritis*, *portāverītis*, *portāverō*, *portāvērunt*, *portāvī*, *portāvimus*, *portāvisse*, *portāvissem*, *portāvissēmus*, *portāvissent*, *portāvissēs*, *portāvisset*, *portāvissētis*, *portāvistī*, *portāvistis*, *portāvit*, *portem*, *portēminī*, *portēmur*, *portēmus*, *portent*, *portentur*, *porter*, *portēre*, *portēris*, *portēs*, *portet*, *portētis*, *portētur*, *portō*, *portor* — "convey";

secā, *secābam*, *secābāminī*, *secābāmur*, *secābāmus*, *secābant*, *secābantur*, *secābar*, *secābāre*, *secābāris*, *secābās*, *secābat*, *secābātis*, *secābātur*, *secābere*, *secāberis*, *secābiminī*, *secābimur*, *secābimus*, *secābis*, *secābit*, *secābitis*, *secābitur*, *secābō*, *secābor*, *secābunt*, *secābuntur*, *secāminī*, *secāmur*, *secāmus*, *secandī*, *secandō*, *secandum*, *secandus*, *secāns*, *secant*, *secantō*, *secantor*, *secantur*, *secāre*, *secārem*, *secārēminī*, *secārēmur*, *secārēmus*, *secārent*, *secārentur*, *secārer*, *secārēre*, *secārēris*, *secārēs*, *secāret*, *secārētis*, *secārētur*, *secārī*, *secāris*, *secās*, *secat*, *secāte*, *secātis*, *secātō*, *secātor*, *secātōte*, *secātur*, *secem*, *secēminī*, *secēmur*, *secēmus*, *secent*, *secentur*, *secer*, *secēre*, *secēris*, *secēs*, *secet*, *secētis*, *secētur*, *secō*, *secor*, *secta*, *sectā*, *sectae*, *sectam*, *sectārum*, *sectās*, *secte*, *sectī*, *sectīs*, *sectō*, *sectōrum*, *sectōs*, *sectū*, *sectum*, *sectūrus*, *sectus*, *secueram*, *secuerāmus*, *secuerant*, *secuerās*, *secuerat*, *secuerātis*, *secuēre*, *secuerim*, *secuerimus*, *secuerīmus*, *secuerint*, *secueris*, *secuerīs*, *secuerit*, *secueritis*, *secuerītis*, *secuerō*, *secuērunt*, *secuī*, *secuimus*, *secuisse*, *secuissem*, *secuissēmus*, *secuissent*, *secuissēs*, *secuisset*, *secuissētis*, *secuistī*, *secuistis*, *secuit* — "cut";

stā, *stābam*, *stābāmus*, *stābant*, *stābās*, *stābat*, *stābātis*, *stābātur*, *stābimus*, *stābis*, *stābit*, *stābitis*, *stābitur*, *stābō*, *stābunt*, *stāmus*, *standī*, *standō*, *standum*, *standus*, *stāns*, *stant*, *stantō*, *stāre*, *stārem*, *stārēmus*, *stārent*, *stārēs*, *stāret*, *stārētis*, *stārētur*, *stārī*, *stās*, *stat*, *stata*, *statā*, *statae*, *statam*, *statārum*, *statās*, *state*, *stāte*, *statī*, *statīs*, *stātis*, *statō*, *stātō*, *statōrum*, *statōs*, *stātōte*, *statū*, *statum*, *stātur*, *statūrus*, *status*, *stem*, *stēmus*, *stent*, *stēs*, *stet*, *steteram*, *steterāmus*, *steterant*, *steterās*, *steterat*, *steterātis*, *stetēre*, *steterim*, *steterimus*, *steterīmus*, *steterint*, *steteris*, *steterīs*, *steterit*, *steteritis*, *steterītis*, *steterō*, *stetērunt*, *stetī*, *stetimus*, *stētis*, *stetisse*, *stetissem*, *stetissēmus*, *stetissent*, *stetissēs*, *stetisset*, *stetissētis*, *stetistī*, *stetistis*, *stetit*, *stētur*, *stō* — "stand";

vetā, *vetābam*, *vetābāminī*, *vetābāmur*, *vetābāmus*, *vetābant*, *vetābantur*, *vetābar*, *vetābāre*, *vetābāris*, *vetābās*, *vetābat*, *vetābātis*, *vetābātur*, *vetābere*, *vetāberis*, *vetābiminī*, *vetābimur*, *vetābimus*, *vetābis*, *vetābit*, *vetābitis*, *vetābitur*, *vetābō*, *vetābor*, *vetābunt*, *vetābuntur*, *vetāminī*, *vetāmur*, *vetāmus*, *vetandī*, *vetandō*, *vetandum*, *vetandus*, *vetāns*, *vetant*, *vetantō*, *vetantor*, *vetantur*, *vetāre*, *vetārem*, *vetārēminī*, *vetārēmur*, *vetārēmus*, *vetārent*, *vetārentur*, *vetārer*, *vetārēre*, *vetārēris*, *vetārēs*, *vetāret*, *vetārētis*, *vetārētur*, *vetārī*, *vetāris*, *vetās*, *vetat*, *vetāte*, *vetātis*, *vetātō*, *vetātor*, *vetātōte*, *vetātur*, *vetem*, *vetēminī*, *vetēmur*, *vetēmus*, *vetent*, *vetentur*, *veter*, *vetēre*, *vetēris*, *vetēs*, *vetet*, *vetētis*, *vetētur*, *vetita*, *vetitā*, *vetitae*, *vetitam*, *vetitārum*, *vetitās*, *vetite*, *vetitī*, *vetitīs*, *vetitō*, *vetitōrum*, *vetitōs*, *vetitū*, *vetitum*, *vetitūrus*, *vetitus*, *vetō*, *vetor*, *vetueram*, *vetuerāmus*, *vetuerant*, *vetuerās*, *vetuerat*, *vetuerātis*, *vetuēre*, *vetuerim*, *vetuerimus*, *vetuerīmus*, *vetuerint*, *vetueris*, *vetuerīs*, *vetuerit*, *vetueritis*, *vetuerītis*, *vetuerō*, *vetuērunt*, *vetuī*, *vetuimus*, *vetuisse*, *vetuissem*, *vetuissēmus*, *vetuissent*, *vetuissēs*, *vetuisset*, *vetuissētis*, *vetuistī*, *vetuistis*, *vetuit* — "forbid".

`https://en.wikipedia.org/wiki/Latin_conjugation#Second_conjugation`

aucta, *auctā*, *auctae*, *auctam*, *auctārum*, *auctās*, *aucte*, *auctī*, *auctīs*, *auctō*, *auctōrum*, *auctōs*, *auctū*, *auctum*, *auctūrus*, *auctus*, *augē*, *augeam*, *augeāminī*, *augeāmur*, *augeāmus*, *augeant*, *augeantur*, *augear*, *augeāre*, *augeāris*, *augeās*, *augeat*, *augeātis*, *augeātur*, *augēbam*, *augēbāminī*, *augēbāmur*, *augēbāmus*, *augēbant*, *augēbantur*, *augēbar*, *augēbāre*, *augēbāris*, *augēbās*, *augēbat*, *augēbātis*, *augēbātur*, *augēbere*, *augēberis*, *augēbiminī*, *augēbimur*, *augēbimus*, *augēbis*, *augēbit*, *augēbitis*, *augēbitur*, *augēbō*, *augēbor*, *augēbunt*, *augēbuntur*, *augēminī*,

augēmur, *augēmus*, *augendī*, *augendō*, *augendum*, *augendus*, *augēns*, *augent*, *augentō*, *augentor*, *augentur*, *augeō*, *augeor*, *augēre*, *augērem*, *augērēminī*, *augērēmur*, *augērēmus*, *augērent*, *augērentur*, *augērer*, *augērēre*, *augērēris*, *augērēs*, *augēret*, *augērētis*, *augērētur*, *augērī*, *augēris*, *augēs*, *auget*, *augēte*, *augētis*, *augētō*, *augētor*, *augētōte*, *augētur*, *auxeram*, *auxerāmus*, *auxerant*, *auxerās*, *auxerat*, *auxerātis*, *auxēre*, *auxerim*, *auxerimus*, *auxerīmus*, *auxerint*, *auxeris*, *auxerīs*, *auxerit*, *auxeritis*, *auxerītis*, *auxerō*, *auxērunt*, *auxī*, *auximus*, *auxisse*, *auxissem*, *auxissēmus*, *auxissent*, *auxissēs*, *auxisset*, *auxissētis*, *auxistī*, *auxistis*, *auxit* — “increase”;

ciē, *cieam*, *cieāminī*, *cieāmur*, *cieāmus*, *cieant*, *cieantur*, *ciear*, *cieāre*, *cieāris*, *cieās*, *cieat*, *cieātis*, *cieātur*, *ciēbam*, *ciēbāminī*, *ciēbāmur*, *ciēbāmus*, *ciēbant*, *ciēbantur*, *ciēbar*, *ciēbāre*, *ciēbāris*, *ciēbās*, *ciēbat*, *ciēbātis*, *ciēbātur*, *ciēbere*, *ciēberis*, *ciēbiminī*, *ciēbimur*, *ciēbimus*, *ciēbis*, *ciēbit*, *ciēbitis*, *ciēbitur*, *ciēbō*, *ciēbor*, *ciēbunt*, *ciēbuntur*, *ciēminī*, *ciēmur*, *ciēmus*, *ciendī*, *ciendō*, *ciendum*, *ciendus*, *ciēns*, *cient*, *cientō*, *cientor*, *cientur*, *cieō*, *cieor*, *ciēre*, *ciērem*, *ciērēminī*, *ciērēmur*, *ciērēmus*, *ciērent*, *ciērentur*, *ciērer*, *ciērēre*, *ciērēris*, *ciērēs*, *ciēret*, *ciērētis*, *ciērētur*, *ciērī*, *ciēris*, *ciēs*, *ciet*, *ciēte*, *ciētis*, *ciētō*, *ciētor*, *ciētōte*, *ciētur*, *cita*, *citā*, *citae*, *citam*, *citārum*, *citās*, *cite*, *citī*, *citīs*, *citō*, *citōrum*, *citōs*, *citū*, *citum*, *citūrus*, *citus*, *cīveram*, *cīverāmus*, *cīverant*, *cīverās*, *cīverat*, *cīverātis*, *cīvēre*, *cīverim*, *cīverimus*, *cīverīmus*, *cīverint*, *cīveris*, *cīverīs*, *cīverit*, *cīveritis*, *cīverītis*, *cīverō*, *cīvērunt*, *cīvī*, *cīvimus*, *cīvisse*, *cīvissem*, *cīvissēmus*, *cīvissent*, *cīvissēs*, *cīvisset*, *cīvissētis*, *cīvistī*, *cīvistis*, *cīvit* — “arouse”;

dēlē, *dēleam*, *dēleāminī*, *dēleāmur*, *dēleāmus*, *dēleant*, *dēleantur*, *dēlear*, *dēleāre*, *dēleāris*, *dēleās*, *dēleat*, *dēleātis*, *dēleātur*, *dēlēbam*, *dēlēbāminī*, *dēlēbāmur*, *dēlēbāmus*, *dēlēbant*, *dēlēbantur*, *dēlēbar*, *dēlēbāre*, *dēlēbāris*, *dēlēbās*, *dēlēbat*, *dēlēbātis*, *dēlēbātur*, *dēlēbere*, *dēlēberis*, *dēlēbiminī*, *dēlēbimur*, *dēlēbimus*, *dēlēbis*, *dēlēbit*, *dēlēbitis*, *dēlēbitur*, *dēlēbō*, *dēlēbor*, *dēlēbunt*, *dēlēbuntur*, *dēlēminī*, *dēlēmur*, *dēlēmus*, *dēlendī*, *dēlendō*, *dēlendum*, *dēlendus*, *dēlēns*, *dēlent*, *dēlentō*, *dēlentor*, *dēlentur*, *dēleō*, *dēleor*, *dēlēre*, *dēlērem*, *dēlērēminī*, *dēlērēmur*, *dēlērēmus*, *dēlērent*, *dēlērentur*, *dēlērer*, *dēlērēre*, *dēlērēris*, *dēlērēs*, *dēlēret*, *dēlērētis*, *dēlērētur*, *dēlērī*, *dēlēris*, *dēlēs*, *dēlet*, *dēlēta*, *dēlētā*, *dēlētae*, *dēlētam*, *dēlētārum*, *dēlētās*, *dēlēte*, *dēlētī*, *dēlētis*, *dēlētīs*, *dēlētō*, *dēlētor*, *dēlētōrum*, *dēlētōs*, *dēlētōte*, *dēlētū*, *dēlētum*, *dēlētur*, *dēlētūrus*, *dēlētus*, *dēlēveram*, *dēlēverāmus*, *dēlēverant*, *dēlēverās*, *dēlēverat*, *dēlēverātis*, *dēlēvēre*, *dēlēverim*, *dēlēverimus*, *dēlēverīmus*, *dēlēverint*, *dēlēveris*, *dēlēverīs*, *dēlēverit*, *dēlēveritis*, *dēlēverītis*, *dēlēverō*, *dēlēvērunt*, *dēlēvī*, *dēlēvimus*, *dēlēvisse*, *dēlēvissem*, *dēlēvissēmus*, *dēlēvissent*, *dēlēvissēs*, *dēlēvisset*, *dēlēvissētis*, *dēlēvistī*, *dēlēvistis*, *dēlēvit* — “destroy”;

docē, *doceam*, *doceāminī*, *doceāmur*, *doceāmus*, *doceant*, *doceantur*, *docear*, *doceāre*, *doceāris*, *doceās*, *doceat*, *doceātis*, *doceātur*, *docēbam*, *docēbāminī*, *docēbāmur*, *docēbāmus*, *docēbant*, *docēbantur*, *docēbar*, *docēbāre*, *docēbāris*, *docēbās*, *docēbat*, *docēbātis*, *docēbātur*, *docēbere*, *docēberis*, *docēbiminī*, *docēbimur*, *docēbimus*, *docēbis*, *docēbit*, *docēbitis*, *docēbitur*, *docēbō*, *docēbor*, *docēbunt*, *docēbuntur*, *docēminī*, *docēmur*, *docēmus*, *docendī*, *docendō*, *docendum*, *docendus*, *docēns*, *docent*, *docentō*, *docentor*, *docentur*, *doceō*, *doceor*, *docēre*, *docērem*, *docērēminī*, *docērēmur*, *docērēmus*, *docērent*, *docērentur*, *docērer*, *docērēre*, *docērēris*, *docērēs*, *docēret*, *docērētis*, *docērētur*, *docērī*, *docēris*, *docēs*, *docet*, *docēte*, *docētis*, *docētō*, *docētor*, *docētōte*, *docētur*, *docta*, *doctā*, *doctae*, *doctam*, *doctārum*, *doctās*, *docte*, *doctī*, *doctīs*, *doctō*, *doctōrum*, *doctōs*, *doctū*, *doctum*, *doctūrus*, *doctus*, *docueram*, *docuerāmus*, *docuerant*, *docuerās*, *docuerat*, *docuerātis*, *docuēre*, *docuerim*, *docuerimus*, *docuerīmus*, *docuerint*, *docueris*, *docuerīs*, *docuerit*, *docueritis*, *docuerītis*, *docuerō*, *docuērunt*, *docuī*, *docuimus*, *docuisse*, *docuissem*, *docuissēmus*, *docuissent*, *docuissēs*, *docuisset*, *docuissētis*, *docuistī*, *docuistis*, *docuit* — “teach”;

ferbueram, *ferbuerāmus*, *ferbuerant*, *ferbuerās*, *ferbuerat*, *ferbuerātis*, *ferbuēre*, *ferbuerim*, *ferbuerimus*, *ferbuerīmus*, *ferbuerint*, *ferbueris*, *ferbuerīs*, *ferbuerit*, *ferbueritis*, *ferbuerītis*, *ferbuerō*, *ferbuērunt*, *ferbuī*, *ferbuimus*, *ferbuisse*, *ferbuissem*, *ferbuissēmus*, *ferbuissent*, *ferbuissēs*, *ferbuisset*, *ferbuissētis*, *ferbuistī*, *ferbuistis*, *ferbuit*, *fervē*, *ferveam*, *ferveāmus*, *ferveant*, *ferveās*, *ferveat*, *ferveātis*, *fervēbam*, *fervēbāmus*, *fervēbant*, *fervēbās*, *fervēbat*, *fervēbātis*, *fervēbimus*, *fervēbis*, *fervēbit*, *fervēbitis*, *fervēbō*, *fervēbunt*, *fervēmus*, *fervendī*, *fervendō*, *fervendum*, *fervēns*, *fervent*, *ferventō*, *ferveō*, *fervēre*, *fervērem*, *fervērēmus*, *fervērent*, *fervērēs*, *fervēret*, *fervērētis*, *fervēs*, *fervet*, *fervēte*, *fervētis*, *fervētō*, *fervētōte*, *fervitū*, *fervitum*, *fervitūrus* — “boil”;

fōta, *fōtā*, *fōtae*, *fōtam*, *fōtārum*, *fōtās*, *fōte*, *fōtī*, *fōtīs*, *fōtō*, *fōtōrum*, *fōtōs*, *fōtū*, *fōtum*, *fōtūrus*, *fōtus*, *fovē*, *foveam*, *foveāminī*, *foveāmur*, *foveāmus*, *foveant*, *foveantur*, *fovear*, *foveāre*, *foveāris*, *foveās*, *foveat*, *foveātis*, *foveātur*, *fovēbam*, *fovēbāminī*, *fovēbāmur*, *fovēbāmus*, *fovēbant*, *fovēbantur*, *fovēbar*, *fovēbāre*, *fovēbāris*, *fovēbās*, *fovēbat*, *fovēbātis*, *fovēbātur*, *fovēbere*, *fovēberis*, *fovēbiminī*, *fovēbimur*, *fovēbimus*, *fovēbis*, *fovēbit*, *fovēbitis*, *fovēbitur*, *fovēbō*, *fovēbor*, *fovēbunt*, *fovēbuntur*, *fovēminī*, *fovēmur*, *fovēmus*, *fovendī*, *fovendō*, *fovendum*, *fovendus*, *fovēns*, *fovent*, *foventō*, *foventor*, *foventur*, *foveō*, *foveor*, *fōveram*, *fōverāmus*, *fōverant*, *fōverās*, *fōverat*, *fōverātis*, *fovēre*, *fōvēre*, *fovērem*, *fovērēminī*, *fovērēmur*, *fovērēmus*, *fovērent*, *fovērentur*, *fovērer*, *fovērēre*, *fovērēris*, *fovērēs*, *fovēret*, *fovērētis*, *fovērētur*, *fovērī*, *fōverim*, *fōverimus*, *fōverīmus*, *fōverint*, *fovēris*, *fōveris*, *fōverīs*, *fōverit*,

fōveritis, *fōverītis*, *fōverō*, *fōvērunt*, *fovēs*, *fovet*, *fovēte*, *fovētis*, *fovētō*, *fovētor*, *fovētōte*, *fovētur*, *fōvī*, *fōvimus*, *fōvisse*, *fōvissem*, *fōvissēmus*, *fōvissent*, *fōvissēs*, *fōvisset*, *fōvissētis*, *fōvistī*, *fōvistis*, *fōvit* — "caress";

iubē, *iubeam*, *iubeāminī*, *iubeāmur*, *iubeāmus*, *iubeant*, *iubeantur*, *iubear*, *iubeāre*, *iubeāris*, *iubeās*, *iubeat*, *iubeātis*, *iubeātur*, *iubēbam*, *iubēbāminī*, *iubēbāmur*, *iubēbāmus*, *iubēbant*, *iubēbantur*, *iubēbar*, *iubēbāre*, *iubēbāris*, *iubēbās*, *iubēbat*, *iubēbātis*, *iubēbātur*, *iubēbere*, *iubēberis*, *iubēbiminī*, *iubēbimur*, *iubēbimus*, *iubēbis*, *iubēbit*, *iubēbitis*, *iubēbitur*, *iubēbō*, *iubēbor*, *iubēbunt*, *iubēbuntur*, *iubēminī*, *iubēmur*, *iubēmus*, *iubendī*, *iubendō*, *iubendum*, *iubendus*, *iubēns*, *iubent*, *iubentō*, *iubentor*, *iubentur*, *iubeō*, *iubeor*, *iubēre*, *iubērem*, *iubērēminī*, *iubērēmur*, *iubērēmus*, *iubērent*, *iubērentur*, *iubērer*, *iubērēre*, *iubērēris*, *iubērēs*, *iubēret*, *iubērētis*, *iubērētur*, *iubērī*, *iubēris*, *iubēs*, *iubet*, *iubēte*, *iubētis*, *iubētō*, *iubētor*, *iubētōte*, *iubētur*, *iussa*, *iussā*, *iussae*, *iussam*, *iussārum*, *iussās*, *iusse*, *iusseram*, *iusserāmus*, *iusserant*, *iusserās*, *iusserat*, *iusserātis*, *iussēre*, *iusserim*, *iusserimus*, *iusserīmus*, *iusserint*, *iusseris*, *iusserīs*, *iusserit*, *iusseritis*, *iusserītis*, *iusserō*, *iussērunt*, *iussī*, *iussimus*, *iussīs*, *iussisse*, *iussissem*, *iussissēmus*, *iussissent*, *iussissēs*, *iussisset*, *iussissētis*, *iussistī*, *iussistis*, *iussit*, *iussō*, *iussōrum*, *iussōs*, *iussū*, *iussum*, *iussūrus*, *iussus* — "order";

momorderam, *momorderāmus*, *momorderant*, *momorderās*, *momorderat*, *momorderātis*, *momordēre*, *momorderim*, *momorderimus*, *momorderīmus*, *momorderint*, *momorderis*, *momorderīs*, *momorderit*, *momorderitis*, *momorderītis*, *momorderō*, *momordērunt*, *momordī*, *momordimus*, *momordisse*, *momordissem*, *momordissēmus*, *momordissent*, *momordissēs*, *momordisset*, *momordissētis*, *momordistī*, *momordistis*, *momordit*, *mordē*, *mordeam*, *mordeāminī*, *mordeāmur*, *mordeāmus*, *mordeant*, *mordeantur*, *mordear*, *mordeāre*, *mordeāris*, *mordeās*, *mordeat*, *mordeātis*, *mordeātur*, *mordēbam*, *mordēbāminī*, *mordēbāmur*, *mordēbāmus*, *mordēbant*, *mordēbantur*, *mordēbar*, *mordēbāre*, *mordēbāris*, *mordēbās*, *mordēbat*, *mordēbātis*, *mordēbātur*, *mordēbere*, *mordēberis*, *mordēbiminī*, *mordēbimur*, *mordēbimus*, *mordēbis*, *mordēbit*, *mordēbitis*, *mordēbitur*, *mordēbō*, *mordēbor*, *mordēbunt*, *mordēbuntur*, *mordēminī*, *mordēmur*, *mordēmus*, *mordendī*, *mordendō*, *mordendum*, *mordendus*, *mordēns*, *mordent*, *mordentō*, *mordentor*, *mordentur*, *mordeō*, *mordeor*, *mordēre*, *mordērem*, *mordērēminī*, *mordērēmur*, *mordērēmus*, *mordērent*, *mordērentur*, *mordērer*, *mordērēre*, *mordērēris*, *mordērēs*, *mordēret*, *mordērētis*, *mordērētur*, *mordērī*, *mordēris*, *mordēs*, *mordet*, *mordēte*, *mordētis*, *mordētō*, *mordētor*, *mordētōte*, *mordētur*, *morsa*, *morsā*, *morsae*, *morsam*, *morsārum*, *morsās*, *morse*, *morsī*, *morsīs*, *morsō*, *morsōrum*, *morsōs*, *morsū*, *morsum*, *morsūrus*, *morsus* — "bite";

monē, *moneam*, *moneāminī*, *moneāmur*, *moneāmus*, *moneant*, *moneantur*, *monear*, *moneāre*, *moneāris*, *moneās*, *moneat*, *moneātis*, *moneātur*, *monēbam*, *monēbāminī*, *monēbāmur*, *monēbāmus*, *monēbant*, *monēbantur*, *monēbar*, *monēbāre*, *monēbāris*, *monēbās*, *monēbat*, *monēbātis*, *monēbātur*, *monēbere*, *monēberis*, *monēbiminī*, *monēbimur*, *monēbimus*, *monēbis*, *monēbit*, *monēbitis*, *monēbitur*, *monēbō*, *monēbor*, *monēbunt*, *monēbuntur*, *monēminī*, *monēmur*, *monēmus*, *monendī*, *monendō*, *monendum*, *monendus*, *monēns*, *monent*, *monentō*, *monentor*, *monentur*, *moneō*, *moneor*, *monēre*, *monērem*, *monērēminī*, *monērēmur*, *monērēmus*, *monērent*, *monērentur*, *monērer*, *monērēre*, *monērēris*, *monērēs*, *monēret*, *monērētis*, *monērētur*, *monērī*, *monēris*, *monēs*, *monet*, *monēte*, *monētis*, *monētō*, *monētor*, *monētōte*, *monētur*, *monitū*, *monitum*, *monitūrus*, *monitus*, *monueram*, *monuerāmus*, *monuerant*, *monuerās*, *monuerat*, *monuerātis*, *monuēre*, *monuerim*, *monuerimus*, *monuerīmus*, *monuerint*, *monueris*, *monuerīs*, *monuerit*, *monueritis*, *monuerītis*, *monuerō*, *monuērunt*, *monuī*, *monuimus*, *monuisse*, *monuissem*, *monuissēmus*, *monuissent*, *monuissēs*, *monuisset*, *monuissētis*, *monuistī*, *monuistis*, *monuit* — "warn";

polliceāminī, *polliceāmur*, *polliceantur*, *pollicear*, *polliceāre*, *polliceāris*, *polliceātur*, *pollicēbāminī*, *pollicēbāmur*, *pollicēbantur*, *pollicēbar*, *pollicēbāre*, *pollicēbāris*, *pollicēbātur*, *pollicēbere*, *pollicēberis*, *pollicēbiminī*, *pollicēbimur*, *pollicēbitur*, *pollicēbor*, *pollicēbuntur*, *pollicēminī*, *pollicēmur*, *pollicendī*, *pollicendō*, *pollicendum*, *pollicendus*, *pollicēns*, *pollicentor*, *pollicentur*, *polliceor*, *pollicēre*, *pollicērēminī*, *pollicērēmur*, *pollicērentur*, *pollicērer*, *pollicērēre*, *pollicērēris*, *pollicērētur*, *pollicērī*, *pollicēris*, *pollicētor*, *pollicētur*, *pollicita*, *pollicitā*, *pollicitae*, *pollicitam*, *pollicitārum*, *pollicitās*, *pollicite*, *pollicitī*, *pollicitīs*, *pollicitō*, *pollicitōrum*, *pollicitōs*, *pollicitū*, *pollicitum*, *pollicitūrus*, *pollicitus* — "promise";

spondē, *spondeam*, *spondeāminī*, *spondeāmur*, *spondeāmus*, *spondeant*, *spondeantur*, *spondear*, *spondeāre*, *spondeāris*, *spondeās*, *spondeat*, *spondeātis*, *spondeātur*, *spondēbam*, *spondēbāminī*, *spondēbāmur*, *spondēbāmus*, *spondēbant*, *spondēbantur*, *spondēbar*, *spondēbāre*, *spondēbāris*, *spondēbās*, *spondēbat*, *spondēbātis*, *spondēbātur*, *spondēbere*, *spondēberis*, *spondēbiminī*, *spondēbimur*, *spondēbimus*, *spondēbis*, *spondēbit*, *spondēbitis*, *spondēbitur*, *spondēbō*, *spondēbor*, *spondēbunt*, *spondēbuntur*, *spondēminī*, *spondēmur*, *spondēmus*, *spondendī*, *spondendō*, *spondendum*, *spondendus*, *spondēns*, *spondent*, *spondentō*, *spondentor*, *spondentur*, *spondeō*, *spondeor*, *spondēre*, *spondērem*, *spondērēminī*, *spondērēmur*, *spondērēmus*, *spondērent*, *spondērentur*, *spondērer*,

spondērēre, *spondērēris*, *spondērēs*, *spondēret*, *spondērētis*, *spondērētur*, *spondērī*, *spondēris*, *spondēs*, *spondet*, *spondēte*, *spondētis*, *spondētō*, *spondētor*, *spondētōte*, *spondētur*, *spōnse*, *spōnsī*, *spōnsīs*, *spōnsō*, *spōnsōrum*, *spōnsōs*, *spōnsū*, *spōnsum*, *spōnsūrus*, *spōnsus*, *spoponderam*, *spoponderāmus*, *spoponderant*, *spoponderās*, *spoponderat*, *spoponderātis*, *spopondēre*, *spoponderim*, *spoponderimus*, *spoponderīmus*, *spoponderint*, *spoponderis*, *spoponderīs*, *spoponderit*, *spoponderitis*, *spoponderītis*, *spoponderō*, *spopondērunt*, *spopondī*, *spopondimus*, *spopondisse*, *spopondissem*, *spopondissēmus*, *spopondissent*, *spopondissēs*, *spopondisset*, *spopondissētis*, *spopondistī*, *spopondistis*, *spopondit* — "vow";

stridē, *strideam*, *strideāminī*, *strideāmur*, *strideāmus*, *strideant*, *strideantur*, *stridear*, *strideāre*, *strideāris*, *strideās*, *strideat*, *strideātis*, *strideātur*, *stridēbam*, *stridēbāminī*, *stridēbāmur*, *stridēbāmus*, *stridēbant*, *stridēbantur*, *stridēbar*, *stridēbāre*, *stridēbāris*, *stridēbās*, *stridēbat*, *stridēbātis*, *stridēbātur*, *stridēbere*, *stridēberis*, *stridēbiminī*, *stridēbimur*, *stridēbimus*, *stridēbis*, *stridēbit*, *stridēbitis*, *stridēbitur*, *stridēbō*, *stridēbor*, *stridēbunt*, *stridēbuntur*, *stridēminī*, *stridēmur*, *stridēmus*, *stridendī*, *stridendō*, *stridendum*, *stridendus*, *stridēns*, *strident*, *stridentō*, *stridentor*, *stridentur*, *strideō*, *strideor*, *stridēre*, *stridērem*, *stridērēminī*, *stridērēmur*, *stridērēmus*, *stridērent*, *stridērentur*, *stridērer*, *stridērēre*, *stridērēris*, *stridērēs*, *stridēret*, *stridērētis*, *stridērētur*, *stridērī*, *stridēris*, *stridēs*, *stridet*, *stridēte*, *stridētis*, *stridētō*, *stridētor*, *stridētōte*, *stridētur*, *stridū*, *stridum*, *stridūrus* — "hiss";

tenē, *teneam*, *teneāminī*, *teneāmur*, *teneāmus*, *teneant*, *teneantur*, *tenear*, *teneāre*, *teneāris*, *teneās*, *teneat*, *teneātis*, *teneātur*, *tenēbam*, *tenēbāminī*, *tenēbāmur*, *tenēbāmus*, *tenēbant*, *tenēbantur*, *tenēbar*, *tenēbāre*, *tenēbāris*, *tenēbās*, *tenēbat*, *tenēbātis*, *tenēbātur*, *tenēbere*, *tenēberis*, *tenēbiminī*, *tenēbimur*, *tenēbimus*, *tenēbis*, *tenēbit*, *tenēbitis*, *tenēbitur*, *tenēbō*, *tenēbor*, *tenēbunt*, *tenēbuntur*, *tenēminī*, *tenēmur*, *tenēmus*, *tenendī*, *tenendō*, *tenendum*, *tenendus*, *tenēns*, *tenent*, *tenentō*, *tenentor*, *tenentur*, *teneō*, *teneor*, *tenēre*, *tenērem*, *tenērēminī*, *tenērēmur*, *tenērēmus*, *tenērent*, *tenērentur*, *tenērer*, *tenērēre*, *tenērēris*, *tenērēs*, *tenēret*, *tenērētis*, *tenērētur*, *tenērī*, *tenēris*, *tenēs*, *tenet*, *tenēte*, *tenētis*, *tenētō*, *tenētor*, *tenētōte*, *tenētur*, *tenueram*, *tenuerāmus*, *tenuerant*, *tenuerās*, *tenuerat*, *tenuerātis*, *tenuēre*, *tenuerim*, *tenuerimus*, *tenuerīmus*, *tenuerint*, *tenueris*, *tenuerīs*, *tenuerit*, *tenueritis*, *tenuerītis*, *tenuerō*, *tenuērunt*, *tenuī*, *tenuimus*, *tenuisse*, *tenuissem*, *tenuissēmus*, *tenuissent*, *tenuissēs*, *tenuisset*, *tenuissētis*, *tenuistī*, *tenuistis*, *tenuit* — "hold";

terrē, *terream*, *terreāminī*, *terreāmur*, *terreāmus*, *terreant*, *terreantur*, *terrear*, *terreāre*, *terreāris*, *terreās*, *terreat*, *terreātis*, *terreātur*, *terrēbam*, *terrēbāminī*, *terrēbāmur*, *terrēbāmus*, *terrēbant*, *terrēbantur*, *terrēbar*, *terrēbāre*, *terrēbāris*, *terrēbās*, *terrēbat*, *terrēbātis*, *terrēbātur*, *terrēbere*, *terrēberis*, *terrēbiminī*, *terrēbimur*, *terrēbimus*, *terrēbis*, *terrēbit*, *terrēbitis*, *terrēbitur*, *terrēbō*, *terrēbor*, *terrēbunt*, *terrēbuntur*, *terrēminī*, *terrēmur*, *terrēmus*, *terrendī*, *terrendō*, *terrendum*, *terrendus*, *terrēns*, *terrent*, *terrentō*, *terrentor*, *terrentur*, *terreō*, *terreor*, *terrēre*, *terrērem*, *terrērēminī*, *terrērēmur*, *terrērēmus*, *terrērent*, *terrērentur*, *terrērer*, *terrērēre*, *terrērēris*, *terrērēs*, *terrēret*, *terrērētis*, *terrērētur*, *terrērī*, *terrēris*, *terrēs*, *terret*, *terrēte*, *terrētis*, *terrētō*, *terrētor*, *terrētōte*, *terrētur*, *territa*, *territā*, *territae*, *territam*, *territārum*, *territās*, *territe*, *territī*, *territīs*, *territō*, *territōrum*, *territōs*, *territū*, *territum*, *territūrus*, *territus*, *terrueram*, *terruerāmus*, *terruerant*, *terruerās*, *terruerat*, *terruerātis*, *terruēre*, *terruerim*, *terruerimus*, *terruerīmus*, *terruerint*, *terrueris*, *terruerīs*, *terruerit*, *terrueritis*, *terruerītis*, *terruerō*, *terruērunt*, *terruī*, *terruimus*, *terruisse*, *terruissem*, *terruissēmus*, *terruissent*, *terruissēs*, *terruisset*, *terruissētis*, *terruistī*, *terruistis*, *terruit* — "frighten";

vidē, *videam*, *videāminī*, *videāmur*, *videāmus*, *videant*, *videantur*, *videar*, *videāre*, *videāris*, *videās*, *videat*, *videātis*, *videātur*, *vidēbam*, *vidēbāminī*, *vidēbāmur*, *vidēbāmus*, *vidēbant*, *vidēbantur*, *vidēbar*, *vidēbāre*, *vidēbāris*, *vidēbās*, *vidēbat*, *vidēbātis*, *vidēbātur*, *vidēbere*, *vidēberis*, *vidēbiminī*, *vidēbimur*, *vidēbimus*, *vidēbis*, *vidēbit*, *vidēbitis*, *vidēbitur*, *vidēbō*, *vidēbor*, *vidēbunt*, *vidēbuntur*, *vidēminī*, *vidēmur*, *vidēmus*, *videndī*, *videndō*, *videndum*, *videndus*, *vidēns*, *vident*, *videntō*, *videntor*, *videntur*, *videō*, *videor*, *vīderam*, *vīderāmus*, *vīderant*, *vīderās*, *vīderat*, *vīderātis*, *vidēre*, *vīdēre*, *vidērem*, *vidērēminī*, *vidērēmur*, *vidērēmus*, *vidērent*, *vidērentur*, *vidērer*, *vidērēre*, *vidērēris*, *vidērēs*, *vidēret*, *vidērētis*, *vidērētur*, *vidērī*, *vīderim*, *vīderimus*, *vīderīmus*, *vīderint*, *vidēris*, *vīderis*, *vīderīs*, *vīderit*, *vīderitis*, *vīderītis*, *vīderō*, *vīdērunt*, *vidēs*, *videt*, *vidēte*, *vidētis*, *vidētō*, *vidētor*, *vidētōte*, *vidētur*, *vīdī*, *vīdimus*, *vīdisse*, *vīdissem*, *vīdissēmus*, *vīdissent*, *vīdissēs*, *vīdisset*, *vīdissētis*, *vīdistī*, *vīdistis*, *vīdit*, *vīsa*, *vīsā*, *vīsae*, *vīsam*, *vīsārum*, *vīsās*, *vīse*, *vīsī*, *vīsīs*, *vīsō*, *vīsōrum*, *vīsōs*, *vīsū*, *vīsum*, *vīsūrus*, *vīsus* — "see".

`https://en.wikipedia.org/wiki/Latin_conjugation#Third_conjugation`

ācta, *āctā*, *āctae*, *āctam*, *āctārum*, *āctās*, *ācte*, *āctī*, *āctīs*, *āctō*, *āctōrum*, *āctōs*, *āctū*, *āctum*, *āctūrus*, *āctus*, *agam*, *agāminī*, *agāmur*, *agāmus*, *agant*, *agantur*, *agar*, *agāre*, *agāris*, *agās*, *agat*, *agātis*, *agātur*, *age*, *agēbam*, *agēbāminī*, *agēbāmur*, *agēbāmus*, *agēbant*, *agēbantur*, *agēbar*, *agēbāre*, *agēbāris*, *agēbās*, *agēbat*, *agēbātis*, *agēbātur*, *agēminī*, *agēmur*, *agēmus*, *agendī*, *agendō*, *agendum*, *agendus*, *agēns*, *agent*, *agentur*, *agere*, *agēre*, *agerem*, *agerēminī*, *agerēmur*, *agerēmus*, *agerent*, *agerentur*, *agerer*, *agerēre*, *agerēris*, *agerēs*, *ageret*, *agerētis*,

agerētur, *ageris*, *agēris*, *agēs*, *aget*, *agētis*, *agētur*, *agī*, *agiminī*, *agimur*, *agimus*, *agis*, *agit*, *agite*, *agitis*, *agitō*, *agitor*, *agitōte*, *agitur*, *agō*, *agor*, *agunt*, *aguntō*, *aguntor*, *aguntur*, *ēgeram*, *ēgerāmus*, *ēgerant*, *ēgerās*, *ēgerat*, *ēgerātis*, *ēgēre*, *ēgerim*, *ēgerimus*, *ēgerīmus*, *ēgerint*, *ēgeris*, *ēgerīs*, *ēgerit*, *ēgeritis*, *ēgerītis*, *ēgerō*, *ēgērunt*, *ēgī*, *ēgimus*, *ēgisse*, *ēgissem*, *ēgissēmus*, *ēgissent*, *ēgissēs*, *ēgisset*, *ēgissētis*, *ēgistī*, *ēgistis*, *ēgit* — “drive”;

adhaerēscam, *adhaerēscāminī*, *adhaerēscāmur*, *adhaerēscāmus*, *adhaerēscant*, *adhaerēscantur*, *adhaerēscar*, *adhaerēscāre*, *adhaerēscāris*, *adhaerēscās*, *adhaerēscat*, *adhaerēscātis*, *adhaerēscātur*, *adhaerēsce*, *adhaerēscēbam*, *adhaerēscēbāminī*, *adhaerēscēbāmur*, *adhaerēscēbāmus*, *adhaerēscēbant*, *adhaerēscēbantur*, *adhaerēscēbar*, *adhaerēscēbāre*, *adhaerēscēbāris*, *adhaerēscēbās*, *adhaerēscēbat*, *adhaerēscēbātis*, *adhaerēscēbātur*, *adhaerēscēminī*, *adhaerēscēmur*, *adhaerēscēmus*, *adhaerēscendī*, *adhaerēscendō*, *adhaerēscendum*, *adhaerēscendus*, *adhaerēscēns*, *adhaerēscent*, *adhaerēscentur*, *adhaerēscere*, *adhaerēscēre*, *adhaerēscerem*, *adhaerēscerēminī*, *adhaerēscerēmur*, *adhaerēscerēmus*, *adhaerēscerent*, *adhaerēscerentur*, *adhaerēscerer*, *adhaerēscerēre*, *adhaerēscerēris*, *adhaerēscerēs*, *adhaerēsceret*, *adhaerēscerētis*, *adhaerēscerētur*, *adhaerēsceris*, *adhaerēscēris*, *adhaerēscēs*, *adhaerēscet*, *adhaerēscētis*, *adhaerēscētur*, *adhaerēscī*, *adhaerēsciminī*, *adhaerēscimur*, *adhaerēscimus*, *adhaerēscis*, *adhaerēscit*, *adhaerēscite*, *adhaerēscitis*, *adhaerēscitō*, *adhaerēscitor*, *adhaerēscitōte*, *adhaerēscitur*, *adhaerēscō*, *adhaerēscor*, *adhaerēscunt*, *adhaerēscuntō*, *adhaerēscuntor*, *adhaerēscuntur* — “adhere”;

adolēscam, *adolēscāminī*, *adolēscāmur*, *adolēscāmus*, *adolēscant*, *adolēscantur*, *adolēscar*, *adolēscāre*, *adolēscāris*, *adolēscās*, *adolēscat*, *adolēscātis*, *adolēscātur*, *adolēsce*, *adolēscēbam*, *adolēscēbāminī*, *adolēscēbāmur*, *adolēscēbāmus*, *adolēscēbant*, *adolēscēbantur*, *adolēscēbar*, *adolēscēbāre*, *adolēscēbāris*, *adolēscēbās*, *adolēscēbat*, *adolēscēbātis*, *adolēscēbātur*, *adolēscēminī*, *adolēscēmur*, *adolēscēmus*, *adolēscendī*, *adolēscendō*, *adolēscendum*, *adolēscendus*, *adolēscēns*, *adolēscent*, *adolēscentur*, *adolēscere*, *adolēscēre*, *adolēscerem*, *adolēscerēminī*, *adolēscerēmur*, *adolēscerēmus*, *adolēscerent*, *adolēscerentur*, *adolēscerer*, *adolēscerēre*, *adolēscerēris*, *adolēscerēs*, *adolēsceret*, *adolēscerētis*, *adolēscerētur*, *adolēsceris*, *adolēscēris*, *adolēscēs*, *adolēscet*, *adolēscētis*, *adolēscētur*, *adolēscī*, *adolēsciminī*, *adolēscimur*, *adolēscimus*, *adolēscis*, *adolēscit*, *adolēscite*, *adolēscitis*, *adolēscitō*, *adolēscitor*, *adolēscitōte*, *adolēscitur*, *adolēscō*, *adolēscor*, *adolēscunt*, *adolēscuntō*, *adolēscuntor*, *adolēscuntur*, *adolēveram*, *adolēverāmus*, *adolēverant*, *adolēverās*, *adolēverat*, *adolēverātis*, *adolēvēre*, *adolēverim*, *adolēverimus*, *adolēverīmus*, *adolēverint*, *adolēveris*, *adolēverīs*, *adolēverit*, *adolēveritis*, *adolēverītis*, *adolēverō*, *adolēvērunt*, *adolēvī*, *adolēvimus*, *adolēvisse*, *adolēvissem*, *adolēvissēmus*, *adolēvissent*, *adolēvissēs*, *adolēvisset*, *adolēvissētis*, *adolēvistī*, *adolēvistis*, *adolēvit*, *adulta*, *adultā*, *adultae*, *adultam*, *adultārum*, *adultās*, *adulte*, *adultī*, *adultīs*, *adultō*, *adultōrum*, *adultōs*, *adultū*, *adultum*, *adultūrus*, *adultus* — “mature”;

caedam, *caedāminī*, *caedāmur*, *caedāmus*, *caedant*, *caedantur*, *caedar*, *caedāre*, *caedāris*, *caedās*, *caedat*, *caedātis*, *caedātur*, *caede*, *caedēbam*, *caedēbāminī*, *caedēbāmur*, *caedēbāmus*, *caedēbant*, *caedēbantur*, *caedēbar*, *caedēbāre*, *caedēbāris*, *caedēbās*, *caedēbat*, *caedēbātis*, *caedēbātur*, *caedēminī*, *caedēmur*, *caedēmus*, *caedendī*, *caedendō*, *caedendum*, *caedendus*, *caedēns*, *caedent*, *caedentur*, *caedere*, *caedēre*, *caederem*, *caederēminī*, *caederēmur*, *caederēmus*, *caederent*, *caederentur*, *caederer*, *caederēre*, *caederēris*, *caederēs*, *caederet*, *caederētis*, *caederētur*, *caederis*, *caedēris*, *caedēs*, *caedet*, *caedētis*, *caedētur*, *caedī*, *caediminī*, *caedimur*, *caedimus*, *caedis*, *caedit*, *caedite*, *caeditis*, *caeditō*, *caeditor*, *caeditōte*, *caeditur*, *caedō*, *caedor*, *caedunt*, *caeduntō*, *caeduntor*, *caeduntur*, *caesa*, *caesā*, *caesae*, *caesam*, *caesārum*, *caesās*, *caese*, *caesī*, *caesīs*, *caesō*, *caesōrum*, *caesōs*, *caesū*, *caesum*, *caesūrus*, *caesus*, *cecīderam*, *cecīderāmus*, *cecīderant*, *cecīderās*, *cecīderat*, *cecīderātis*, *cecīdēre*, *cecīderim*, *cecīderimus*, *cecīderīmus*, *cecīderint*, *cecīderis*, *cecīderīs*, *cecīderit*, *cecīderitis*, *cecīderītis*, *cecīderō*, *cecīdērunt*, *cecīdī*, *cecīdimus*, *cecīdisse*, *cecīdissem*, *cecīdissēmus*, *cecīdissent*, *cecīdissēs*, *cecīdisset*, *cecīdissētis*, *cecīdistī*, *cecīdistis*, *cecīdit* — “kill”;

carpam, *carpāminī*, *carpāmur*, *carpāmus*, *carpant*, *carpantur*, *carpar*, *carpāre*, *carpāris*, *carpās*, *carpat*, *carpātis*, *carpātur*, *carpe*, *carpēbam*, *carpēbāminī*, *carpēbāmur*, *carpēbāmus*, *carpēbant*, *carpēbantur*, *carpēbar*, *carpēbāre*, *carpēbāris*, *carpēbās*, *carpēbat*, *carpēbātis*, *carpēbātur*, *carpēminī*, *carpēmur*, *carpēmus*, *carpendī*, *carpendō*, *carpendum*, *carpendus*, *carpēns*, *carpent*, *carpentur*, *carpere*, *carpēre*, *carperem*, *carperēminī*, *carperēmur*, *carperēmus*, *carperent*, *carperentur*, *carperer*, *carperēre*, *carperēris*, *carperēs*, *carperet*, *carperētis*, *carperētur*, *carperis*, *carpēris*, *carpēs*, *carpet*, *carpētis*, *carpētur*, *carpī*, *carpiminī*, *carpimur*, *carpimus*, *carpis*, *carpit*, *carpite*, *carpitis*, *carpitō*, *carpitor*, *carpitōte*, *carpitur*, *carpō*, *carpor*, *carpseram*, *carpserāmus*, *carpserant*, *carpserās*, *carpserat*, *carpserātis*, *carpsēre*, *carpserim*, *carpserimus*, *carpserīmus*, *carpserint*, *carpseris*, *carpserīs*, *carpserit*, *carpseritis*, *carpserītis*, *carpserō*, *carpsērunt*, *carpsī*, *carpsimus*, *carpsisse*, *carpsissem*, *carpsissēmus*, *carpsissent*, *carpsissēs*, *carpsisset*, *carpsissētis*, *carpsistī*, *carpsistis*, *carpsit*, *carpta*, *carptā*, *carptae*, *carptam*, *carptārum*, *carptās*, *carpte*, *carptī*, *carptīs*, *carptō*, *carptōrum*, *carptōs*, *carptū*, *carptum*, *carptūrus*, *carptus*, *carpunt*, *carpuntō*, *carpuntor*, *carpuntur* — “pluck”;

colam, *colāminī*, *colāmur*, *colāmus*, *colant*, *colantur*, *colar*, *colāre*, *colāris*, *colās*, *colat*, *colātis*, *colātur*, *cole*, *colēbam*, *colēbāminī*, *colēbāmur*, *colēbāmus*, *colēbant*, *colēbantur*, *colēbar*, *colēbāre*, *colēbāris*, *colēbās*, *colēbat*, *colēbātis*, *colēbātur*, *colēminī*, *colēmur*, *colēmus*, *colendī*, *colendō*, *colendum*, *colendus*, *colēns*, *colent*, *colentur*, *colere*, *colēre*, *colerem*, *colerēminī*, *colerēmur*, *colerēmus*, *colerent*, *colerentur*, *colerer*, *colerēre*, *colerēris*, *colerēs*, *coleret*, *colerētis*, *colerētur*, *coleris*, *colēris*, *colēs*, *colet*, *colētis*, *colētur*, *colī*, *coliminī*, *colimur*, *colimus*, *colis*, *colit*, *colite*, *colitis*, *colitō*, *colitor*, *colitōte*, *colitur*, *colō*, *color*, *colueram*, *coluerāmus*, *coluerant*, *coluerās*, *coluerat*, *coluerātis*, *coluēre*, *coluerim*, *coluerimus*, *coluerīmus*, *coluerint*, *colueris*, *coluerīs*, *coluerit*, *colueritis*, *coluerītis*, *coluerō*, *coluērunt*, *coluī*, *coluimus*, *coluisse*, *coluissem*, *coluissēmus*, *coluissent*, *coluissēs*, *coluisset*, *coluissētis*, *coluistī*, *coluistis*, *coluit*, *colunt*, *coluntō*, *coluntor*, *coluntur*, *culta*, *cultā*, *cultae*, *cultam*, *cultārum*, *cultās*, *culte*, *cultī*, *cultīs*, *cultō*, *cultōrum*, *cultōs*, *cultū*, *cultum*, *cultūrus*, *cultus* — “cultivate”;

cucurreram, *cucurrerāmus*, *cucurrerant*, *cucurrerās*, *cucurrerat*, *cucurrerātis*, *cucurrēre*, *cucurrerim*, *cucurrerimus*, *cucurrerīmus*, *cucurrerint*, *cucurreris*, *cucurrerīs*, *cucurrerit*, *cucurreritis*, *cucurrerītis*, *cucurrerō*, *cucurrērunt*, *cucurrī*, *cucurrimus*, *cucurrisse*, *cucurrissem*, *cucurrissēmus*, *cucurrissent*, *cucurrissēs*, *cucurrisset*, *cucurrissētis*, *cucurristī*, *cucurristis*, *cucurrit*, *curram*, *currāminī*, *currāmur*, *currāmus*, *currant*, *currantur*, *currar*, *currāre*, *currāris*, *currās*, *currat*, *currātis*, *currātur*, *curre*, *currēbam*, *currēbāminī*, *currēbāmur*, *currēbāmus*, *currēbant*, *currēbantur*, *currēbar*, *currēbāre*, *currēbāris*, *currēbās*, *currēbat*, *currēbātis*, *currēbātur*, *currēminī*, *currēmur*, *currēmus*, *currendī*, *currendō*, *currendum*, *currendus*, *currēns*, *current*, *currentur*, *currere*, *currēre*, *currerem*, *currerēminī*, *currerēmur*, *currerēmus*, *currerent*, *currerentur*, *currerer*, *currerēre*, *currerēris*, *currerēs*, *curreret*, *currerētis*, *currerētur*, *curreris*, *currēris*, *currēs*, *curret*, *currētis*, *currētur*, *currī*, *curriminī*, *currimur*, *currimus*, *curris*, *currit*, *currite*, *curritis*, *curritō*, *curritor*, *curritōte*, *curritur*, *currō*, *curror*, *currunt*, *curruntō*, *curruntor*, *curruntur*, *cursa*, *cursā*, *cursae*, *cursam*, *cursārum*, *cursās*, *curse*, *cursī*, *cursīs*, *cursō*, *cursōrum*, *cursōs*, *cursū*, *cursum*, *cursūrus*, *cursus* — “run”;

emam, *emāminī*, *emāmur*, *emāmus*, *emant*, *emantur*, *emar*, *emāre*, *emāris*, *emās*, *emat*, *emātis*, *emātur*, *eme*, *emēbam*, *emēbāminī*, *emēbāmur*, *emēbāmus*, *emēbant*, *emēbantur*, *emēbar*, *emēbāre*, *emēbāris*, *emēbās*, *emēbat*, *emēbātis*, *emēbātur*, *emēminī*, *emēmur*, *emēmus*, *emendī*, *emendō*, *emendum*, *emendus*, *emēns*, *ement*, *ementur*, *ēmeram*, *ēmerāmus*, *ēmerant*, *ēmerās*, *ēmerat*, *ēmerātis*, *emere*, *emēre*, *ēmēre*, *emerem*, *emerēminī*, *emerēmur*, *emerēmus*, *emerent*, *emerentur*, *emerer*, *emerēre*, *emerēris*, *emerēs*, *emeret*, *emerētis*, *emerētur*, *ēmerim*, *ēmerimus*, *ēmerīmus*, *ēmerint*, *emeris*, *emēris*, *ēmeris*, *ēmerīs*, *ēmerit*, *ēmeritis*, *ēmerītis*, *ēmerō*, *ēmērunt*, *emēs*, *emet*, *emētis*, *emētur*, *emī*, *ēmī*, *emiminī*, *emimur*, *emimus*, *ēmimus*, *emis*, *ēmisse*, *ēmissem*, *ēmissēmus*, *ēmissent*, *ēmissēs*, *ēmisset*, *ēmissētis*, *ēmistī*, *ēmistis*, *emit*, *ēmit*, *emite*, *emitis*, *emitō*, *emitor*, *emitōte*, *emitur*, *emō*, *emor*, *empta*, *emptā*, *emptae*, *emptam*, *emptārum*, *emptās*, *empte*, *emptī*, *emptīs*, *emptō*, *emptōrum*, *emptōs*, *ēmptū*, *emptum*, *ēmptum*, *ēmptūrus*, *emptus*, *ēmptus*, *emunt*, *emuntō*, *emuntor*, *emuntur* — “buy”;

flectam, *flectāminī*, *flectāmur*, *flectāmus*, *flectant*, *flectantur*, *flectar*, *flectāre*, *flectāris*, *flectās*, *flectat*, *flectātis*, *flectātur*, *flecte*, *flectēbam*, *flectēbāminī*, *flectēbāmur*, *flectēbāmus*, *flectēbant*, *flectēbantur*, *flectēbar*, *flectēbāre*, *flectēbāris*, *flectēbās*, *flectēbat*, *flectēbātis*, *flectēbātur*, *flectēminī*, *flectēmur*, *flectēmus*, *flectendī*, *flectendō*, *flectendum*, *flectendus*, *flectēns*, *flectent*, *flectentur*, *flectere*, *flectēre*, *flecterem*, *flecterēminī*, *flecterēmur*, *flecterēmus*, *flecterent*, *flecterentur*, *flecterer*, *flecterēre*, *flecterēris*, *flecterēs*, *flecteret*, *flecterētis*, *flecterētur*, *flecteris*, *flectēris*, *flectēs*, *flectet*, *flectētis*, *flectētur*, *flectī*, *flectiminī*, *flectimur*, *flectimus*, *flectis*, *flectit*, *flectite*, *flectitis*, *flectitō*, *flectitor*, *flectitōte*, *flectitur*, *flectō*, *flector*, *flectunt*, *flectuntō*, *flectuntor*, *flectuntur*, *flexa*, *flexā*, *flexae*, *flexam*, *flexārum*, *flexās*, *flexe*, *flexeram*, *flexerāmus*, *flexerant*, *flexerās*, *flexerat*, *flexerātis*, *flexēre*, *flexerim*, *flexerimus*, *flexerīmus*, *flexerint*, *flexeris*, *flexerīs*, *flexerit*, *flexeritis*, *flexerītis*, *flexerō*, *flexērunt*, *flexī*, *fleximus*, *flexīs*, *flexisse*, *flexissem*, *flexissēmus*, *flexissent*, *flexissēs*, *flexisset*, *flexissētis*, *flexistī*, *flexistis*, *flexit*, *flexō*, *flexōrum*, *flexōs*, *flexū*, *flexum*, *flexūrus*, *flexus* — “bend”;

flōrēscam, *flōrēscāminī*, *flōrēscāmur*, *flōrēscāmus*, *flōrēscant*, *flōrēscantur*, *flōrēscar*, *flōrēscāre*, *flōrēscāris*, *flōrēscās*, *flōrēscat*, *flōrēscātis*, *flōrēscātur*, *flōrēsce*, *flōrēscēbam*, *flōrēscēbāminī*, *flōrēscēbāmur*, *flōrēscēbāmus*, *flōrēscēbant*, *flōrēscēbantur*, *flōrēscēbar*, *flōrēscēbāre*, *flōrēscēbāris*, *flōrēscēbās*, *flōrēscēbat*, *flōrēscēbātis*, *flōrēscēbātur*, *flōrēscēminī*, *flōrēscēmur*, *flōrēscēmus*, *flōrēscendī*, *flōrēscendō*, *flōrēscendum*, *flōrēscendus*, *flōrēscēns*, *flōrēscent*, *flōrēscentur*, *flōrēscere*, *flōrēscēre*, *flōrēscerem*, *flōrēscerēminī*, *flōrēscerēmur*, *flōrēscerēmus*, *flōrēscerent*, *flōrēscerentur*, *flōrēscerer*, *flōrēscerēre*, *flōrēscerēris*, *flōrēscerēs*, *flōrēsceret*, *flōrēscerētis*, *flōrēscerētur*, *flōrēsceris*, *flōrēscēris*, *flōrēscēs*, *flōrēscet*, *flōrēscētis*, *flōrēscētur*, *flōrēscī*, *flōrēsciminī*, *flōrēscimur*, *flōrēscimus*, *flōrēscis*, *flōrēscit*, *flōrēscite*, *flōrēscitis*, *flōrēscitō*, *flōrēscitor*, *flōrēscitōte*, *flōrēscitur*, *flōrēscō*, *flōrēscor*, *flōrēscunt*, *flōrēscuntō*, *flōrēscuntor*, *flōrēscuntur* — “blossom”;

fūderam, *fūderāmus*, *fūderant*, *fūderās*, *fūderat*, *fūderātis*, *fūdēre*, *fūderim*, *fūderimus*, *fūderīmus*, *fūderint*, *fūderis*, *fūderīs*, *fūderit*, *fūderitis*, *fūderītis*, *fūderō*, *fūdērunt*, *fūdī*, *fūdimus*, *fūdisse*, *fūdissem*, *fūdissēmus*, *fūdissent*,

fūdissēs, *fūdisset*, *fūdissētis*, *fūdistī*, *fūdistis*, *fūdit*, *fundam*, *fundāmus*, *fundant*, *fundantur*, *fundās*, *fundat*, *fundātis*, *fundātur*, *funde*, *fundēbam*, *fundēbāmus*, *fundēbant*, *fundēbantur*, *fundēbās*, *fundēbat*, *fundēbātis*, *fundēbātur*, *fundēmus*, *fundendī*, *fundendō*, *fundendum*, *fundendus*, *fundēns*, *fundent*, *fundentur*, *fundere*, *funderem*, *funderēmus*, *funderent*, *funderentur*, *funderēs*, *funderet*, *funderētis*, *funderētur*, *fundēs*, *fundet*, *fundētis*, *fundētur*, *fundī*, *fundimus*, *fundis*, *fundit*, *fundite*, *funditis*, *funditō*, *funditōte*, *funditur*, *fundō*, *fundunt*, *funduntō*, *funduntur*, *fūsa*, *fūsā*, *fūsae*, *fūsam*, *fūsārum*, *fūsās*, *fūse*, *fūsī*, *fūsīs*, *fūsō*, *fūsōrum*, *fūsōs*, *fūsū*, *fūsum*, *fūsūrus*, *fūsus* — “pour”;

genita, *genitā*, *genitae*, *genitam*, *genitārum*, *genitās*, *genite*, *genitī*, *genitīs*, *genitō*, *genitōrum*, *genitōs*, *genitū*, *genitum*, *genitūrus*, *genitus*, *genueram*, *genuerāmus*, *genuerant*, *genuerās*, *genuerat*, *genuerātis*, *genuēre*, *genuerim*, *genuerimus*, *genuerīmus*, *genuerint*, *genueris*, *genuerīs*, *genuerit*, *genueritis*, *genuerītis*, *genuerō*, *genuērunt*, *genuī*, *genuimus*, *genuisse*, *genuissem*, *genuissēmus*, *genuissent*, *genuissēs*, *genuisset*, *genuissētis*, *genuistī*, *genuistis*, *genuit*, *gignam*, *gignāminī*, *gignāmur*, *gignāmus*, *gignant*, *gignantur*, *gignar*, *gignāre*, *gignāris*, *gignās*, *gignat*, *gignātis*, *gignātur*, *gigne*, *gignēbam*, *gignēbāminī*, *gignēbāmur*, *gignēbāmus*, *gignēbant*, *gignēbantur*, *gignēbar*, *gignēbāre*, *gignēbāris*, *gignēbās*, *gignēbat*, *gignēbātis*, *gignēbātur*, *gignēminī*, *gignēmur*, *gignēmus*, *gignendī*, *gignendō*, *gignendum*, *gignendus*, *gignēns*, *gignent*, *gignentur*, *gignere*, *gignēre*, *gignerem*, *gignerēminī*, *gignerēmur*, *gignerēmus*, *gignerent*, *gignerentur*, *gignerer*, *gignerēre*, *gignerēris*, *gignerēs*, *gigneret*, *gignerētis*, *gignerētur*, *gigneris*, *gignēris*, *gignēs*, *gignet*, *gignētis*, *gignētur*, *gignī*, *gigniminī*, *gignimur*, *gignimus*, *gignis*, *gignit*, *gignite*, *gignitis*, *gignitō*, *gignitor*, *gignitōte*, *gignitur*, *gignō*, *gignor*, *gignunt*, *gignuntō*, *gignuntor*, *gignuntur* — “beget”;

geram, *gerāminī*, *gerāmur*, *gerāmus*, *gerant*, *gerantur*, *gerar*, *gerāre*, *gerāris*, *gerās*, *gerat*, *gerātis*, *gerātur*, *gere*, *gerēbam*, *gerēbāminī*, *gerēbāmur*, *gerēbāmus*, *gerēbant*, *gerēbantur*, *gerēbar*, *gerēbāre*, *gerēbāris*, *gerēbās*, *gerēbat*, *gerēbātis*, *gerēbātur*, *gerēminī*, *gerēmur*, *gerēmus*, *gerendī*, *gerendō*, *gerendum*, *gerendus*, *gerēns*, *gerent*, *gerentur*, *gerere*, *gerēre*, *gererem*, *gererēminī*, *gererēmur*, *gererēmus*, *gererent*, *gererentur*, *gererer*, *gererēre*, *gererēris*, *gererēs*, *gereret*, *gererētis*, *gererētur*, *gereris*, *gerēris*, *gerēs*, *geret*, *gerētis*, *gerētur*, *gerī*, *geriminī*, *gerimur*, *gerimus*, *geris*, *gerit*, *gerite*, *geritis*, *geritō*, *geritor*, *geritōte*, *geritur*, *gerō*, *geror*, *gerunt*, *geruntō*, *geruntor*, *geruntur*, *gesseram*, *gesserāmus*, *gesserant*, *gesserās*, *gesserat*, *gesserātis*, *gessēre*, *gesserim*, *gesserimus*, *gesserīmus*, *gesserint*, *gesseris*, *gesserīs*, *gesserit*, *gesseritis*, *gesserītis*, *gesserō*, *gessērunt*, *gessī*, *gessimus*, *gessisse*, *gessissem*, *gessissēmus*, *gessissent*, *gessissēs*, *gessisset*, *gessissētis*, *gessistī*, *gessistis*, *gessit*, *gesta*, *gestā*, *gestae*, *gestam*, *gestārum*, *gestās*, *geste*, *gestī*, *gestīs*, *gestō*, *gestōrum*, *gestōs*, *gestū*, *gestum*, *gestūrus*, *gestus* — “wear”;

īcam, *īcāminī*, *īcāmur*, *īcāmus*, *īcant*, *īcantur*, *īcar*, *īcāre*, *īcāris*, *īcās*, *īcat*, *īcātis*, *īcātur*, *īce*, *īcēbam*, *īcēbāminī*, *īcēbāmur*, *īcēbāmus*, *īcēbant*, *īcēbantur*, *īcēbar*, *īcēbāre*, *īcēbāris*, *īcēbās*, *īcēbat*, *īcēbātis*, *īcēbātur*, *īcēminī*, *īcēmur*, *īcēmus*, *īcendī*, *īcendō*, *īcendum*, *īcendus*, *īcēns*, *īcent*, *īcentur*, *īceram*, *īcerāmus*, *īcerant*, *īcerās*, *īcerat*, *īcerātis*, *īcere*, *īcēre*, *īcerem*, *īcerēminī*, *īcerēmur*, *īcerēmus*, *īcerent*, *īcerentur*, *īcerer*, *īcerēre*, *īcerēris*, *īcerēs*, *īceret*, *īcerētis*, *īcerētur*, *īcerim*, *īcerimus*, *īcerīmus*, *īcerint*, *īceris*, *īcerīs*, *īcēris*, *īcerit*, *īceritis*, *īcerītis*, *īcerō*, *īcērunt*, *īcēs*, *īcet*, *īcētis*, *īcētur*, *īcī*, *īciminī*, *īcimur*, *īcimus*, *īcis*, *īcisse*, *īcissem*, *īcissēmus*, *īcissent*, *īcissēs*, *īcisset*, *īcissētis*, *īcistī*, *īcistis*, *īcit*, *īcite*, *īcitis*, *īcitō*, *īcitor*, *īcitōte*, *īcitur*, *īcō*, *īcor*, *īcta*, *īctā*, *īctae*, *īctam*, *īctārum*, *īctās*, *īcte*, *īctī*, *īctīs*, *īctō*, *īctōrum*, *īctōs*, *īctū*, *īctum*, *īctūrus*, *īctus*, *īcunt*, *īcuntō*, *īcuntor*, *īcuntur* — “strike”;

lēcta, *lēctā*, *lēctae*, *lēctam*, *lēctārum*, *lēctās*, *lēcte*, *lēctī*, *lēctīs*, *lēctō*, *lēctōrum*, *lēctōs*, *lēctū*, *lēctum*, *lēctūrus*, *lēctus*, *legam*, *legāminī*, *legāmur*, *legāmus*, *legant*, *legantur*, *legar*, *legāre*, *legāris*, *legās*, *legat*, *legātis*, *legātur*, *lege*, *legēbam*, *legēbāminī*, *legēbāmur*, *legēbāmus*, *legēbant*, *legēbantur*, *legēbar*, *legēbāre*, *legēbāris*, *legēbās*, *legēbat*, *legēbātis*, *legēbātur*, *legēminī*, *legēmur*, *legēmus*, *legendī*, *legendō*, *legendum*, *legendus*, *legēns*, *legent*, *legentur*, *lēgeram*, *lēgerāmus*, *lēgerant*, *lēgerās*, *lēgerat*, *lēgerātis*, *legere*, *legēre*, *lēgēre*, *legerem*, *legerēminī*, *legerēmur*, *legerēmus*, *legerent*, *legerentur*, *legerer*, *legerēre*, *legerēris*, *legerēs*, *legeret*, *legerētis*, *legerētur*, *lēgerim*, *lēgerimus*, *lēgerīmus*, *lēgerint*, *legeris*, *legēris*, *lēgeris*, *lēgerīs*, *lēgerit*, *lēgeritis*, *lēgerītis*, *lēgerō*, *lēgērunt*, *legēs*, *leget*, *legētis*, *legētur*, *legī*, *lēgī*, *legiminī*, *legimur*, *legimus*, *lēgimus*, *legis*, *lēgisse*, *lēgissem*, *lēgissēmus*, *lēgissent*, *lēgissēs*, *lēgisset*, *lēgissētis*, *lēgistī*, *lēgistis*, *legit*, *lēgit*, *legite*, *legitis*, *legitō*, *legitor*, *legitōte*, *legitur*, *legō*, *legor*, *legunt*, *leguntō*, *leguntor*, *leguntur* — “read”;

lēveram, *lēverāmus*, *lēverant*, *lēverās*, *lēverat*, *lēverātis*, *lēvēre*, *lēverim*, *lēverimus*, *lēverīmus*, *lēverint*, *lēveris*, *lēverīs*, *lēverit*, *lēveritis*, *lēverītis*, *lēverō*, *lēvērunt*, *lēvī*, *lēvimus*, *lēvisse*, *lēvissem*, *lēvissēmus*, *lēvissent*, *lēvissēs*, *lēvisset*, *lēvissētis*, *lēvistī*, *lēvistis*, *lēvit*, *linam*, *lināminī*, *lināmur*, *lināmus*, *linant*, *linantur*, *linar*, *lināre*, *lināris*, *linās*, *linat*, *linātis*, *linātur*, *line*, *linēbam*, *linēbāminī*, *linēbāmur*, *linēbāmus*, *linēbant*, *linēbantur*, *linēbar*, *linēbāre*, *linēbāris*, *linēbās*, *linēbat*, *linēbātis*, *linēbātur*, *linēminī*, *linēmur*, *linēmus*, *linendī*, *linendō*, *linendum*, *linendus*, *linēns*, *linent*, *linentur*, *linere*, *linēre*, *linerem*, *linerēminī*, *linerēmur*, *linerēmus*, *linerent*, *linerentur*, *linerer*, *linerēre*, *linerēris*, *linerēs*, *lineret*, *linerētis*, *linerētur*, *lineris*, *linēris*, *linēs*, *linet*, *linētis*, *linētur*, *linī*, *liniminī*, *linimur*, *linimus*, *linis*, *linit*, *linite*, *linitis*, *linitō*, *linitor*, *linitōte*, *linitur*, *linō*, *linor*, *linunt*, *linuntō*, *linuntor*, *linuntur*, *litū*, *litum*, *litūrus*, *litus* — “smear”;

locūta, *locūtā*, *locūtae*, *locūtam*, *locūtārum*, *locūtās*, *locūte*, *locūtī*, *locūtīs*, *locūtō*, *locūtōrum*, *locūtōs*, *locūtū*, *locūtum*, *locūtūrus*, *locūtus*, *loquāminī*, *loquāmur*, *loquantur*, *loquar*, *loquāre*, *loquāris*, *loquātur*, *loquēbāminī*, *loquēbāmur*, *loquēbantur*, *loquēbar*, *loquēbāre*, *loquēbāris*, *loquēbātur*, *loquēminī*, *loquēmur*, *loquendī*, *loquendō*, *loquendum*, *loquendus*, *loquēns*, *loquentur*, *loquere*, *loquēre*, *loquerēminī*, *loquerēmur*, *loquerentur*, *loquerer*, *loquerēre*, *loquerēris*, *loquerētur*, *loqueris*, *loquēris*, *loquētur*, *loquī*, *loquiminī*, *loquimur*, *loquitor*, *loquitur*, *loquor*, *loquuntor*, *loquuntur* — "speak";

messa, *messā*, *messae*, *messam*, *messārum*, *messās*, *messe*, *messī*, *messīs*, *messō*, *messōrum*, *messōs*, *messū*, *messueram*, *messuerāmus*, *messuerant*, *messuerās*, *messuerat*, *messuerātis*, *messuēre*, *messuerim*, *messuerimus*, *messuerīmus*, *messuerint*, *messueris*, *messuerīs*, *messuerit*, *messueritis*, *messuerītis*, *messuerō*, *messuērunt*, *messuī*, *messuimus*, *messuisse*, *messuissem*, *messuissēmus*, *messuissent*, *messuissēs*, *messuisset*, *messuissētis*, *messuistī*, *messuistis*, *messuit*, *messum*, *messūrus*, *messus*, *metam*, *metāminī*, *metāmur*, *metāmus*, *metant*, *metantur*, *metar*, *metāre*, *metāris*, *metās*, *metat*, *metātis*, *metātur*, *mete*, *metēbam*, *metēbāminī*, *metēbāmur*, *metēbāmus*, *metēbant*, *metēbantur*, *metēbar*, *metēbāre*, *metēbāris*, *metēbās*, *metēbat*, *metēbātis*, *metēbātur*, *metēminī*, *metēmur*, *metēmus*, *metendī*, *metendō*, *metendum*, *metendus*, *metēns*, *metent*, *metentur*, *metere*, *metēre*, *meterem*, *meterēminī*, *meterēmur*, *meterēmus*, *meterent*, *meterentur*, *meterer*, *meterēre*, *meterēris*, *meterēs*, *meteret*, *meterētis*, *meterētur*, *meteris*, *metēris*, *metēs*, *metet*, *metētis*, *metētur*, *metī*, *metiminī*, *metimur*, *metimus*, *metis*, *metit*, *metite*, *metitis*, *metitō*, *metitor*, *metitōte*, *metitur*, *metō*, *metor*, *metunt*, *metuntō*, *metuntor*, *metuntur* — "reap";

mīseram, *mīserāmus*, *mīserant*, *mīserās*, *mīserat*, *mīserātis*, *mīsēre*, *mīserim*, *mīserimus*, *mīserīmus*, *mīserint*, *mīseris*, *mīserīs*, *mīserit*, *mīseritis*, *mīserītis*, *mīserō*, *mīsērunt*, *mīsī*, *mīsimus*, *mīsisse*, *mīsissem*, *mīsissēmus*, *mīsissent*, *mīsissēs*, *mīsisset*, *mīsissētis*, *mīsistī*, *mīsistis*, *mīsit*, *missa*, *missā*, *missae*, *missam*, *missārum*, *missās*, *misse*, *missī*, *missīs*, *missō*, *missōrum*, *missōs*, *missū*, *missum*, *missūrus*, *missus*, *mittam*, *mittāminī*, *mittāmur*, *mittāmus*, *mittant*, *mittantur*, *mittar*, *mittāre*, *mittāris*, *mittās*, *mittat*, *mittātis*, *mittātur*, *mitte*, *mittēbam*, *mittēbāminī*, *mittēbāmur*, *mittēbāmus*, *mittēbant*, *mittēbantur*, *mittēbar*, *mittēbāre*, *mittēbāris*, *mittēbās*, *mittēbat*, *mittēbātis*, *mittēbātur*, *mittēminī*, *mittēmur*, *mittēmus*, *mittendī*, *mittendō*, *mittendum*, *mittendus*, *mittēns*, *mittent*, *mittentur*, *mittere*, *mittēre*, *mitterem*, *mitterēminī*, *mitterēmur*, *mitterēmus*, *mitterent*, *mitterentur*, *mitterer*, *mitterēre*, *mitterēris*, *mitterēs*, *mitteret*, *mitterētis*, *mitterētur*, *mitteris*, *mittēris*, *mittēs*, *mittet*, *mittētis*, *mittētur*, *mittī*, *mittiminī*, *mittimur*, *mittimus*, *mittis*, *mittit*, *mittite*, *mittitis*, *mittitō*, *mittitor*, *mittitōte*, *mittitur*, *mittō*, *mittor*, *mittunt*, *mittuntō*, *mittuntor*, *mittuntur* — "send";

nōscam, *nōscāminī*, *nōscāmur*, *nōscāmus*, *nōscant*, *nōscantur*, *nōscar*, *nōscāre*, *nōscāris*, *nōscās*, *nōscat*, *nōscātis*, *nōscātur*, *nōsce*, *nōscēbam*, *nōscēbāminī*, *nōscēbāmur*, *nōscēbāmus*, *nōscēbant*, *nōscēbantur*, *nōscēbar*, *nōscēbāre*, *nōscēbāris*, *nōscēbās*, *nōscēbat*, *nōscēbātis*, *nōscēbātur*, *nōscēminī*, *nōscēmur*, *nōscēmus*, *nōscendī*, *nōscendō*, *nōscendum*, *nōscendus*, *nōscēns*, *nōscent*, *nōscentur*, *nōscere*, *nōscēre*, *nōscerem*, *nōscerēminī*, *nōscerēmur*, *nōscerēmus*, *nōscerent*, *nōscerentur*, *nōscerer*, *nōscerēre*, *nōscerēris*, *nōscerēs*, *nōsceret*, *nōscerētis*, *nōscerētur*, *nōsceris*, *nōscēris*, *nōscēs*, *nōscet*, *nōscētis*, *nōscētur*, *nōscī*, *nōsciminī*, *nōscimur*, *nōscimus*, *nōscis*, *nōscit*, *nōscite*, *nōscitis*, *nōscitō*, *nōscitor*, *nōscitōte*, *nōscitur*, *nōscō*, *nōscor*, *nōscunt*, *nōscuntō*, *nōscuntor*, *nōscuntur*, *nōta*, *nōtā*, *nōtae*, *nōtam*, *nōtārum*, *nōtās*, *nōte*, *nōtī*, *nōtīs*, *nōtō*, *nōtōrum*, *nōtōs*, *nōtū*, *nōtum*, *nōtūrus*, *nōtus*, *nōveram*, *nōverāmus*, *nōverant*, *nōverās*, *nōverat*, *nōverātis*, *nōvēre*, *nōverim*, *nōverimus*, *nōverīmus*, *nōverint*, *nōveris*, *nōverīs*, *nōverit*, *nōveritis*, *nōverītis*, *nōverō*, *nōvērunt*, *nōvī*, *nōvimus*, *nōvisse*, *nōvissem*, *nōvissēmus*, *nōvissent*, *nōvissēs*, *nōvisset*, *nōvissētis*, *nōvistī*, *nōvistis*, *nōvit* — "know";

pāscam, *pāscāminī*, *pāscāmur*, *pāscāmus*, *pāscant*, *pāscantur*, *pāscar*, *pāscāre*, *pāscāris*, *pāscās*, *pāscat*, *pāscātis*, *pāscātur*, *pāsce*, *pāscēbam*, *pāscēbāminī*, *pāscēbāmur*, *pāscēbāmus*, *pāscēbant*, *pāscēbantur*, *pāscēbar*, *pāscēbāre*, *pāscēbāris*, *pāscēbās*, *pāscēbat*, *pāscēbātis*, *pāscēbātur*, *pāscēminī*, *pāscēmur*, *pāscēmus*, *pāscendī*, *pāscendō*, *pāscendum*, *pāscendus*, *pāscēns*, *pāscent*, *pāscentur*, *pāscere*, *pāscēre*, *pāscerem*, *pāscerēminī*, *pāscerēmur*, *pāscerēmus*, *pāscerent*, *pāscerentur*, *pāscerer*, *pāscerēre*, *pāscerēris*, *pāscerēs*, *pāsceret*, *pāscerētis*, *pāscerētur*, *pāsceris*, *pāscēris*, *pāscēs*, *pāscet*, *pāscētis*, *pāscētur*, *pāscī*, *pāsciminī*, *pāscimur*, *pāscimus*, *pāscis*, *pāscit*, *pāscite*, *pāscitis*, *pāscitō*, *pāscitor*, *pāscitōte*, *pāscitur*, *pāscō*, *pāscor*, *pāscunt*, *pāscuntō*, *pāscuntor*, *pāscuntur*, *pasta*, *pastā*, *pastae*, *pastam*, *pastārum*, *pastās*, *paste*, *pastī*, *pastīs*, *pastō*, *pastōrum*, *pastōs*, *pastū*, *pastum*, *pastūrus*, *pastus*, *pāveram*, *pāverāmus*, *pāverant*, *pāverās*, *pāverat*, *pāverātis*, *pāvēre*, *pāverim*, *pāverimus*, *pāverīmus*, *pāverint*, *pāveris*, *pāverīs*, *pāverit*, *pāveritis*, *pāverītis*, *pāverō*, *pāvērunt*, *pāvī*, *pāvimus*, *pāvisse*, *pāvissem*, *pāvissēmus*, *pāvissent*, *pāvissēs*, *pāvisset*, *pāvissētis*, *pāvistī*, *pāvistis*, *pāvit* — "feed";

pellam, *pellāminī*, *pellāmur*, *pellāmus*, *pellant*, *pellantur*, *pellar*, *pellāre*, *pellāris*, *pellās*, *pellat*, *pellātis*, *pellātur*, *pelle*, *pellēbam*, *pellēbāminī*, *pellēbāmur*, *pellēbāmus*, *pellēbant*, *pellēbantur*, *pellēbar*, *pellēbāre*, *pellēbāris*, *pellēbās*, *pellēbat*, *pellēbātis*, *pellēbātur*, *pellēminī*, *pellēmur*, *pellēmus*, *pellendī*, *pellendō*, *pellendum*, *pellendus*, *pellēns*, *pellent*, *pellentur*, *pellere*, *pellēre*, *pellerem*, *pellerēminī*, *pellerēmur*, *pellerēmus*, *pellerent*, *pellerentur*,

pellerer, *pellerēre*, *pellerēris*, *pellerēs*, *pelleret*, *pellerētis*, *pellerētur*, *pelleris*, *pellēris*, *pellēs*, *pellet*, *pellētis*, *pellētur*, *pellī*, *pelliminī*, *pellimur*, *pellimus*, *pellis*, *pellit*, *pellite*, *pellitis*, *pellitō*, *pellitor*, *pellitōte*, *pellitur*, *pellō*, *pellor*, *pellunt*, *pelluntō*, *pelluntor*, *pelluntur*, *pepuleram*, *pepulerāmus*, *pepulerant*, *pepulerās*, *pepulerat*, *pepulerātis*, *pepulēre*, *pepulerim*, *pepulerimus*, *pepulerīmus*, *pepulerint*, *pepuleris*, *pepulerīs*, *pepulerit*, *pepuleritis*, *pepulerītis*, *pepulerō*, *pepulērunt*, *pepulī*, *pepulimus*, *pepulisse*, *pepulissem*, *pepulissēmus*, *pepulissent*, *pepulissēs*, *pepulisset*, *pepulissētis*, *pepulistī*, *pepulistis*, *pepulit*, *pulsa*, *pulsā*, *pulsae*, *pulsam*, *pulsārum*, *pulsās*, *pulse*, *pulsī*, *pulsīs*, *pulsō*, *pulsōrum*, *pulsōs*, *pulsū*, *pulsum*, *pulsūrus*, *pulsus* — "beat";

petam, *petāminī*, *petāmur*, *petāmus*, *petant*, *petantur*, *petar*, *petāre*, *petāris*, *petās*, *petat*, *petātis*, *petātur*, *pete*, *petēbam*, *petēbāminī*, *petēbāmur*, *petēbāmus*, *petēbant*, *petēbantur*, *petēbar*, *petēbāre*, *petēbāris*, *petēbās*, *petēbat*, *petēbātis*, *petēbātur*, *petēminī*, *petēmur*, *petēmus*, *petendī*, *petendō*, *petendum*, *petendus*, *petēns*, *petent*, *petentur*, *petere*, *petēre*, *peterem*, *peterēminī*, *peterēmur*, *peterēmus*, *peterent*, *peterentur*, *peterer*, *peterēre*, *peterēris*, *peterēs*, *peteret*, *peterētis*, *peterētur*, *peteris*, *petēris*, *petēs*, *petet*, *petētis*, *petētur*, *petī*, *petiminī*, *petimur*, *petimus*, *petis*, *petit*, *petīta*, *petītā*, *petītae*, *petītam*, *petītārum*, *petītās*, *petite*, *petīte*, *petītī*, *petitis*, *petītīs*, *petitō*, *petītō*, *petitor*, *petītōrum*, *petītōs*, *petitōte*, *petītū*, *petītum*, *petitur*, *petītūrus*, *petītus*, *petīveram*, *petīverāmus*, *petīverant*, *petīverās*, *petīverat*, *petīverātis*, *petīvēre*, *petīverim*, *petīverimus*, *petīverīmus*, *petīverint*, *petīveris*, *petīverīs*, *petīverit*, *petīveritis*, *petīverītis*, *petīverō*, *petīvērunt*, *petīvī*, *petīvimus*, *petīvisse*, *petīvissem*, *petīvissēmus*, *petīvissent*, *petīvissēs*, *petīvisset*, *petīvissētis*, *petīvistī*, *petīvistis*, *petīvit*, *petō*, *petor*, *petunt*, *petuntō*, *petuntor*, *petuntur* — "seek";

sata, *satā*, *satae*, *satam*, *satārum*, *satās*, *sate*, *satī*, *satīs*, *satō*, *satōrum*, *satōs*, *satū*, *satum*, *satūrus*, *satus*, *seram*, *serāminī*, *serāmur*, *serāmus*, *serant*, *serantur*, *serar*, *serāre*, *serāris*, *serās*, *serat*, *serātis*, *serātur*, *sere*, *serēbam*, *serēbāminī*, *serēbāmur*, *serēbāmus*, *serēbant*, *serēbantur*, *serēbar*, *serēbāre*, *serēbāris*, *serēbās*, *serēbat*, *serēbātis*, *serēbātur*, *serēminī*, *serēmur*, *serēmus*, *serendī*, *serendō*, *serendum*, *serendus*, *serēns*, *serent*, *serentur*, *serere*, *serēre*, *sererem*, *sererēminī*, *sererēmur*, *sererēmus*, *sererent*, *sererentur*, *sererer*, *sererēre*, *sererēris*, *sererēs*, *sereret*, *sererētis*, *sererētur*, *sereris*, *serēris*, *serēs*, *seret*, *serētis*, *serētur*, *serī*, *seriminī*, *serimur*, *serimus*, *seris*, *serit*, *serite*, *seritis*, *seritō*, *seritor*, *seritōte*, *seritur*, *serō*, *seror*, *serunt*, *seruntō*, *seruntor*, *seruntur*, *sēveram*, *sēverāmus*, *sēverant*, *sēverās*, *sēverat*, *sēverātis*, *sēvēre*, *sēverim*, *sēverimus*, *sēverīmus*, *sēverint*, *sēveris*, *sēverīs*, *sēverit*, *sēveritis*, *sēverītis*, *sēverō*, *sēvērunt*, *sēvī*, *sēvimus*, *sēvisse*, *sēvissem*, *sēvissēmus*, *sēvissent*, *sēvissēs*, *sēvisset*, *sēvissētis*, *sēvistī*, *sēvistis*, *sēvit* — "sow";

sternam, *sternāminī*, *sternāmur*, *sternāmus*, *sternant*, *sternantur*, *sternar*, *sternāre*, *sternāris*, *sternās*, *sternat*, *sternātis*, *sternātur*, *sterne*, *sternēbam*, *sternēbāminī*, *sternēbāmur*, *sternēbāmus*, *sternēbant*, *sternēbantur*, *sternēbar*, *sternēbāre*, *sternēbāris*, *sternēbās*, *sternēbat*, *sternēbātis*, *sternēbātur*, *sternēminī*, *sternēmur*, *sternēmus*, *sternendī*, *sternendō*, *sternendum*, *sternendus*, *sternēns*, *sternent*, *sternentur*, *sternere*, *sternēre*, *sternerem*, *sternerēminī*, *sternerēmur*, *sternerēmus*, *sternerent*, *sternerentur*, *sternerer*, *sternerēre*, *sternerēris*, *sternerēs*, *sterneret*, *sternerētis*, *sternerētur*, *sterneris*, *sternēris*, *sternēs*, *sternet*, *sternētis*, *sternētur*, *sternī*, *sterniminī*, *sternimur*, *sternimus*, *sternis*, *sternit*, *sternite*, *sternitis*, *sternitō*, *sternitor*, *sternitōte*, *sternitur*, *sternō*, *sternor*, *sternunt*, *sternuntō*, *sternuntor*, *sternuntur*, *strāta*, *strātā*, *strātae*, *strātam*, *strātārum*, *strātās*, *strāte*, *strātī*, *strātīs*, *strātō*, *strātōrum*, *strātōs*, *strātū*, *strātum*, *strātūrus*, *strātus*, *strāveram*, *strāverāmus*, *strāverant*, *strāverās*, *strāverat*, *strāverātis*, *strāvēre*, *strāverim*, *strāverimus*, *strāverīmus*, *strāverint*, *strāveris*, *strāverīs*, *strāverit*, *strāveritis*, *strāverītis*, *strāverō*, *strāvērunt*, *strāvī*, *strāvimus*, *strāvisse*, *strāvissem*, *strāvissēmus*, *strāvissent*, *strāvissēs*, *strāvisset*, *strāvissētis*, *strāvistī*, *strāvistis*, *strāvit* — "spread";

tācta, *tāctā*, *tāctae*, *tāctam*, *tāctārum*, *tāctās*, *tācte*, *tāctī*, *tāctīs*, *tāctō*, *tāctōrum*, *tāctōs*, *tāctū*, *tāctum*, *tāctūrus*, *tāctus*, *tangam*, *tangāminī*, *tangāmur*, *tangāmus*, *tangant*, *tangantur*, *tangar*, *tangāre*, *tangāris*, *tangās*, *tangat*, *tangātis*, *tangātur*, *tange*, *tangēbam*, *tangēbāminī*, *tangēbāmur*, *tangēbāmus*, *tangēbant*, *tangēbantur*, *tangēbar*, *tangēbāre*, *tangēbāris*, *tangēbās*, *tangēbat*, *tangēbātis*, *tangēbātur*, *tangēminī*, *tangēmur*, *tangēmus*, *tangendī*, *tangendō*, *tangendum*, *tangendus*, *tangēns*, *tangent*, *tangentur*, *tangere*, *tangēre*, *tangerem*, *tangerēminī*, *tangerēmur*, *tangerēmus*, *tangerent*, *tangerentur*, *tangerer*, *tangerēre*, *tangerēris*, *tangerēs*, *tangeret*, *tangerētis*, *tangerētur*, *tangeris*, *tangēris*, *tangēs*, *tanget*, *tangētis*, *tangētur*, *tangī*, *tangiminī*, *tangimur*, *tangimus*, *tangis*, *tangit*, *tangite*, *tangitis*, *tangitō*, *tangitor*, *tangitōte*, *tangitur*, *tangō*, *tangor*, *tangunt*, *tanguntō*, *tanguntor*, *tanguntur* — "touch";

teram, *terāminī*, *terāmur*, *terāmus*, *terant*, *terantur*, *terar*, *terāre*, *terāris*, *terās*, *terat*, *terātis*, *terātur*, *tere*, *terēbam*, *terēbāminī*, *terēbāmur*, *terēbāmus*, *terēbant*, *terēbantur*, *terēbar*, *terēbāre*, *terēbāris*, *terēbās*, *terēbat*, *terēbātis*, *terēbātur*, *terēminī*, *terēmur*, *terēmus*, *terendī*, *terendō*, *terendum*, *terendus*, *terēns*, *terent*, *terentur*, *terere*, *terēre*, *tererem*, *tererēminī*, *tererēmur*, *tererēmus*, *tererent*, *tererentur*, *tererer*, *tererēre*, *tererēris*, *tererēs*, *tereret*, *tererētis*, *tererētur*, *tereris*, *terēris*, *terēs*, *teret*, *terētis*, *terētur*, *terī*, *teriminī*, *terimur*, *terimus*, *teris*, *terit*, *terite*, *teritis*, *teritō*, *teritor*, *teritōte*, *teritur*, *terō*, *teror*, *terunt*, *teruntō*, *teruntor*, *teruntur*, *trīta*, *trītā*, *trītae*,

trītam, *trītārum*, *trītās*, *trīte*, *trītī*, *trītīs*, *trītō*, *trītōrum*, *trītōs*, *trītū*, *trītum*, *trītūrus*, *trītus*, *trīveram*, *trīverāmus*, *trīverant*, *trīverās*, *trīverat*, *trīverātis*, *trīvēre*, *trīverim*, *trīverimus*, *trīverīmus*, *trīverint*, *trīveris*, *trīverīs*, *trīverit*, *trīveritis*, *trīverītis*, *trīverō*, *trīvērunt*, *trīvī*, *trīvimus*, *trīvisse*, *trīvissem*, *trīvissēmus*, *trīvissent*, *trīvissēs*, *trīvisset*, *trīvissētis*, *trīvistī*, *trīvistis*, *trīvit* — “triturate”;

texam, *texāminī*, *texāmur*, *texāmus*, *texant*, *texantur*, *texar*, *texāre*, *texāris*, *texās*, *texat*, *texātis*, *texātur*, *texe*, *texēbam*, *texēbāminī*, *texēbāmur*, *texēbāmus*, *texēbant*, *texēbantur*, *texēbar*, *texēbāre*, *texēbāris*, *texēbās*, *texēbat*, *texēbātis*, *texēbātur*, *texēminī*, *texēmur*, *texēmus*, *texendī*, *texendō*, *texendum*, *texendus*, *texēns*, *texent*, *texentur*, *texere*, *texēre*, *texerem*, *texerēminī*, *texerēmur*, *texerēmus*, *texerent*, *texerentur*, *texerer*, *texerēre*, *texerēris*, *texerēs*, *texeret*, *texerētis*, *texerētur*, *texeris*, *texēris*, *texēs*, *texet*, *texētis*, *texētur*, *texī*, *teximinī*, *teximur*, *teximus*, *texis*, *texit*, *texite*, *texitis*, *texitō*, *texitor*, *texitōte*, *texitur*, *texō*, *texor*, *texta*, *textā*, *textae*, *textam*, *textārum*, *textās*, *texte*, *textī*, *textīs*, *textō*, *textōrum*, *textōs*, *textū*, *textum*, *textūrus*, *textus*, *texueram*, *texuerāmus*, *texuerant*, *texuerās*, *texuerat*, *texuerātis*, *texuēre*, *texuerim*, *texuerimus*, *texuerīmus*, *texuerint*, *texueris*, *texuerīs*, *texuerit*, *texueritis*, *texuerītis*, *texuerō*, *texuērunt*, *texuī*, *texuimus*, *texuisse*, *texuissem*, *texuissēmus*, *texuissent*, *texuissēs*, *texuisset*, *texuissētis*, *texuistī*, *texuistis*, *texuit*, *texunt*, *texuntō*, *texuntor*, *texuntur* — “weave”;

tracta, *tractā*, *tractae*, *tractam*, *tractārum*, *tractās*, *tracte*, *tractī*, *tractīs*, *tractō*, *tractōrum*, *tractōs*, *tractū*, *tractum*, *tractūrus*, *tractus*, *traham*, *trahāminī*, *trahāmur*, *trahāmus*, *trahant*, *trahantur*, *trahar*, *trahāre*, *trahāris*, *trahās*, *trahat*, *trahātis*, *trahātur*, *trahe*, *trahēbam*, *trahēbāminī*, *trahēbāmur*, *trahēbāmus*, *trahēbant*, *trahēbantur*, *trahēbar*, *trahēbāre*, *trahēbāris*, *trahēbās*, *trahēbat*, *trahēbātis*, *trahēbātur*, *trahēminī*, *trahēmur*, *trahēmus*, *trahendī*, *trahendō*, *trahendum*, *trahendus*, *trahēns*, *trahent*, *trahentur*, *trahere*, *trahēre*, *traherem*, *traherēminī*, *traherēmur*, *traherēmus*, *traherent*, *traherentur*, *traherer*, *traherēre*, *traherēris*, *traherēs*, *traheret*, *traherētis*, *traherētur*, *traheris*, *trahēris*, *trahēs*, *trahet*, *trahētis*, *trahētur*, *trahī*, *trahiminī*, *trahimur*, *trahimus*, *trahis*, *trahit*, *trahite*, *trahitis*, *trahitō*, *trahitor*, *trahitōte*, *trahitur*, *trahō*, *trahor*, *trahunt*, *trahuntō*, *trahuntor*, *trahuntur* — “drag”;

versa, *versā*, *versae*, *versam*, *versārum*, *versās*, *verse*, *versī*, *versīs*, *versō*, *versōrum*, *versōs*, *versū*, *versum*, *versūrus*, *versus*, *vertam*, *vertāminī*, *vertāmur*, *vertāmus*, *vertant*, *vertantur*, *vertar*, *vertāre*, *vertāris*, *vertās*, *vertat*, *vertātis*, *vertātur*, *verte*, *vertēbam*, *vertēbāminī*, *vertēbāmur*, *vertēbāmus*, *vertēbant*, *vertēbantur*, *vertēbar*, *vertēbāre*, *vertēbāris*, *vertēbās*, *vertēbat*, *vertēbātis*, *vertēbātur*, *vertēminī*, *vertēmur*, *vertēmus*, *vertendī*, *vertendō*, *vertendum*, *vertendus*, *vertēns*, *vertent*, *vertentur*, *verteram*, *verterāmus*, *verterant*, *verterās*, *verterat*, *verterātis*, *vertere*, *vertēre*, *verterem*, *verterēminī*, *verterēmur*, *verterēmus*, *verterent*, *verterentur*, *verterer*, *verterēre*, *verterēris*, *verterēs*, *verteret*, *verterētis*, *verterētur*, *verterim*, *verterimus*, *verterīmus*, *verterint*, *verteris*, *verterīs*, *vertēris*, *verterit*, *verteritis*, *verterītis*, *verterō*, *vertērunt*, *vertēs*, *vertet*, *vertētis*, *vertētur*, *vertī*, *vertiminī*, *vertimur*, *vertimus*, *vertis*, *vertisse*, *vertissem*, *vertissēmus*, *vertissent*, *vertissēs*, *vertisset*, *vertissētis*, *vertistī*, *vertistis*, *vertit*, *vertite*, *vertitis*, *vertitō*, *vertitor*, *vertitōte*, *vertitur*, *vertō*, *vertor*, *vertunt*, *vertuntō*, *vertuntor*, *vertuntur* — “turn”;

vīceram, *vīcerāmus*, *vīcerant*, *vīcerās*, *vīcerat*, *vīcerātis*, *vīcēre*, *vīcerim*, *vīcerimus*, *vīcerīmus*, *vīcerint*, *vīceris*, *vīcerīs*, *vīcerit*, *vīceritis*, *vīcerītis*, *vīcerō*, *vīcērunt*, *vīcī*, *vīcimus*, *vīcisse*, *vīcissem*, *vīcissēmus*, *vīcissent*, *vīcissēs*, *vīcisset*, *vīcissētis*, *vīcistī*, *vīcistis*, *vīcit*, *vīcta*, *vīctā*, *vīctae*, *vīctam*, *vīctārum*, *vīctās*, *vīcte*, *vīctī*, *vīctīs*, *vīctō*, *vīctōrum*, *vīctōs*, *victū*, *victum*, *vīctum*, *victūrus*, *victus*, *vīctus*, *vincam*, *vincāminī*, *vincāmur*, *vincāmus*, *vincant*, *vincantur*, *vincar*, *vincāre*, *vincāris*, *vincās*, *vincat*, *vincātis*, *vincātur*, *vince*, *vincēbam*, *vincēbāminī*, *vincēbāmur*, *vincēbāmus*, *vincēbant*, *vincēbantur*, *vincēbar*, *vincēbāre*, *vincēbāris*, *vincēbās*, *vincēbat*, *vincēbātis*, *vincēbātur*, *vincēminī*, *vincēmur*, *vincēmus*, *vincendī*, *vincendō*, *vincendum*, *vincendus*, *vincēns*, *vincent*, *vincentur*, *vincere*, *vincēre*, *vincerem*, *vincerēminī*, *vincerēmur*, *vincerēmus*, *vincerent*, *vincerentur*, *vincerer*, *vincerēre*, *vincerēris*, *vincerēs*, *vinceret*, *vincerētis*, *vincerētur*, *vinceris*, *vincēris*, *vincēs*, *vincet*, *vincētis*, *vincētur*, *vincī*, *vinciminī*, *vincimur*, *vincimus*, *vincis*, *vincit*, *vincite*, *vincitis*, *vincitō*, *vincitor*, *vincitōte*, *vincitur*, *vincō*, *vincor*, *vincunt*, *vincuntō*, *vincuntor*, *vincuntur* — “conquer”;

vīsa, *vīsā*, *vīsae*, *vīsam*, *vīsāminī*, *vīsāmur*, *vīsāmus*, *vīsant*, *vīsantur*, *vīsar*, *vīsāre*, *vīsāris*, *vīsārum*, *vīsās*, *vīsat*, *vīsātis*, *vīsātur*, *vīse*, *vīsēbam*, *vīsēbāminī*, *vīsēbāmur*, *vīsēbāmus*, *vīsēbant*, *vīsēbantur*, *vīsēbar*, *vīsēbāre*, *vīsēbāris*, *vīsēbās*, *vīsēbat*, *vīsēbātis*, *vīsēbātur*, *vīsēminī*, *vīsēmur*, *vīsēmus*, *vīsendī*, *vīsendō*, *vīsendum*, *vīsendus*, *vīsēns*, *vīsent*, *vīsentur*, *vīseram*, *vīserāmus*, *vīserant*, *vīserās*, *vīserat*, *vīserātis*, *vīsere*, *vīsēre*, *vīserem*, *vīserēminī*, *vīserēmur*, *vīserēmus*, *vīserent*, *vīserentur*, *vīserer*, *vīserēre*, *vīserēris*, *vīserēs*, *vīseret*, *vīserētis*, *vīserētur*, *vīserim*, *vīserimus*, *vīserīmus*, *vīserint*, *vīseris*, *vīserīs*, *vīsēris*, *vīserit*, *vīseritis*, *vīserītis*, *vīserō*, *vīsērunt*, *vīsēs*, *vīset*, *vīsētis*, *vīsētur*, *vīsī*, *vīsiminī*, *vīsimur*, *vīsimus*, *vīsis*, *vīsīs*, *vīsisse*, *vīsissem*, *vīsissēmus*, *vīsissent*, *vīsissēs*, *vīsisset*, *vīsissētis*, *vīsistī*, *vīsistis*, *vīsit*, *vīsite*, *vīsitis*, *vīsitō*, *vīsitor*, *vīsitōte*, *vīsitur*, *vīsō*, *vīsor*, *vīsōrum*, *vīsōs*, *vīsū*, *vīsum*, *vīsunt*, *vīsuntō*, *vīsuntor*, *vīsuntur*, *vīsūrus*, *vīsus* — “visit”;

vomam, *vomāminī*, *vomāmur*, *vomāmus*, *vomant*, *vomantur*, *vomar*, *vomāre*, *vomāris*, *vomās*, *vomat*, *vomātis*, *vomātur*, *vome*, *vomēbam*, *vomēbāminī*, *vomēbāmur*, *vomēbāmus*, *vomēbant*, *vomēbantur*, *vomēbar*, *vomēbāre*, *vomēbāris*, *vomēbās*, *vomēbat*, *vomēbātis*, *vomēbātur*, *vomēminī*, *vomēmur*, *vomēmus*, *vomendī*, *vomendō*, *vomendum*, *vomendus*, *vomēns*, *voment*, *vomentur*, *vomere*, *vomēre*, *vomerem*, *vomerēminī*, *vomerēmur*, *vomerēmus*, *vomerent*, *vomerentur*, *vomerer*, *vomerēre*, *vomerēris*, *vomerēs*, *vomeret*, *vomerētis*, *vomerētur*, *vomeris*, *vomēris*, *vomēs*, *vomet*, *vomētis*, *vomētur*, *vomī*, *vomiminī*, *vomimur*, *vomimus*, *vomis*, *vomit*, *vomita*, *vomitā*, *vomitae*, *vomitam*, *vomitārum*, *vomitās*, *vomite*, *vomitī*, *vomitis*, *vomitīs*, *vomitō*, *vomitor*, *vomitōrum*, *vomitōs*, *vomitōte*, *vomitū*, *vomitum*, *vomitur*, *vomitūrus*, *vomitus*, *vomō*, *vomor*, *vomueram*, *vomuerāmus*, *vomuerant*, *vomuerās*, *vomuerat*, *vomuerātis*, *vomuēre*, *vomuerim*, *vomuerimus*, *vomuerīmus*, *vomuerint*, *vomueris*, *vomuerīs*, *vomuerit*, *vomueritis*, *vomuerītis*, *vomuerō*, *vomuērunt*, *vomuī*, *vomuimus*, *vomuisse*, *vomuissem*, *vomuissēmus*, *vomuissent*, *vomuissēs*, *vomuisset*, *vomuissētis*, *vomuistī*, *vomuistis*, *vomuit*, *vomunt*, *vomuntō*, *vomuntor*, *vomuntur* — “vomit”.

`https://en.wikipedia.org/wiki/Latin_conjugation#Fourth_conjugation`

aperī, *aperiam*, *aperiāminī*, *aperiāmur*, *aperiāmus*, *aperiant*, *aperiantur*, *aperiar*, *aperiāre*, *aperiāris*, *aperiās*, *aperiat*, *aperiātis*, *aperiātur*, *aperiēbam*, *aperiēbāminī*, *aperiēbāmur*, *aperiēbāmus*, *aperiēbant*, *aperiēbantur*, *aperiēbar*, *aperiēbāre*, *aperiēbāris*, *aperiēbās*, *aperiēbat*, *aperiēbātis*, *aperiēbātur*, *aperiēminī*, *aperiēmur*, *aperiēmus*, *aperiendī*, *aperiendō*, *aperiendum*, *aperiendus*, *aperiēns*, *aperient*, *aperientur*, *aperiēre*, *aperiēris*, *aperiēs*, *aperiet*, *aperiētis*, *aperiētur*, *aperīminī*, *aperīmur*, *aperīmus*, *aperiō*, *aperior*, *aperīre*, *aperīrem*, *aperīrēminī*, *aperīrēmur*, *aperīrēmus*, *aperīrent*, *aperīrentur*, *aperīrer*, *aperīrēre*, *aperīrēris*, *aperīrēs*, *aperīret*, *aperīrētis*, *aperīrētur*, *aperīrī*, *aperīris*, *aperīs*, *aperit*, *aperīte*, *aperītis*, *aperītō*, *aperītor*, *aperītōte*, *aperītur*, *aperiunt*, *aperiuntō*, *aperiuntor*, *aperiuntur*, *aperta*, *apertā*, *apertae*, *apertam*, *apertārum*, *apertās*, *aperte*, *apertī*, *apertīs*, *apertō*, *apertōrum*, *apertōs*, *apertū*, *apertum*, *apertūrus*, *apertus*, *aperueram*, *aperuerāmus*, *aperuerant*, *aperuerās*, *aperuerat*, *aperuerātis*, *aperuēre*, *aperuerim*, *aperuerimus*, *aperuerīmus*, *aperuerint*, *aperueris*, *aperuerīs*, *aperuerit*, *aperueritis*, *aperuerītis*, *aperuerō*, *aperuērunt*, *aperuī*, *aperuimus*, *aperuisse*, *aperuissem*, *aperuissēmus*, *aperuissent*, *aperuissēs*, *aperuisset*, *aperuissētis*, *aperuistī*, *aperuistis*, *aperuit* — “open”;

audī, *audiam*, *audiāminī*, *audiāmur*, *audiāmus*, *audiant*, *audiantur*, *audiar*, *audiāre*, *audiāris*, *audiās*, *audiat*, *audiātis*, *audiātur*, *audiēbam*, *audiēbāminī*, *audiēbāmur*, *audiēbāmus*, *audiēbant*, *audiēbantur*, *audiēbar*, *audiēbāre*, *audiēbāris*, *audiēbās*, *audiēbat*, *audiēbātis*, *audiēbātur*, *audiēminī*, *audiēmur*, *audiēmus*, *audiendī*, *audiendō*, *audiendum*, *audiendus*, *audiēns*, *audient*, *audientur*, *audiēre*, *audiēris*, *audiēs*, *audiet*, *audiētis*, *audiētur*, *audīminī*, *audīmur*, *audīmus*, *audiō*, *audior*, *audīre*, *audīrem*, *audīrēminī*, *audīrēmur*, *audīrēmus*, *audīrent*, *audīrentur*, *audīrer*, *audīrēre*, *audīrēris*, *audīrēs*, *audīret*, *audīrētis*, *audīrētur*, *audīrī*, *audīris*, *audīs*, *audit*, *audīta*, *audītā*, *audītae*, *audītam*, *audītārum*, *audītās*, *audīte*, *audītī*, *audītis*, *audītīs*, *audītō*, *audītor*, *audītōrum*, *audītōs*, *audītōte*, *audītū*, *audītum*, *audītur*, *audītūrus*, *audītus*, *audiunt*, *audiuntō*, *audiuntor*, *audiuntur*, *audīveram*, *audīverāmus*, *audīverant*, *audīverās*, *audīverat*, *audīverātis*, *audīvēre*, *audīverim*, *audīverimus*, *audīverīmus*, *audīverint*, *audīveris*, *audīverīs*, *audīverit*, *audīveritis*, *audīverītis*, *audīverō*, *audīvērunt*, *audīvī*, *audīvimus*, *audīvisse*, *audīvissem*, *audīvissēmus*, *audīvissent*, *audīvissēs*, *audīvisset*, *audīvissētis*, *audīvistī*, *audīvistis*, *audīvit* — “hear”;

mūnī, *mūniam*, *mūniāminī*, *mūniāmur*, *mūniāmus*, *mūniant*, *mūniantur*, *mūniar*, *mūniāre*, *mūniāris*, *mūniās*, *mūniat*, *mūniātis*, *mūniātur*, *mūniēbam*, *mūniēbāminī*, *mūniēbāmur*, *mūniēbāmus*, *mūniēbant*, *mūniēbantur*, *mūniēbar*, *mūniēbāre*, *mūniēbāris*, *mūniēbās*, *mūniēbat*, *mūniēbātis*, *mūniēbātur*, *mūniēminī*, *mūniēmur*, *mūniēmus*, *mūniendī*, *mūniendō*, *mūniendum*, *mūniendus*, *mūniēns*, *mūnient*, *mūnientur*, *mūniēre*, *mūniēris*, *mūniēs*, *mūniet*, *mūniētis*, *mūniētur*, *mūnīminī*, *mūnīmur*, *mūnīmus*, *mūniō*, *mūnior*, *mūnīre*, *mūnīrem*, *mūnīrēminī*, *mūnīrēmur*, *mūnīrēmus*, *mūnīrent*, *mūnīrentur*, *mūnīrer*, *mūnīrēre*, *mūnīrēris*, *mūnīrēs*, *mūnīret*, *mūnīrētis*, *mūnīrētur*, *mūnīrī*, *mūnīris*, *mūnīs*, *mūnit*, *mūnita*, *mūnitā*, *mūnitae*, *mūnitam*, *mūnitārum*, *mūnitās*, *mūnite*, *mūnīte*, *mūnitī*, *mūnitīs*, *mūnītis*, *mūnitō*, *mūnītō*, *mūnītor*, *mūnitōrum*, *mūnitōs*, *mūnītōte*, *mūnītū*, *mūnitum*, *mūnītum*, *mūnītur*, *mūnītūrus*, *mūnitus*, *mūnītus*, *mūniunt*, *mūniuntō*, *mūniuntor*, *mūniuntur*, *mūnīveram*, *mūnīverāmus*, *mūnīverant*, *mūnīverās*, *mūnīverat*, *mūnīverātis*, *mūnīvēre*, *mūnīverim*, *mūnīverimus*, *mūnīverīmus*, *mūnīverint*, *mūnīveris*, *mūnīverīs*, *mūnīverit*, *mūnīveritis*, *mūnīverītis*, *mūnīverō*, *mūnīvērunt*, *mūnīvī*, *mūnīvimus*, *mūnīvisse*, *mūnīvissem*, *mūnīvissēmus*, *mūnīvissent*, *mūnīvissēs*, *mūnīvisset*, *mūnīvissētis*, *mūnīvistī*, *mūnīvistis*, *mūnīvit* — “fortify”;

oriāminī, *oriāmur*, *oriantur*, *oriar*, *oriāre*, *oriāris*, *oriātur*, *oriēbāminī*, *oriēbāmur*, *oriēbantur*, *oriēbar*, *oriēbāre*, *oriēbāris*, *oriēbātur*, *oriēminī*, *oriēmur*, *oriendī*, *oriendō*, *oriendum*, *oriendus*, *oriēns*, *orientur*, *oriēre*, *oriēris*, *oriētur*, *orīminī*, *orīmur*, *orior*, *orīre*, *orīrēminī*, *orīrēmur*, *orīrentur*, *orīrer*, *orīrēre*, *orīrēris*, *orīrētur*, *orīrī*, *orīris*, *orītor*, *orītur*, *oritūrus*, *oriuntor*, *oriuntur*, *orta*, *ortā*, *ortae*, *ortam*, *ortārum*, *ortās*, *orte*, *ortī*, *ortīs*, *ortō*, *ortōrum*, *ortōs*, *ortū*, *ortum*, *ortus* — “rise”;

saepī, *saepiam*, *saepiāminī*, *saepiāmur*, *saepiāmus*, *saepiant*, *saepiantur*, *saepiar*, *saepiāre*, *saepiāris*, *saepiās*, *saepiat*, *saepiātis*, *saepiātur*, *saepiēbam*, *saepiēbāminī*, *saepiēbāmur*, *saepiēbāmus*, *saepiēbant*, *saepiēbantur*, *saepiēbar*, *saepiēbāre*, *saepiēbāris*, *saepiēbās*, *saepiēbat*, *saepiēbātis*, *saepiēbātur*, *saepiēminī*, *saepiēmur*, *saepiēmus*, *saepiendī*, *saepiendō*, *saepiendum*, *saepiendus*, *saepiēns*, *saepient*, *saepientur*, *saepiēre*, *saepiēris*, *saepiēs*, *saepiet*, *saepiētis*, *saepiētur*, *saepīminī*, *saepīmur*, *saepīmus*, *saepiō*, *saepior*, *saepīre*, *saepīrem*, *saepīrēminī*, *saepīrēmur*, *saepīrēmus*, *saepīrent*, *saepīrentur*, *saepīrer*, *saepīrēre*, *saepīrēris*, *saepīrēs*, *saepīret*, *saepīrētis*, *saepīrētur*, *saepīrī*, *saepīris*, *saepīs*, *saepit*, *saepīte*, *saepītis*, *saepītō*, *saepītor*, *saepītōte*, *saepītur*, *saepiunt*, *saepiuntō*, *saepiuntor*, *saepiuntur*, *saepseram*, *saepserāmus*, *saepserant*, *saepserās*, *saepserat*, *saepserātis*, *saepsēre*, *saepserim*, *saepserimus*, *saepserīmus*, *saepserint*, *saepseris*, *saepserīs*, *saepserit*, *saepseritis*, *saepserītis*, *saepserō*, *saepsērunt*, *saepsī*, *saepsimus*, *saepsisse*, *saepsissem*, *saepsissēmus*, *saepsissent*, *saepsissēs*, *saepsisset*, *saepsissētis*, *saepsistī*, *saepsistis*, *saepsit*, *saepta*, *saeptā*, *saeptae*, *saeptam*, *saeptārum*, *saeptās*, *saepte*, *saeptī*, *saeptīs*, *saeptō*, *saeptōrum*, *saeptōs*, *saeptū*, *saeptum*, *saeptūrus*, *saeptus* — “surround”;

sancī, *sanciam*, *sanciāminī*, *sanciāmur*, *sanciāmus*, *sanciant*, *sanciantur*, *sanciar*, *sanciāre*, *sanciāris*, *sanciās*, *sanciat*, *sanciātis*, *sanciātur*, *sanciēbam*, *sanciēbāminī*, *sanciēbāmur*, *sanciēbāmus*, *sanciēbant*, *sanciēbantur*, *sanciēbar*, *sanciēbāre*, *sanciēbāris*, *sanciēbās*, *sanciēbat*, *sanciēbātis*, *sanciēbātur*, *sanciēminī*, *sanciēmur*, *sanciēmus*, *sanciendī*, *sanciendō*, *sanciendum*, *sanciendus*, *sanciēns*, *sancient*, *sancientur*, *sanciēre*, *sanciēris*, *sanciēs*, *sanciet*, *sanciētis*, *sanciētur*, *sancīminī*, *sancīmur*, *sancīmus*, *sanciō*, *sancior*, *sancīre*, *sancīrem*, *sancīrēminī*, *sancīrēmur*, *sancīrēmus*, *sancīrent*, *sancīrentur*, *sancīrer*, *sancīrēre*, *sancīrēris*, *sancīrēs*, *sancīret*, *sancīrētis*, *sancīrētur*, *sancīrī*, *sancīris*, *sancīs*, *sancit*, *sancīte*, *sancītis*, *sancītō*, *sancītor*, *sancītōte*, *sancītur*, *sanciunt*, *sanciuntō*, *sanciuntor*, *sanciuntur*, *sāncta*, *sānctā*, *sānctae*, *sānctam*, *sānctārum*, *sānctās*, *sāncte*, *sānctī*, *sānctīs*, *sānctō*, *sānctōrum*, *sānctōs*, *sānctū*, *sānctum*, *sānctūrus*, *sānctus*, *sānxeram*, *sānxerāmus*, *sānxerant*, *sānxerās*, *sānxerat*, *sānxerātis*, *sānxēre*, *sānxerim*, *sānxerimus*, *sānxerīmus*, *sānxerint*, *sānxeris*, *sānxerīs*, *sānxerit*, *sānxeritis*, *sānxerītis*, *sānxerō*, *sānxērunt*, *sānxī*, *sānximus*, *sānxisse*, *sānxissem*, *sānxissēmus*, *sānxissent*, *sānxissēs*, *sānxisset*, *sānxissētis*, *sānxistī*, *sānxistis*, *sānxit* — “confirm”;

sēnsa, *sēnsā*, *sēnsae*, *sēnsam*, *sēnsārum*, *sēnsās*, *sēnse*, *sēnseram*, *sēnserāmus*, *sēnserant*, *sēnserās*, *sēnserat*, *sēnserātis*, *sēnsēre*, *sēnserim*, *sēnserimus*, *sēnserīmus*, *sēnserint*, *sēnseris*, *sēnserīs*, *sēnserit*, *sēnseritis*, *sēnserītis*, *sēnserō*, *sēnsērunt*, *sēnsī*, *sēnsimus*, *sēnsīs*, *sēnsisse*, *sēnsissem*, *sēnsissēmus*, *sēnsissent*, *sēnsissēs*, *sēnsisset*, *sēnsissētis*, *sēnsistī*, *sēnsistis*, *sēnsit*, *sēnsō*, *sēnsōrum*, *sēnsōs*, *sēnsū*, *sēnsum*, *sēnsūrus*, *sēnsus*, *sentī*, *sentiam*, *sentiāminī*, *sentiāmur*, *sentiāmus*, *sentiant*, *sentiantur*, *sentiar*, *sentiāre*, *sentiāris*, *sentiās*, *sentiat*, *sentiātis*, *sentiātur*, *sentiēbam*, *sentiēbāminī*, *sentiēbāmur*, *sentiēbāmus*, *sentiēbant*, *sentiēbantur*, *sentiēbar*, *sentiēbāre*, *sentiēbāris*, *sentiēbās*, *sentiēbat*, *sentiēbātis*, *sentiēbātur*, *sentiēminī*, *sentiēmur*, *sentiēmus*, *sentiendī*, *sentiendō*, *sentiendum*, *sentiendus*, *sentiēns*, *sentient*, *sentientur*, *sentiēre*, *sentiēris*, *sentiēs*, *sentiet*, *sentiētis*, *sentiētur*, *sentīminī*, *sentīmur*, *sentīmus*, *sentiō*, *sentior*, *sentīre*, *sentīrem*, *sentīrēminī*, *sentīrēmur*, *sentīrēmus*, *sentīrent*, *sentīrentur*, *sentīrer*, *sentīrēre*, *sentīrēris*, *sentīrēs*, *sentīret*, *sentīrētis*, *sentīrētur*, *sentīrī*, *sentīris*, *sentīs*, *sentit*, *sentīte*, *sentītis*, *sentītō*, *sentītor*, *sentītōte*, *sentītur*, *sentiunt*, *sentiuntō*, *sentiuntor*, *sentiuntur* — “feel”;

vēneram, *vēnerāmus*, *vēnerant*, *vēnerās*, *vēnerat*, *vēnerātis*, *vēnēre*, *vēnerim*, *vēnerimus*, *vēnerīmus*, *vēnerint*, *vēneris*, *vēnerīs*, *vēnerit*, *vēneritis*, *vēnerītis*, *vēnerō*, *vēnērunt*, *venī*, *vēnī*, *veniam*, *veniāmus*, *veniant*, *veniās*, *veniat*, *veniātis*, *veniātur*, *veniēbam*, *veniēbāmus*, *veniēbant*, *veniēbās*, *veniēbat*, *veniēbātis*, *veniēbātur*, *veniēmus*, *veniendī*, *veniendō*, *veniendum*, *veniendus*, *veniēns*, *venient*, *veniēs*, *veniet*, *veniētis*, *veniētur*, *venīmus*, *vēnimus*, *veniō*, *venīre*, *venīrem*, *venīrēmus*, *venīrent*, *venīrēs*, *venīret*, *venīrētis*, *venīrētur*, *venīrī*, *venīs*, *vēnisse*, *vēnissem*, *vēnissēmus*, *vēnissent*, *vēnissēs*, *vēnisset*, *vēnissētis*, *vēnistī*, *vēnistis*, *venit*, *vēnit*, *venīte*, *venītis*, *venītō*, *venītōte*, *venītur*, *veniunt*, *veniuntō*, *venta*, *ventā*, *ventae*, *ventam*, *ventārum*, *ventās*, *vente*, *ventī*, *ventīs*, *ventō*, *ventōrum*, *ventōs*, *ventū*, *ventum*, *ventūrus*, *ventus* — “come”.

As we throw each group of test verbs into our algorithm, the verb forms from the same conjugation groups are clustered correctly.

We note, however, that the inflected forms of *vetus* “old” and *vetō* “forbid” will be partly conflated if they appear simultaneously:

{*vetā*, *vetābam*, *vetābāminī*, *vetābāmur*, *vetābāmus*, *vetābant*, *vetābantur*, *vetābar*, *vetābāre*, *vetābāris*, *vetābās*, *vetābat*, *vetābātis*, *vetābātur*, *vetābere*, *vetāberis*, *vetābiminī*, *vetābimur*, *vetābimus*, *vetābis*, *vetābit*, *vetābitis*, *vetābitur*, *vetābō*, *vetābor*, *vetābunt*, *vetābuntur*, *vetāminī*, *vetāmur*, *vetāmus*, *vetandī*, *vetandō*, *vetandum*, *vetandus*, *vetāns*, *vetant*, *vetantō*, *vetantor*, *vetantur*, *vetāre*, *vetārem*, *vetārēminī*, *vetārēmur*, *vetārēmus*, *vetārent*, *vetārentur*, *vetārer*, *vetārēre*, *vetārēris*, *vetārēs*, *vetāret*, *vetārētis*, *vetārētur*, *vetārī*, *vetāris*, *vetās*, *vetat*, *vetāte*, *vetātis*, *vetātō*, *vetātor*, *vetātōte*, *vetātur*, *vetem*, *vetēminī*, *vetēmur*, *vetēmus*, *vetent*, *vetentur*, *veter*, *vetera*, *vetere*, *vetēre*, *veterem*, *veterēs*, *veterī*, *veteribus*, *veteris*, *vetēris*, *veterrima*, *veterrimā*, *veterrimae*, *veterrimam*,

veterrimārum, *veterrimās*, *veterrime*, *veterrimī*, *veterrimīs*, *veterrimō*, *veterrimōrum*, *veterrimōs*, *veterrimum*, *veterrimus*, *veterum*, *vetēs*, *vetet*, *vetētis*, *vetētur*, *vetita*, *vetitā*, *vetitae*, *vetitam*, *vetitārum*, *vetitās*, *vetite*, *vetitī*, *vetitīs*, *vetitō*, *vetitōrum*, *vetitōs*, *vetitū*, *vetitum*, *vetitūrus*, *vetitus*, *vetō*, *vetor*, *vetueram*, *vetuerāmus*, *vetuerant*, *vetuerās*, *vetuerat*, *vetuerātis*, *vetuēre*, *vetuerim*, *vetuerimus*, *vetuerīmus*, *vetuerint*, *vetueris*, *vetuerīs*, *vetuerit*, *vetueritis*, *vetuerītis*, *vetuerō*, *vetuērunt*, *vetuī*, *vetuimus*, *vetuisse*, *vetuissem*, *vetuissēmus*, *vetuissent*, *vetuissēs*, *vetuisset*, *vetuissētis*, *vetuistī*, *vetuistis*, *vetuit*, *vetus*, *vetustissimī*, *vetustissimīs*},

{*vetustior*, *vetustiōra*, *vetustiōre*, *vetustiōrem*, *vetustiōrēs*, *vetustiōrī*, *vetustiōribus*, *vetustiōris*, *vetustiōrum*, *vetustissima*, *vetustissimā*, *vetustissimae*, *vetustissimam*, *vetustissimārum*, *vetustissimās*, *vetustissime*, *vetustissimō*, *vetustissimōrum*, *vetustissimōs*, *vetustissimum*, *vetustissimus*, *vetustius*}.

Such unfortunate conflations seem unavoidable. Observe that *veteris* is the genitive singular form of *vetus*, while *vetēris* is the second-person singular present passive subjunctive of *vetō*. Both *veteris* and *vetēris* share the same essential root *vet*, and they are practically indistinguishable without context, if the document in question does not employ the diacritical mark for long vowels. Some other Latin words also behave similarly: *portā* is both the ablative singular of *porta* "gate" and the singular present active imperative of *portō* "convey".

Example 6.29.3. We pick a list of typical irregular verbs from the Wikipedia links indicated below.

`https://en.wikipedia.org/wiki/Latin_conjugation#Irregular_verbs`

`https://en.wikipedia.org/wiki/Latin_conjugation#Third_conjugation_.E2.80.93i.C5.8D_verbs`

cape, *capere*, *caperem*, *caperēminī*, *caperēmur*, *caperēmus*, *caperent*, *caperentur*, *caperer*, *caperēre*, *caperēris*, *caperēs*, *caperet*, *caperētis*, *caperētur*, *caperis*, *capī*, *capiam*, *capiāminī*, *capiāmur*, *capiāmus*, *capiant*, *capiantur*, *capiar*, *capiāre*, *capiāris*, *capiās*, *capiat*, *capiātis*, *capiātur*, *capiēbam*, *capiēbāminī*, *capiēbāmur*, *capiēbāmus*, *capiēbant*, *capiēbantur*, *capiēbar*, *capiēbāre*, *capiēbāris*, *capiēbās*, *capiēbat*, *capiēbātis*, *capiēbātur*, *capiēminī*, *capiēmur*, *capiēmus*, *capiendī*, *capiendō*, *capiendum*, *capiendus*, *capiēns*, *capient*, *capientur*, *capiēre*, *capiēris*, *capiēs*, *capiet*, *capiētis*, *capiētur*, *capiminī*, *capimur*, *capimus*, *capiō*, *capior*, *capis*, *capit*, *capite*, *capitis*, *capitō*, *capitor*, *capitōte*, *capitur*, *capiunt*, *capiuntō*, *capiuntor*, *capiuntur*, *capta*, *captā*, *captae*, *captam*, *captārum*, *captās*, *capte*, *captī*, *captīs*, *captō*, *captōrum*, *captōs*, *captū*, *captum*, *captūrus*, *captus*, *cēperam*, *cēperāmus*, *cēperant*, *cēperās*, *cēperat*, *cēperātis*, *cēpēre*, *cēperim*, *cēperimus*, *cēperīmus*, *cēperint*, *cēperis*, *cēperīs*, *cēperit*, *cēperitis*, *cēperītis*, *cēperō*, *cēpērunt*, *cēpī*, *cēpimus*, *cēpisse*, *cēpissem*, *cēpissēmus*, *cēpissent*, *cēpissēs*, *cēpisset*, *cēpissētis*, *cēpistī*, *cēpistis*, *cēpit* — "capture";

cupe, *cupere*, *cuperem*, *cuperēminī*, *cuperēmur*, *cuperēmus*, *cuperent*, *cuperentur*, *cuperer*, *cuperēre*, *cuperēris*, *cuperēs*, *cuperet*, *cuperētis*, *cuperētur*, *cuperis*, *cupī*, *cupiam*, *cupiāminī*, *cupiāmur*, *cupiāmus*, *cupiant*, *cupiantur*, *cupiar*, *cupiāre*, *cupiāris*, *cupiās*, *cupiat*, *cupiātis*, *cupiātur*, *cupiēbam*, *cupiēbāminī*, *cupiēbāmur*, *cupiēbāmus*, *cupiēbant*, *cupiēbantur*, *cupiēbar*, *cupiēbāre*, *cupiēbāris*, *cupiēbās*, *cupiēbat*, *cupiēbātis*, *cupiēbātur*, *cupiēminī*, *cupiēmur*, *cupiēmus*, *cupiendī*, *cupiendō*, *cupiendum*, *cupiendus*, *cupiēns*, *cupient*, *cupientur*, *cupiēre*, *cupiēris*, *cupiēs*, *cupiet*, *cupiētis*, *cupiētur*, *cupiminī*, *cupimur*, *cupimus*, *cupiō*, *cupior*, *cupis*, *cupit*, *cupīta*, *cupītā*, *cupītae*, *cupītam*, *cupītārum*, *cupītās*, *cupite*, *cupīte*, *cupītī*, *cupitis*, *cupītīs*, *cupitō*, *cupītō*, *cupitor*, *cupītōrum*, *cupītōs*, *cupitōte*, *cupītū*, *cupītum*, *cupitur*, *cupītūrus*, *cupītus*, *cupiunt*, *cupiuntō*, *cupiuntor*, *cupiuntur*, *cupīveram*, *cupīverāmus*, *cupīverant*, *cupīverās*, *cupīverat*, *cupīverātis*, *cupīvēre*, *cupīverim*, *cupīverimus*, *cupīverīmus*, *cupīverint*, *cupīveris*, *cupīverīs*, *cupīverit*, *cupīveritis*, *cupīverītis*, *cupīverō*, *cupīvērunt*, *cupīvī*, *cupīvimus*, *cupīvisse*, *cupīvissem*, *cupīvissēmus*, *cupīvissent*, *cupīvissēs*, *cupīvisset*, *cupīvissētis*, *cupīvistī*, *cupīvistis*, *cupīvit* — "desire";

eam, *eamus*, *eant*, *eas*, *eat*, *eatis*, *eatur*, *eo*, *eundi*, *eundo*, *eundum*, *eundus*, *eunt*, *eunto*, *i*, *ibam*, *ibamus*, *ibant*, *ibas*, *ibat*, *ibatis*, *ibatur*, *ibimus*, *ibis*, *ibit*, *ibitis*, *ibitur*, *ibo*, *ibunt*, *iens*, *ieram*, *ieramus*, *ierant*, *ieras*, *ierat*, *ieratis*, *iere*, *ierim*, *ierimus*, *ierint*, *ieris*, *ierit*, *ieritis*, *iero*, *ierunt*, *ii*, *iimus*, *iit*, *imus*, *ire*, *irem*, *iremus*, *irent*, *ires*, *iret*, *iretis*, *iretur*, *iri*, *is*, *isse*, *issem*, *issemus*, *issent*, *isses*, *isset*, *issetis*, *isti*, *istis*, *it*, *ite*, *itis*, *ito*, *itote*, *itu*, *itum*, *itur*, *iturus*, *itus*, *ivi*, *ivisti*, *ivit* — "go";

edam, *edāminī*, *edāmur*, *edāmus*, *edant*, *edantur*, *edar*, *edāre*, *edāris*, *edās*, *edat*, *edātis*, *edātur*, *ede*, *edēbam*, *edēbāminī*, *edēbāmur*, *edēbāmus*, *edēbant*, *edēbantur*, *edēbar*, *edēbāre*, *edēbāris*, *edēbās*, *edēbat*, *edēbātis*, *edēbātur*, *edēminī*, *edēmur*, *edēmus*, *edendī*, *edendō*, *edendum*, *edendus*, *edēns*, *edent*, *edentur*, *ēderam*, *ēderāmus*, *ēderant*, *ēderās*, *ēderat*, *ēderātis*, *edere*, *edēre*, *ēdēre*, *ederem*, *ederēminī*, *ederēmur*, *ederēmus*, *ederent*, *ederentur*, *ederer*, *ederēre*, *ederēris*, *ederēs*, *ederet*, *ederētis*, *ederētur*, *ēderim*, *ēderimus*, *ēderīmus*, *ēderint*, *ederis*, *edēris*, *ēderis*, *ēderīs*, *ēderit*, *ēderitis*, *ēderītis*, *ēderō*, *ēdērunt*, *edēs*, *edet*, *edētis*, *edētur*, *edī*, *ēdī*, *edim*, *ediminī*, *edimur*, *edimus*,

edīmus, *ēdimus*, *edint*, *edis*, *edīs*, *ēdisse*, *ēdissem*, *ēdissēmus*, *ēdissent*, *ēdissēs*, *ēdisset*, *ēdissētis*, *ēdistī*, *ēdistis*, *edit*, *ēdit*, *edite*, *editis*, *edītis*, *editō*, *editor*, *editōte*, *editur*, *edō*, *edor*, *edunt*, *eduntō*, *eduntor*, *eduntur*, *ēs*, *ēsa*, *ēsā*, *ēsae*, *ēsam*, *ēsārum*, *ēsās*, *ēse*, *ēsī*, *ēsīs*, *ēsō*, *ēsōrum*, *ēsōs*, *esse*, *ēsse*, *ēssem*, *ēssēmus*, *ēssent*, *ēssēs*, *ēsset*, *ēssētis*, *ēst*, *ēste*, *ēstis*, *ēstō*, *ēstōte*, *ēstur*, *ēsū*, *ēsum*, *ēsūrus*, *ēsus* — “eat”;

fer, *feram*, *ferāminī*, *ferāmur*, *ferāmus*, *ferant*, *ferantur*, *ferar*, *ferāre*, *ferāris*, *ferās*, *ferat*, *ferātis*, *ferātur*, *ferēbam*, *ferēbāminī*, *ferēbāmur*, *ferēbāmus*, *ferēbant*, *ferēbantur*, *ferēbar*, *ferēbāre*, *ferēbāris*, *ferēbās*, *ferēbat*, *ferēbātis*, *ferēbātur*, *ferēminī*, *ferēmur*, *ferēmus*, *ferendī*, *ferendō*, *ferendum*, *ferendus*, *ferēns*, *ferent*, *ferentur*, *ferēre*, *ferēris*, *ferēs*, *feret*, *ferētis*, *ferētur*, *feriminī*, *ferimur*, *ferimus*, *ferō*, *feror*, *ferre*, *ferrem*, *ferrēminī*, *ferrēmur*, *ferrēmus*, *ferrent*, *ferrentur*, *ferrer*, *ferrēre*, *ferrēris*, *ferrēs*, *ferret*, *ferrētis*, *ferrētur*, *ferrī*, *ferris*, *fers*, *fert*, *ferte*, *fertis*, *fertō*, *fertor*, *fertōte*, *fertur*, *ferunt*, *feruntō*, *feruntor*, *feruntur*, *lāta*, *lātā*, *lātae*, *lātam*, *lātārum*, *lātās*, *lāte*, *lātī*, *lātīs*, *lātō*, *lātōrum*, *lātōs*, *lātū*, *lātum*, *lātūrus*, *lātus*, *tuleram*, *tulerāmus*, *tulerant*, *tulerās*, *tulerat*, *tulerātis*, *tulēre*, *tulerim*, *tulerimus*, *tulerīmus*, *tulerint*, *tuleris*, *tulerīs*, *tulerit*, *tuleritis*, *tulerītis*, *tulerō*, *tulērunt*, *tulī*, *tulimus*, *tulisse*, *tulissem*, *tulissēmus*, *tulissent*, *tulissēs*, *tulisset*, *tulissētis*, *tulistī*, *tulistis*, *tulit* — “carry”;

mālam, *mālēbam*, *mālēbāmus*, *mālēbant*, *mālēbās*, *mālēbat*, *mālēbātis*, *mālēmus*, *mālent*, *mālēs*, *mālet*, *mālētis*, *mālim*, *mālīmus*, *mālint*, *mālīs*, *mālit*, *mālītis*, *mālle*, *māllem*, *māllēmus*, *māllent*, *māllēs*, *māllet*, *māllētis*, *mālō*, *mālueram*, *māluerāmus*, *māluerant*, *māluerās*, *māluerat*, *māluerātis*, *māluēre*, *māluerim*, *māluerimus*, *māluerīmus*, *māluerint*, *mālueris*, *māluerīs*, *māluerit*, *mālueritis*, *māluerītis*, *māluerō*, *māluērunt*, *māluī*, *māluimus*, *māluisse*, *māluissem*, *māluissēmus*, *māluissent*, *māluissēs*, *māluisset*, *māluissētis*, *māluistī*, *māluistis*, *māluit*, *mālumus*, *mālunt*, *māvīs*, *māvult*, *māvultis* — “prefer”;

morere, *morerēminī*, *morerēmur*, *morerentur*, *morerer*, *morerēre*, *morerēris*, *morerētur*, *moreris*, *morī*, *moriāminī*, *moriāmur*, *moriantur*, *moriar*, *moriāre*, *moriāris*, *moriātur*, *moriēbāminī*, *moriēbāmur*, *moriēbantur*, *moriēbar*, *moriēbāre*, *moriēbāris*, *moriēbātur*, *moriēminī*, *moriēmur*, *moriendī*, *moriendō*, *moriendum*, *moriendus*, *moriēns*, *morientur*, *moriēre*, *moriēris*, *moriētur*, *moriminī*, *morimur*, *morior*, *moritor*, *moritur*, *moritūrus*, *moriuntor*, *moriuntur*, *mortua*, *mortuā*, *mortuae*, *mortuam*, *mortuārum*, *mortuās*, *mortue*, *mortuī*, *mortuīs*, *mortuō*, *mortuōrum*, *mortuōs*, *mortuū*, *mortuum*, *mortuus* — “die”;

nōlam, *nōlēbam*, *nōlēbāmus*, *nōlēbant*, *nōlēbās*, *nōlēbat*, *nōlēbātis*, *nōlēmus*, *nōlēns*, *nōlent*, *nōlēs*, *nōlet*, *nōlētis*, *nōlī*, *nōlim*, *nōlīmus*, *nōlint*, *nōlīs*, *nōlit*, *nōlīte*, *nōlītis*, *nōlitō*, *nōlitōte*, *nōlle*, *nōllem*, *nōllēmus*, *nōllent*, *nōllēs*, *nōllet*, *nōllētis*, *nōlō*, *nōlueram*, *nōluerāmus*, *nōluerant*, *nōluerās*, *nōluerat*, *nōluerātis*, *nōluēre*, *nōluerim*, *nōluerimus*, *nōluerīmus*, *nōluerint*, *nōlueris*, *nōluerīs*, *nōluerit*, *nōlueritis*, *nōluerītis*, *nōluerō*, *nōluērunt*, *nōluī*, *nōluimus*, *nōluisse*, *nōluissem*, *nōluissēmus*, *nōluissent*, *nōluissēs*, *nōluisset*, *nōluissētis*, *nōluistī*, *nōluistis*, *nōluit*, *nōlumus*, *nōlunt*, *nōluntō* — “refuse”;

passa, *passā*, *passae*, *passam*, *passārum*, *passās*, *passe*, *passī*, *passīs*, *passō*, *passōrum*, *passōs*, *passū*, *passum*, *passūrus*, *passus*, *patere*, *paterēminī*, *paterēmur*, *paterentur*, *paterer*, *paterēre*, *paterēris*, *paterētur*, *pateris*, *patī*, *patiāminī*, *patiāmur*, *patiantur*, *patiar*, *patiāre*, *patiāris*, *patiātur*, *patiēbāminī*, *patiēbāmur*, *patiēbantur*, *patiēbar*, *patiēbāre*, *patiēbāris*, *patiēbātur*, *patiēminī*, *patiēmur*, *patiendī*, *patiendō*, *patiendum*, *patiendus*, *patiēns*, *patientur*, *patiēre*, *patiēris*, *patiētur*, *patiminī*, *patimur*, *patior*, *patitor*, *patitur*, *patiuntor*, *patiuntur* — “suffer”;

rape, *rapere*, *raperem*, *raperēminī*, *raperēmur*, *raperēmus*, *raperent*, *raperentur*, *raperer*, *raperēre*, *raperēris*, *raperēs*, *raperet*, *raperētis*, *raperētur*, *raperis*, *rapī*, *rapiam*, *rapiāminī*, *rapiāmur*, *rapiāmus*, *rapiant*, *rapiantur*, *rapiar*, *rapiāre*, *rapiāris*, *rapiās*, *rapiat*, *rapiātis*, *rapiātur*, *rapiēbam*, *rapiēbāminī*, *rapiēbāmur*, *rapiēbāmus*, *rapiēbant*, *rapiēbantur*, *rapiēbar*, *rapiēbāre*, *rapiēbāris*, *rapiēbās*, *rapiēbat*, *rapiēbātis*, *rapiēbātur*, *rapiēminī*, *rapiēmur*, *rapiēmus*, *rapiendī*, *rapiendō*, *rapiendum*, *rapiendus*, *rapiēns*, *rapient*, *rapientur*, *rapiēre*, *rapiēris*, *rapiēs*, *rapiet*, *rapiētis*, *rapiētur*, *rapiminī*, *rapimur*, *rapimus*, *rapiō*, *rapior*, *rapis*, *rapit*, *rapite*, *rapitis*, *rapitō*, *rapitor*, *rapitōte*, *rapitur*, *rapiunt*, *rapiuntō*, *rapiuntor*, *rapiuntur*, *rapta*, *raptā*, *raptae*, *raptam*, *raptārum*, *raptās*, *rapte*, *raptī*, *raptīs*, *raptō*, *raptōrum*, *raptōs*, *raptū*, *raptum*, *raptūrus*, *raptus*, *rapueram*, *rapuerāmus*, *rapuerant*, *rapuerās*, *rapuerat*, *rapuerātis*, *rapuēre*, *rapuerim*, *rapuerimus*, *rapuerīmus*, *rapuerint*, *rapueris*, *rapuerīs*, *rapuerit*, *rapueritis*, *rapuerītis*, *rapuerō*, *rapuērunt*, *rapuī*, *rapuimus*, *rapuisse*, *rapuissem*, *rapuissēmus*, *rapuissent*, *rapuissēs*, *rapuisset*, *rapuissētis*, *rapuistī*, *rapuistis*, *rapuit* — “snatch”;

velim, *velīmus*, *velint*, *velīs*, *velit*, *velītis*, *velle*, *vellem*, *vellēmus*, *vellent*, *vellēs*, *vellet*, *vellētis*, *vīs*, *volam*, *volēbam*, *volēbāmus*, *volēbant*, *volēbās*, *volēbat*, *volēbātis*, *volēmus*, *volēns*, *volent*, *volēs*, *volet*, *volētis*, *volō*, *volt*, *voltis*, *volueram*, *voluerāmus*, *voluerant*, *voluerās*, *voluerat*, *voluerātis*, *voluēre*, *voluerim*, *voluerimus*, *voluerīmus*, *voluerint*, *volueris*, *voluerīs*, *voluerit*, *volueritis*, *voluerītis*, *voluerō*, *voluērunt*, *voluī*, *voluimus*, *voluisse*, *voluissem*, *voluissēmus*, *voluissent*, *voluissēs*, *voluisset*, *voluissētis*, *voluistī*, *voluistis*, *voluit*, *volumus*, *volunt*, *vult*, *vultis* — “want”.

The clustering result is satisfactory:

{*cape*, *capere*, *caperem*, *caperēminī*, *caperēmur*, *caperēmus*, *caperent*, *caperentur*, *caperer*, *caperēre*, *caperēris*, *caperēs*, *caperet*, *caperētis*, *caperētur*, *caperis*, *capī*, *capiam*, *capiāminī*, *capiāmur*, *capiāmus*, *capiant*, *capiantur*, *capiar*, *capiāre*, *capiāris*, *capiās*, *capiat*, *capiātis*, *capiātur*, *capiēbam*, *capiēbāminī*, *capiēbāmur*, *capiēbāmus*, *capiēbant*, *capiēbantur*, *capiēbar*, *capiēbāre*, *capiēbāris*, *capiēbās*, *capiēbat*, *capiēbātis*, *capiēbātur*, *capiēminī*, *capiēmur*, *capiēmus*, *capiendī*, *capiendō*, *capiendum*, *capiendus*, *capiēns*, *capient*, *capientur*, *capiēre*, *capiēris*, *capiēs*, *capiet*, *capiētis*, *capiētur*, *capiminī*, *capimur*, *capimus*, *capiō*, *capior*, *capis*, *capit*, *capite*, *capitis*, *capitō*, *capitor*, *capitōte*, *capitur*, *capiunt*, *capiuntō*, *capiuntor*, *capiuntur*, *capta*, *captā*, *captae*, *captam*, *captārum*, *captās*, *capte*, *captī*, *captīs*, *captō*, *captōrum*, *captōs*, *captū*, *captum*, *captūrus*, *captus*, *cēperam*, *cēperāmus*, *cēperant*, *cēperās*, *cēperat*, *cēperātis*, *cēpēre*, *cēperim*, *cēperimus*, *cēperīmus*, *cēperint*, *cēperis*, *cēperīs*, *cēperit*, *cēperitis*, *cēperītis*, *cēperō*, *cēpērunt*, *cēpī*, *cēpimus*, *cēpisse*, *cēpissem*, *cēpissēmus*, *cēpissent*, *cēpissēs*, *cēpisset*, *cēpissētis*, *cēpistī*, *cēpistis*, *cēpit*},

{*cupe*, *cupere*, *cuperem*, *cuperēminī*, *cuperēmur*, *cuperēmus*, *cuperent*, *cuperentur*, *cuperer*, *cuperēre*, *cuperēris*, *cuperēs*, *cuperet*, *cuperētis*, *cuperētur*, *cuperis*, *cupī*, *cupiam*, *cupiāminī*, *cupiāmur*, *cupiāmus*, *cupiant*, *cupiantur*, *cupiar*, *cupiāre*, *cupiāris*, *cupiās*, *cupiat*, *cupiātis*, *cupiātur*, *cupiēbam*, *cupiēbāminī*, *cupiēbāmur*, *cupiēbāmus*, *cupiēbant*, *cupiēbantur*, *cupiēbar*, *cupiēbāre*, *cupiēbāris*, *cupiēbās*, *cupiēbat*, *cupiēbātis*, *cupiēbātur*, *cupiēminī*, *cupiēmur*, *cupiēmus*, *cupiendī*, *cupiendō*, *cupiendum*, *cupiendus*, *cupiēns*, *cupient*, *cupientur*, *cupiēre*, *cupiēris*, *cupiēs*, *cupiet*, *cupiētis*, *cupiētur*, *cupiminī*, *cupimur*, *cupimus*, *cupiō*, *cupior*, *cupis*, *cupit*, *cupīta*, *cupītā*, *cupītae*, *cupītam*, *cupītārum*, *cupītās*, *cupite*, *cupīte*, *cupītī*, *cupitis*, *cupītīs*, *cupitō*, *cupītō*, *cupitor*, *cupītōrum*, *cupītōs*, *cupitōte*, *cupītū*, *cupītum*, *cupitur*, *cupītūrus*, *cupītus*, *cupiunt*, *cupiuntō*, *cupiuntor*, *cupiuntur*, *cupīveram*, *cupīverāmus*, *cupīverant*, *cupīverās*, *cupīverat*, *cupīverātis*, *cupīvēre*, *cupīverim*, *cupīverimus*, *cupīverīmus*, *cupīverint*, *cupīveris*, *cupīverīs*, *cupīverit*, *cupīveritis*, *cupīverītis*, *cupīverō*, *cupīvērunt*, *cupīvī*, *cupīvimus*, *cupīvisse*, *cupīvissem*, *cupīvissēmus*, *cupīvissent*, *cupīvissēs*, *cupīvisset*, *cupīvissētis*, *cupīvistī*, *cupīvistis*, *cupīvit*},

{*eam*, *eamus*, *eant*, *eas*, *eat*, *eatis*, *eatur*, *eo*, *ēs*, *ēsa*, *ēsā*, *ēsae*, *ēsam*, *ēsī*, *ēsō*, *ēsū*, *eundi*, *eundo*, *eundum*, *eundus*, *eunt*, *eunto*, *i*, *ibam*, *ibamus*, *ibant*, *ibas*, *ibat*, *ibatis*, *ibatur*, *ibimus*, *ibis*, *ibit*, *ibitis*, *ibitur*, *ibo*, *ibunt*, *iens*, *ieram*, *ieramus*, *ierant*, *ieras*, *ierat*, *ieratis*, *iere*, *ierim*, *ierimus*, *ierint*, *ieris*, *ierit*, *ieritis*, *iero*, *ierunt*, *ii*, *iimus*, *iit*, *imus*, *ire*, *irem*, *iremus*, *irent*, *ires*, *iret*, *iretis*, *iretur*, *iri*, *is*, *isse*, *issem*, *issemus*, *issent*, *isses*, *isset*, *issetis*, *isti*, *istis*, *it*, *ite*, *itis*, *ito*, *itote*, *itu*, *itum*, *itur*, *iturus*, *itus*, *ivi*, *ivisti*, *ivit*},

{*edam*, *edāminī*, *edāmur*, *edāmus*, *edant*, *edantur*, *edar*, *edāre*, *edāris*, *edās*, *edat*, *edātis*, *edātur*, *ede*, *edēbam*, *edēbāminī*, *edēbāmur*, *edēbāmus*, *edēbant*, *edēbantur*, *edēbar*, *edēbāre*, *edēbāris*, *edēbās*, *edēbat*, *edēbātis*, *edēbātur*, *edēminī*, *edēmur*, *edēmus*, *edendī*, *edendō*, *edendum*, *edendus*, *edēns*, *edent*, *edentur*, *ēderam*, *ēderāmus*, *ēderant*, *ēderās*, *ēderat*, *ēderātis*, *edere*, *edēre*, *ēdēre*, *ederem*, *ederēminī*, *ederēmur*, *ederēmus*, *ederent*, *ederentur*, *ederer*, *ederēre*, *ederēris*, *ederēs*, *ederet*, *ederētis*, *ederētur*, *ēderim*, *ēderimus*, *ēderīmus*, *ēderint*, *ederis*, *edēris*, *ēderis*, *ēderīs*, *ēderit*, *ēderitis*, *ēderītis*, *ēderō*, *ēdērunt*, *edēs*, *edet*, *edētis*, *edētur*, *edī*, *ēdī*, *edim*, *ediminī*, *edimur*, *edimus*, *edīmus*, *ēdimus*, *edint*, *edis*, *edīs*, *ēdisse*, *ēdissem*, *ēdissēmus*, *ēdissent*, *ēdissēs*, *ēdisset*, *ēdissētis*, *ēdistī*, *ēdistis*, *edit*, *ēdit*, *edite*, *editis*, *edītis*, *editō*, *editor*, *editōte*, *editur*, *edō*, *edor*, *edunt*, *eduntō*, *eduntor*, *eduntur*, *ēsārum*, *ēsās*, *ēse*, *ēsīs*, *ēsōrum*, *ēsōs*, *esse*, *ēsse*, *ēssem*, *ēssēmus*, *ēssent*, *ēssēs*, *ēsset*, *ēssētis*, *ēst*, *ēste*, *ēstis*, *ēstō*, *ēstōte*, *ēstur*, *ēsum*, *ēsūrus*, *ēsus*},

{*fer*, *feram*, *ferāminī*, *ferāmur*, *ferāmus*, *ferant*, *ferantur*, *ferar*, *ferāre*, *ferāris*, *ferās*, *ferat*, *ferātis*, *ferātur*, *ferēbam*, *ferēbāminī*, *ferēbāmur*, *ferēbāmus*, *ferēbant*, *ferēbantur*, *ferēbar*, *ferēbāre*, *ferēbāris*, *ferēbās*, *ferēbat*, *ferēbātis*, *ferēbātur*, *ferēminī*, *ferēmur*, *ferēmus*, *ferendī*, *ferendō*, *ferendum*, *ferendus*, *ferēns*, *ferent*, *ferentur*, *ferēre*, *ferēris*, *ferēs*, *feret*, *ferētis*, *ferētur*, *feriminī*, *ferimur*, *ferimus*, *ferō*, *feror*, *ferre*, *ferrem*, *ferrēminī*, *ferrēmur*, *ferrēmus*, *ferrent*, *ferrentur*, *ferrer*, *ferrēre*, *ferrēris*, *ferrēs*, *ferret*, *ferrētis*, *ferrētur*, *ferrī*, *ferris*, *fers*, *fert*, *ferte*, *fertis*, *fertō*, *fertor*, *fertōte*, *fertur*, *ferunt*, *feruntō*, *feruntor*, *feruntur*, *lāta*, *lātā*, *lātae*, *lātam*, *lātārum*, *lātās*, *lāte*, *lātī*, *lātīs*, *lātō*, *lātōrum*, *lātōs*, *lātū*, *lātum*, *lātūrus*, *lātus*, *tuleram*, *tulerāmus*, *tulerant*, *tulerās*, *tulerat*, *tulerātis*, *tulēre*, *tulerim*, *tulerimus*, *tulerīmus*, *tulerint*, *tuleris*, *tulerīs*, *tulerit*, *tuleritis*, *tulerītis*, *tulerō*, *tulērunt*, *tulī*, *tulimus*, *tulisse*, *tulissem*, *tulissēmus*, *tulissent*, *tulissēs*, *tulisset*, *tulissētis*, *tulistī*, *tulistis*, *tulit*},

{*mālam*, *mālēbam*, *mālēbāmus*, *mālēbant*, *mālēbās*, *mālēbat*, *mālēbātis*, *mālēmus*, *mālent*, *mālēs*, *mālet*, *mālētis*, *mālim*, *mālīmus*, *mālint*, *mālīs*, *mālit*, *mālītis*, *mālle*, *māllem*, *māllēmus*, *māllent*, *māllēs*, *māllet*, *māllētis*, *mālō*, *mālueram*, *māluerāmus*, *māluerant*, *māluerās*, *māluerat*, *māluerātis*, *māluēre*, *māluerim*, *māluerimus*, *māluerīmus*, *māluerint*, *mālueris*, *māluerīs*, *māluerit*, *mālueritis*, *māluerītis*, *māluerō*, *māluērunt*, *māluī*, *māluimus*, *māluisse*, *māluissem*, *māluissēmus*, *māluissent*, *māluissēs*, *māluisset*, *māluissētis*, *māluistī*, *māluistis*, *māluit*, *mālumus*, *mālunt*, *māvīs*, *māvult*, *māvultis*},

{*morere*, *morerēminī*, *morerēmur*, *morerentur*, *morerer*, *morerēre*, *morerēris*, *morerētur*, *moreris*, *morī*, *moriāminī*, *moriāmur*, *moriantur*, *moriar*, *moriāre*, *moriāris*, *moriātur*, *moriēbāminī*, *moriēbāmur*, *moriēbantur*, *moriēbar*, *moriēbāre*, *moriēbāris*, *moriēbātur*, *moriēminī*, *moriēmur*, *moriendī*, *moriendō*, *moriendum*, *moriendus*, *moriēns*, *morientur*, *moriēre*, *moriēris*, *moriētur*, *moriminī*, *morimur*, *morior*, *moritor*, *moritur*, *moritūrus*, *moriuntor*, *moriuntur*, *mortua*, *mortuā*, *mortuae*, *mortuam*, *mortuārum*, *mortuās*, *mortue*, *mortuī*, *mortuīs*, *mortuō*, *mortuōrum*, *mortuōs*, *mortuū*, *mortuum*, *mortuus*},

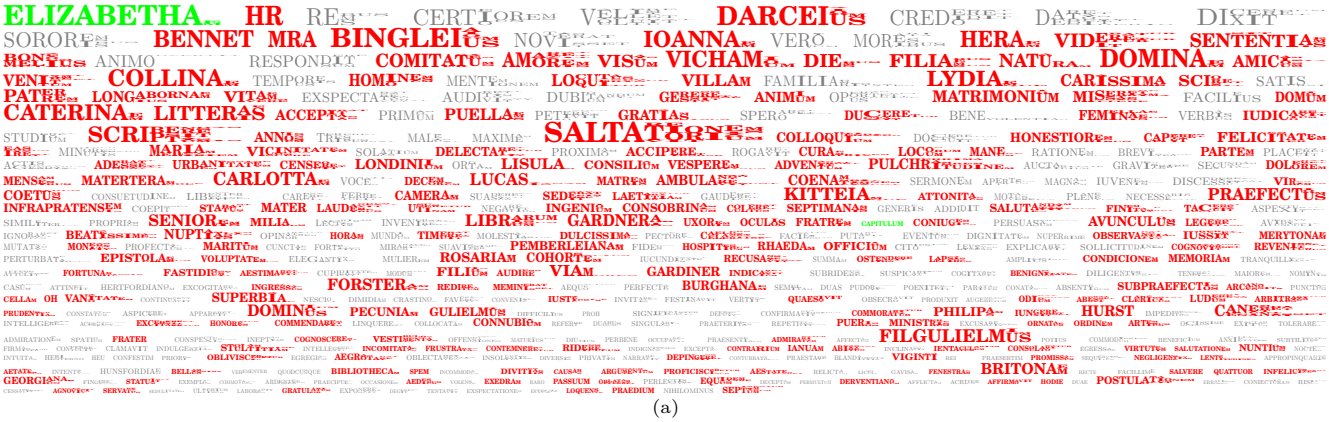
{*nōlam*, *nōlēbam*, *nōlēbāmus*, *nōlēbant*, *nōlēbās*, *nōlēbat*, *nōlēbātis*, *nōlēmus*, *nōlēns*, *nōlent*, *nōlēs*, *nōlet*, *nōlētis*, *nōlī*, *nōlim*, *nōlīmus*, *nōlint*, *nōlīs*, *nōlit*, *nōlīte*, *nōlītis*, *nōlitō*, *nōlitōte*, *nōlle*, *nōllem*, *nōllēmus*, *nōllent*, *nōllēs*, *nōllet*, *nōllētis*, *nōlō*, *nōlueram*, *nōluerāmus*, *nōluerant*, *nōluerās*, *nōluerat*, *nōluerātis*, *nōluēre*, *nōluerim*, *nōluerimus*, *nōluerīmus*, *nōluerint*, *nōlueris*, *nōluerīs*, *nōluerit*, *nōlueritis*, *nōluerītis*, *nōluerō*, *nōluērunt*, *nōluī*, *nōluimus*, *nōluisse*, *nōluissem*, *nōluissēmus*, *nōluissent*, *nōluissēs*, *nōluisset*, *nōluissētis*, *nōluistī*, *nōluistis*, *nōluit*, *nōlumus*, *nōlunt*, *nōluntō*},

{*passa*, *passā*, *passae*, *passam*, *passārum*, *passās*, *passe*, *passī*, *passīs*, *passō*, *passōrum*, *passōs*, *passū*, *passum*, *passūrus*, *passus*, *patere*, *paterēminī*, *paterēmur*, *paterentur*, *paterer*, *paterēre*, *paterēris*, *paterētur*, *pateris*, *patī*, *patiāminī*, *patiāmur*, *patiantur*, *patiar*, *patiāre*, *patiāris*, *patiātur*, *patiēbāminī*, *patiēbāmur*, *patiēbantur*, *patiēbar*, *patiēbāre*, *patiēbāris*, *patiēbātur*, *patiēminī*, *patiēmur*, *patiendī*, *patiendō*, *patiendum*, *patiendus*, *patiēns*, *patientur*, *patiēre*, *patiēris*, *patiētur*, *patiminī*, *patimur*, *patior*, *patitor*, *patitur*, *patiuntor*, *patiuntur*},

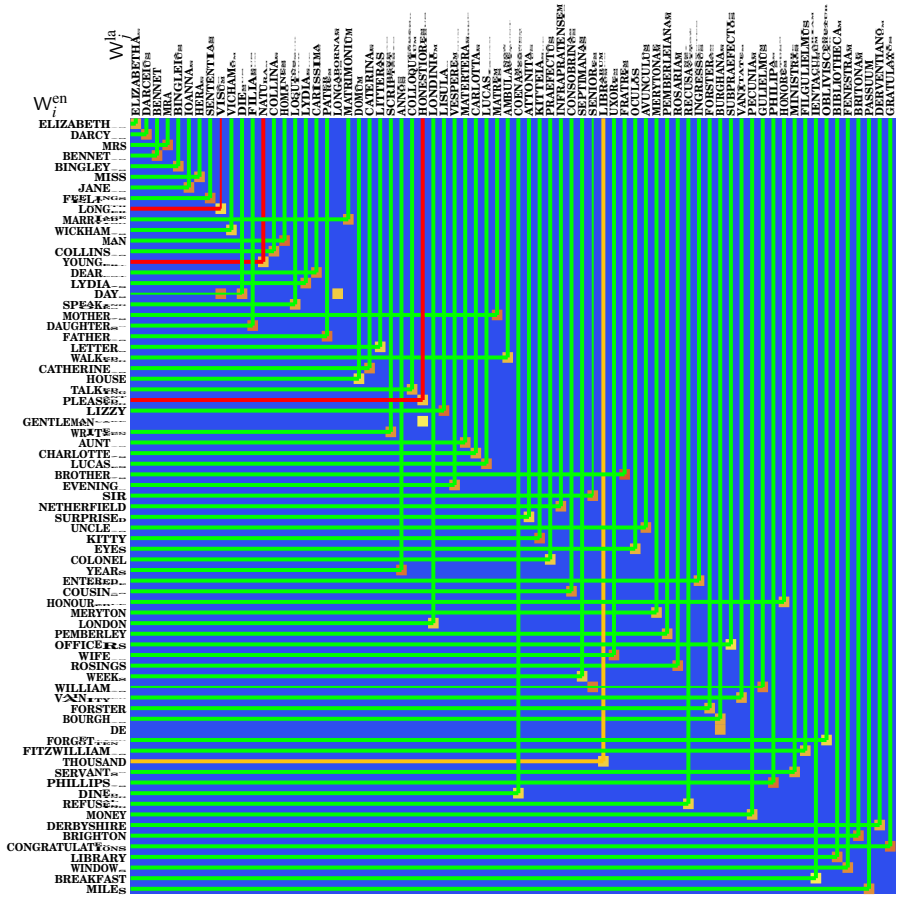
{*rape*, *rapere*, *raperem*, *raperēminī*, *raperēmur*, *raperēmus*, *raperent*, *raperentur*, *raperer*, *raperēre*, *raperēris*, *raperēs*, *raperet*, *raperētis*, *raperētur*, *raperis*, *rapī*, *rapiam*, *rapiāminī*, *rapiāmur*, *rapiāmus*, *rapiant*, *rapiantur*, *rapiar*, *rapiāre*, *rapiāris*, *rapiās*, *rapiat*, *rapiātis*, *rapiātur*, *rapiēbam*, *rapiēbāminī*, *rapiēbāmur*, *rapiēbāmus*, *rapiēbant*, *rapiēbantur*, *rapiēbar*, *rapiēbāre*, *rapiēbāris*, *rapiēbās*, *rapiēbat*, *rapiēbātis*, *rapiēbātur*, *rapiēminī*, *rapiēmur*, *rapiēmus*, *rapiendī*, *rapiendō*, *rapiendum*, *rapiendus*, *rapiēns*, *rapient*, *rapientur*, *rapiēre*, *rapiēris*, *rapiēs*, *rapiet*, *rapiētis*, *rapiētur*, *rapiminī*, *rapimur*, *rapimus*, *rapiō*, *rapior*, *rapis*, *rapit*, *rapite*, *rapitis*, *rapitō*, *rapitor*, *rapitōte*, *rapitur*, *rapiunt*, *rapiuntō*, *rapiuntor*, *rapiuntur*, *rapta*, *raptā*, *raptae*, *raptam*, *raptārum*, *raptās*, *rapte*, *raptī*, *raptīs*, *raptō*, *raptōrum*, *raptōs*, *raptū*, *raptum*, *raptūrus*, *raptus*, *rapueram*, *rapuerāmus*, *rapuerant*, *rapuerās*, *rapuerat*, *rapuerātis*, *rapuēre*, *rapuerim*, *rapuerimus*, *rapuerīmus*, *rapuerint*, *rapueris*, *rapuerīs*, *rapuerit*, *rapueritis*, *rapuerītis*, *rapuerō*, *rapuērunt*, *rapuī*, *rapuimus*, *rapuisse*, *rapuissem*, *rapuissēmus*, *rapuissent*, *rapuissēs*, *rapuisset*, *rapuissētis*, *rapuistī*, *rapuistis*, *rapuit*},

{*velim*, *velīmus*, *velint*, *velīs*, *velit*, *velītis*, *velle*, *vellem*, *vellēmus*, *vellent*, *vellēs*, *vellet*, *vellētis*, *vīs*, *volam*, *volēbam*, *volēbāmus*, *volēbant*, *volēbās*, *volēbat*, *volēbātis*, *volēmus*, *volēns*, *volent*, *volēs*, *volet*, *volētis*, *volō*, *volt*, *voltis*, *volueram*, *voluerāmus*, *voluerant*, *voluerās*, *voluerat*, *voluerātis*, *voluēre*, *voluerim*, *voluerimus*, *voluerīmus*, *voluerint*, *volueris*, *voluerīs*, *voluerit*, *volueritis*, *voluerītis*, *voluerō*, *voluērunt*, *voluī*, *voluimus*, *voluisse*, *voluissem*, *voluissēmus*, *voluissent*, *voluissēs*, *voluisset*, *voluissētis*, *voluistī*, *voluistis*, *voluit*, *volumus*, *volunt*, *vult*, *vultis*}.

Example 6.29.4. In Fig. S9, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source). Since vowel lengths are usually not marked in written Latin, conflations of certain concepts are unavoidable in our algorithm. For example, we have *mēnsis* "month" (nominative/genitive/vocative singular) vs. *mēnsīs* "tables" (dative/ablative plural), *miserīs* "poor" (masculine/feminine/neuter ablative plural) vs. *mīseris* "you will have sent" (second person singular, future perfect).



(a)



(b)

Fig. S9. Text mining in Latin. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Latin version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{la})$ between selected topics in English and Latin versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms (such as proper names, kinship terms, or emotions).

7 Approximate word clustering in selected Slavic languages

In this section, we present the approximate clustering algorithms for two representative Slavic languages: Polish and Russian.

Despite differences in the writing systems (Polish uses the Latin script while Russian employs the Cyrillic alphabet), both these languages share some morphological features in common:

- Polish nouns decline in seven cases, Russian six (in singular and plural forms).
- Adjectives are declined in three genders and in all the cases available to nouns.
- The same verb in imperfective and perfective aspects often differ by a prefix or an alternation of consonant/vowel. A limited amount of verbs have etymologically unrelated stems in their imperfective and perfective aspects.
- Both Polish and Russian have a handful of "verbs of motion" that contrast multidirectional and unidirectional forms.

These characterisations also apply to other Slavic languages (Croatian, Czech, Slovak and Slovene written in the Latin script; Belarusian, Bulgarian, Macedonian, Serbian and Ukrainian written in the Cyrillic script), more or less. In addition, the two modern Baltic languages, Latvian and Lithuanian, are similar to Slavic languages at the morphological level.

Some common words in Polish and Russian have very short (even vowelless) stems, and the short stems themselves may allow vowel and/or consonant alternations. At times, words with very different meanings have nearly identical stems. Thus it is highly challenging to design approximate clustering algorithms for these languages, while avoiding confusions of unrelated words. In the methods developed in this section, we make a compromise between a workable algorithm and error-free clustering.

7.1 Modified Porter stemming algorithm for Polish

Definition 7.1 (Polish stop words)**.** If a word belongs to the following list[87]:

a, *aby*, *ach*, *acz*, *aczkolwiek*, *aj*, *albo*, *albowiem*, *ale*, *ależ*, *ani*, *aż*, *bardziej*, *bardzo*, *bądź*, *bądźcie*, *bądźmy*, *bez*, *będą*, *będąc*, *będąca*, *będące*, *będący*, *będę*, *będzie*, *będziecie*, *będziemy*, *będziesz*, *bo*, *bowiem*, *by*, *bycie*, *być*, *byli*, *byliby*, *bylibyście*, *bylibyśmy*, *byliście*, *byliśmy*, *był*, *była*, *byłaby*, *byłabym*, *byłabyś*, *byłam*, *byłaś*, *byłby*, *byłbym*, *byłbyś*, *byłem*, *byłeś*, *było*, *byłoby*, *były*, *byłyby*, *byłybyście*, *byłybyśmy*, *byłyście*, *byłyśmy*, *bym*, *bynajmniej*, *byś*, *byście*, *byśmy*, *bywszy*, *cali*, *cała*, *cały*, *chociaż*, *choć*, *choćby*, *chyba*, *ci*, *ciebie*, *cię*, *co*, *cokolwiek*, *coraz*, *coś*, *cóż*, *czasami*, *czasem*, *czego*, *czegokolwiek*, *czegoś*, *czemu*, *czemukolwiek*, *czemuś*, *często*, *czy*, *czyjś*, *czyli*, *czym*, *czymkolwiek*, *czymś*, *czyż*, *czyżby*, *czyżbym*, *czyżbyś*, *czyżbyście*, *czyżbyśmy*, *daleko*, *dla*, *dlaczego*, *dlatego*, *do*, *dobrze*, *dokąd*, *dokądś*, *doń*, *dopóki*, *doprawdy*, *dosyć*, *dość*, *dużo*, *dwa*, *dwaj*, *dwie*, *dwoje*, *dzięki*, *dzisiaj*, *dziś*, *gdy*, *gdyby*, *gdybym*, *gdybyś*, *gdybyście*, *gdybyśmy*, *gdyż*, *gdzie*, *gdziekolwiek*, *gdzieś*, *go*, *i*, *ich*, *ile*, *ileś*, *iloma*, *ilu*, *im*, *inna*, *inną*, *inne*, *innego*, *innej*, *innemu*, *inni*, *inny*, *innych*, *innym*, *innymi*, *iż*, *ja*, *jacy*, *jacyś*, *jak*, *jaka*, *jakaś*, *jaką*, *jakąś*, *jakby*, *jaki*, *jakich*, *jakichś*, *jakie*, *jakiego*, *jakiegoś*, *jakiej*, *jakiejś*, *jakiemu*, *jakiemuś*, *jakieś*, *jakim*, *jakimi*, *jakimiś*, *jakimś*, *jakiś*, *jakiż*, *jakkolwiek*, *jako*, *jakoś*, *ją*, *je*, *jeden*, *jedna*, *jednak*, *jednakże*, *jedną*, *jedne*, *jednego*, *jednej*, *jednemu*, *jedni*, *jedno*, *jednych*, *jednym*, *jednymi*, *jedyna*, *jedyną*, *jedyne*, *jedynego*, *jedynej*, *jedynemu*, *jedyni*, *jedynie*, *jedyny*, *jedynych*, *jedynym*, *jedynymi*, *jego*, *jej*, *jemu*, *jest*, *jestem*, *jesteś*, *jesteście*, *jesteśmy*, *jeszcze*, *jeśli*, *jeżeli*, *już*, *każda*, *każdą*, *każde*, *każdego*, *każdej*, *każdemu*, *każdy*, *każdych*, *każdym*, *każdymi*, *kiedy*, *kiedyś*, *kilka*, *kim*, *kimkolwiek*, *kimś*, *kogo*, *kogokolwiek*, *kogoś*, *koło*, *komu*, *komukolwiek*, *komuś*, *kto*, *ktokolwiek*, *ktoś*, *która*, *którą*, *które*, *którego*, *której*, *któremu*, *który*, *których*, *którym*, *którymi*, *któryś*, *którzy*, *ku*, *lat*, *lecz*, *lub*, *ma*, *macie*, *mają*, *mając*, *mająca*, *mające*, *mający*, *mało*, *mam*, *mamy*, *masz*, *mą*, *me*, *mego*, *mej*, *memu*, *mi*, *miał*, *miał1*, *miała*, *miałaby*, *miałabym*, *miałabyś*, *miałam*, *miałaś*, *miałby*, *miałbym*, *miałbyś*, *miałem*, *miałeś*, *miało*, *miałoby*, *miały*, *miałyby*, *miałybyście*, *miałybyśmy*, *miałyście*, *miałyśmy*, *miano*, *mieć*, *miej*, *miejcie*, *miejmy*, *mieli*, *mieliby*, *mielibyście*, *mielibyśmy*, *mieliście*, *mieliśmy*, *mię*, *między*, *mimo*, *mną*, *mnie*, *mogą*, *mogąc*, *mogąca*, *mogące*, *mogący*, *mogę*, *mogli*, *mogliby*, *moglibyście*, *moglibyśmy*, *mogliście*, *mogliśmy*, *mogła*, *mogłaby*, *mogłabym*, *mogłabyś*, *mogłam*, *mogłaś*, *mogłem*, *mogłeś*, *mogło*, *mogłoby*, *mogłobym*, *mogłobyś*, *mogłom*, *mogłoś*, *mogły*, *mogłyby*, *mogłybyście*, *mogłybyśmy*, *mogłyście*, *mogłyśmy*, *moi*, *moich*, *moim*, *moimi*, *moja*, *moją*, *moje*, *mojego*, *mojej*, *mojemu*, *może*, *możecie*, *możemy*, *możesz*, *możliwe*, *można*, *móc*, *mógł*, *mógłby*, *mógłbym*, *mógłbyś*, *mój*, *mu*, *musi*, *musiał*, *musiała*, *musiałaby*, *musiałabym*, *musiałabyś*, *musiałam*, *musiałaś*, *musiałby*, *musiałbym*, *musiałbyś*, *musiałem*, *musiałeś*, *musiało*, *musiałoby*, *musiałobym*, *musiałobyś*, *musiałom*, *musiałoś*, *musiały*, *musiałyby*, *musiałybyście*, *musiałybyśmy*, *musiałyście*, *musiałyśmy*, *musicie*, *musieć*, *musieli*, *musieliby*, *musielibyście*, *musielibyśmy*, *musieliście*, *musieliśmy*, *musimy*, *musisz*, *muszą*, *musząc*, *muszącą*, *muszące*, *muszący*, *muszę*, *my*, *mych*, *mym*, *mymi*, *na*, *nad*, *najbardziej*, *nam*, *nami*, *nań*, *naokoło*, *naprzeciw*, *nas*,

[87] Our list of Polish stop words is based on `https://pl.wikipedia.org/wiki/Wikipedia:Stopwords`, with extensive additions to roughly match their counterparts in English. In particular, we have included all the inflected forms of *co* "what", *każdy* "every", *kto* "who", *nikt* "nobody", *nic* "nothing", *wszyscy*, *wszystek*, *wszystkie*, *wszystko* "all (in various genders)" and all the personal pronouns.

nasi, *nasz*, *nasza*, *naszą*, *nasze*, *naszego*, *naszej*, *naszemu*, *naszych*, *naszym*, *naszymi*, *natomiast*, *natychmiast*, *nawet*, *nią*, *nic*, *nich*, *niczego*, *niczemu*, *niczym*, *nie*, *niech*, *niechaj*, *nieco*, *niego*, *niej*, *niektóre*, *niektórych*, *niektórym*, *niektórymi*, *niektórzy*, *niemu*, *nigdy*, *nikim*, *nikogo*, *nikomu*, *nikt*, *nim*, *nimi*, *niż*, *no*, *o*, *oba*, *obaj*, *obie*, *obiema*, *oboje*, *obojga*, *obojgiem*, *obojgu*, *obok*, *oboma*, *obu*, *oby*, *obym*, *obyś*, *obyście*, *obyśmy*, *od*, *około*, *on*, *ona*, *one*, *oni*, *ono*, *oprócz*, *oraz*, *oto*, *owa*, *ową*, *owe*, *owego*, *owej*, *owemu*, *owi*, *owo*, *owszem*, *owych*, *owym*, *owymi*, *ów*, *pan*, *pana*, *pani*, *po*, *pod*, *podczas*, *pomiędzy*, *pomimo*, *ponad*, *ponieważ*, *poprzez*, *potem*, *powinien*, *powinienem*, *powinieneś*, *powinna*, *powinnam*, *powinnaś*, *powinni*, *powinniście*, *powinniśmy*, *powinno*, *powinny*, *powinnyście*, *powinnyśmy*, *poza*, *później*, *prawie*, *przecież*, *przeciw*, *przeciwko*, *przed*, *przede*, *przedtem*, *przez*, *przy*, *razem*, *roku*, *również*, *sam*, *sama*, *samą*, *same*, *samego*, *samej*, *samemu*, *sami*, *samo*, *samych*, *samym*, *samymi*, *są*, *siebie*, *się*, *skąd*, *skądś*, *skoro*, *sobą*, *sobie*, *spomiędzy*, *sponad*, *sposób*, *spośród*, *spoza*, *swa*, *swą*, *swe*, *swego*, *swej*, *swemu*, *swoi*, *swoich*, *swoim*, *swoimi*, *swoja*, *swoją*, *swoje*, *swojego*, *swojej*, *swojemu*, *swój*, *swych*, *swym*, *swymi*, *ta*, *tacy*, *tak*, *taka*, *taką*, *taki*, *takich*, *takie*, *takiego*, *takiej*, *takiemu*, *takim*, *takimi*, *także*, *tam*, *tamci*, *tamta*, *tamtą*, *tamte*, *tamtego*, *tamtej*, *tamtemu*, *tamten*, *tamto*, *tamtych*, *tamtym*, *tamtymi*, *tą*, *te*, *tego*, *tej*, *temu*, *ten*, *teraz*, *też*, *tę*, *to*, *tobą*, *tobie*, *toteż*, *trzeba*, *tu*, *tutaj*, *twa*, *twą*, *twe*, *twego*, *twej*, *twemu*, *twoi*, *twoich*, *twoim*, *twoimi*, *twoja*, *twoją*, *twoje*, *twojego*, *twojej*, *twojemu*, *twój*, *twych*, *twym*, *twymi*, *ty*, *tych*, *tyle*, *tylko*, *tyloma*, *tylu*, *tym*, *tymi*, *tyś*, *u*, *w*, *wam*, *wami*, *was*, *wasi*, *wasz*, *wasza*, *waszą*, *wasze*, *waszego*, *waszej*, *waszemu*, *waszych*, *waszym*, *waszymi*, *wbrew*, *we*, *według*, *wiele*, *wieloma*, *wielu*, *więc*, *więcej*, *wkrótce*, *właśnie*, *wobec*, *wszak*, *wszakże*, *wszyscy*, *wszystek*, *wszystka*, *wszystką*, *wszystkich*, *wszystkie*, *wszystkiego*, *wszystkiej*, *wszystkiemu*, *wszystkim*, *wszystkimi*, *wszystko*, *wśród*, *wtedy*, *wy*, *z*, *za*, *zapewne*, *zaraz*, *zaś*, *zawsze*, *zbyt*, *ze*, *zł*, *znowu*, *znów*, *został*, *zza*, *ż*, *żaden*, *żadna*, *żadne*, *żadnych*, *że*, *żeby*,

then we consider it a Polish stop word. All the Polish stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

In addition, we define string patterns **eatPolish**, **givePolish**, **MissPolish**, **MrsPolish** through inflected forms of four Polish words:

jada, *jadać*, *jadaj*, *jadają*, *jadając*, *jadająca*, *jadające*, *jadający*, *jadali*, *jadał*, *jadała*, *jadałe*, *jadało*, *jadały*, *jadana*, *jadane*, *jadani*, *jadanie*, *jadano*, *jadany*, *jadasz*, *jadł*, *jadła*, *jadłe*, *jadło*, *jadły*, *je*, *jedli*, *jedz*, *jedzą*, *jedząc*, *jedząca*, *jedzące*, *jedzący*, *jedzeni*, *jedzenie*, *jedzona*, *jedzone*, *jedzono*, *jedzony*, *jesz*, *jeść* — "eat";

da, *dacie*, *dać*, *dadzą*, *daj*, *dają*, *dając*, *dająca*, *dające*, *dający*, *dajcie*, *daje*, *dajecie*, *dajemy*, *dajesz*, *daję*, *dajmy*, *dali*, *daliby*, *dalibyście*, *dalibyśmy*, *daliście*, *daliśmy*, *dał*, *dała*, *dałaby*, *dałabym*, *dałabyś*, *dałam*, *dałaś*, *dałby*, *dałbym*, *dałbyś*, *dałem*, *dałeś*, *dało*, *dałoby*, *dały*, *dałyby*, *dałybyście*, *dałybyśmy*, *dałyście*, *dałyśmy*, *dam*, *damy*, *dana*, *dane*, *dani*, *danie*, *dano*, *dany*, *dasz*, *dawać*, *dawaj*, *dawajcie*, *dawajmy*, *dawali*, *dawaliby*, *dawalibyście*, *dawalibyśmy*, *dawaliście*, *dawaliśmy*, *dawał*, *dawała*, *dawałaby*, *dawałabym*, *dawałabyś*, *dawałam*, *dawałaś*, *dawałby*, *dawałbym*, *dawałbyś*, *dawałem*, *dawałeś*, *dawało*, *dawało1*, *dawałoby*, *dawały*, *dawałyby*, *dawałybyście*, *dawałybyśmy*, *dawałyście*, *dawałyśmy*, *dawana*, *dawane*, *dawani*, *dawanie*, *dawano*, *dawany*, *dawszy* — "give";

panna, *pannach*, *pannami*, *panną*, *pannę*, *pannie*, *panno*, *pannom*, *panny* — "Miss";

paniach, *paniami*, *panią*, *panie*, *paniom*, *pań* — "Mrs".

In order to stay consistent with their counterparts in English, we are not going to treat these string patterns as stop words in Polish. Instead, we will exploit them in clustering of Polish content words.

7.1.1 Effective spelling and essential root

It is assumed that all Polish words are converted to lowercase[88] before going through any of the procedures below.

Definition 7.2 (Polish Vowels and Verb Prefixes)**.** Hereafter in §7.1, the symbol **V** stands for any member from the list of Polish vowels {*a*, *ą*, *e*, *ę*, *i*, *o*, *ó*, *u*, *y*}. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Polish vowels.

Dual to the notations above, the symbol **C** stands for any character that does not belong to the list {*a*, *ą*, *e*, *ę*, *i*, *o*, *ó*, *u*, *y*}, and $\mathbf{C}_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list.

The symbol **P** = (∅|*do*|*na*|*nad*|*o*|*ob*|*obe*|*po*|*pod*|*pode*|*prze*|*przed*|*przy*|*roz*|*s*|*sj*|*u*|*w*|*we*|*wy*|*z*|*za*|*ze*) represents the possible prefix of a Polish verb. □

[88] As of v11.0, *Mathematica* does not produce conversions *Ą*→*ą*, *Ę* → *ę*, *Ń* → *ń*, *Ś* → *ś*, *Ź* → *ź*, *Ż* → *ż* upon invoking the `ToLowerCase` operation. Therefore, case conversions for these special Polish letters need to be hand-coded if one implements our algorithm in *Mathematica*.

Algorithm 7.3 (Polish effective spelling). *For a Polish word $\hat{\sigma}$, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in sequential steps:*

(1) *Replace*

P*rze*(*cz*|*k*)	(∅|*co*)*dzienn*~	(∅|*naj*)*czarn*~	(∅|*naj*)*koniecz*~
P*ρṽek*	*dni*	*βλακ*	*ṽσecz*

(∅|*naj*)*wyżs*(∅|*z*)~	(∅|*naj*)*zim*~	(∅|*o*)*żeni*~	(∅|*o*|*po*)(*ślub*|*żen*)*i*~	(∅|*prze*)*czyt*~
ηwykη	*ωiṽτ*	*małżeństwo*	*małżeństwi*	*ρczytδ*

(∅|*wy*)*marz*~	(*mężat*|*zamężn*|*żonat*)~	(*ojcz*|*tat*)*ul*(∅|*e*)*k*~	(*s*|*ś*)*piesz*	(*u*|*z*)*mar*(*l*|*ł*)**X**~
δρiμ	*małżeństw*	*ojc*	*χυρ*	*μορτ*

bogat~	*bram*~	*dłu*(*g*|*ż*)	*dziewię*~	*dzwo*~	*idea*(*l*|*ł*)~	*ident*~	*iii*	*imi*~	*karol*~	*katarz*~
ρbκit	*βραμ*	*λoγ*	*9e*	*δζwo*	*ιδεαλ*	*ιδeṽτ*	*3*	*ṽαμ*	*καρl*	*κατarz*

kieli(*ch*|*sz*)~	*ksi*(*ążecz*|*ędz*|*ęg*)	*marc*~	*martw*	*milicj*~	*od***C**~	*opł*~	*pami*(*ą*|*ę*)~
καλχ	*książ*	*μ3αρ*	*μορτ*	*μιλτ*	*q***C**	*opλ*	*μρετμ*

panowie	*plan*	*płac*	*pocz*	*poje*~	*pok*(*o*|*ó*)*i*	*pok*(*o*|*ó*)*j*	*poka*(*z*|*ż*)~	*popul*~	*pta***X**~	*robi*~
panach	*πλαṽ*	*płatn*	*ποξ*	*je*	*ρυμoi*	*ρυμoj*	*σχοω*	*ποπul*	*βiρδ*	*zrobi*

rozpacz~	*rób*~	*skr*~	*sp*(*o*|*ó*)**C**~	*stan*(*o*|*ó*)*w*	*stos*	*str*~	*sześ*(*c*|*ć*)~	*świad*~
δεσπρ	*zrob*	*σkρ*	*s***C***p*	*κοṽτw*	*stoσ*	*στρ*	*6*	*ωiτṽσ*

świat~	*tajemnicz*~	*trz*(*y*|*e*(*ch*|*j*|*m*|*ma*))~	*tysi*(*ą*|*ę*)*c*~	*uczu*	*ujemn*~
ωświλδ	*μυστicz*	*3e*	*m1il*	*φιλu*	*ωjemn*

um(∅|*a*|*ie*)*r*(∅|*z*)**X**~	*upły*~	*v*	*wiejsk*~	*wizj*~	*woln*	*wy*(*sł*|*sył*|*ściel*|*śl*)~	*wymien*~	*wynajd*~
μορτ	*ωply*	*5*	*ρυρλ*	*ωiσj*	*φρin*	*σeṽδ*	*ξmien*	*βynajd*

wynal(*a*|*e*)(*z*|*ź*)~	*wyso*(*c*|*k*)~	*wzmian*~	*x*	*zaw*(∅|*a*|*ie*)*r*(∅|*z*)**X**~	*zię*(*c*|*ć*)~	*zi*~	*zna*(*cz*|*k*)~
βynajdo	*ηwykη*	*μmeṽτn*	*10*	*iṽκλ*	*ιṽλαω*	*ζi*	*ζṽak*

znal(*a*|*e*)(*z*|*ź*)~	*zys*(*k*|*zcz*)~	*ży*(*j*|*l*|*ł*|*t*)~	**MissPolish**
znajdo	*γaṽ*	*żyć*	*miß*

MrsPolish	(∅|*z*)**eatPolish**(∅|*by*)(∅|*ś*)(∅|*cie*|*m*|*my*)	(*iść*|*szli*)
μiρσ	*εατ*	*idź*

czu(*j*|*l*|*ł*|*t*|~*ć*)	*de*	*l*	*lady*	*rzec*	*ta*(*cie*|*tach*)	*tat*(**V**|*m*|*w*)$_m$	~**C***że*
φιλuć	*δεe*	*50*	*damskich*	*rzekl*	*ojcu*	*ojcu*	**C**

(2) *Replace*

P*miesza*~	**P***trz*	(*do*|*na*|*o*|*od*|*po*|*roz*|*wy*|*za*)(*łoż*|*łóż*)~
mixa	**P***τρ*	*kład*

$^{\mathbf{X}\in}$((*poj*|*przyj*)(*ą*|*ę*|*m*))~	*brać*~	*chud*(*l*|*ł*|*n*)~	*cie*(*ni*|*ń*)~	*ciesz*~	*czc*	*dba*~	*domysł*~	*domyś*~	*doskon*	*drzew*~
*σ***X**	*wziąć*	*ωoτ*	*oμβ*	*ξeσ*	*czt*	*δyβa*	*mysł*	*myś*	*περφ*	*δrzew*

licz(∅|*b*)~	*list*~	*łam*$\hat{\chi}$	*łaśc*	*maj*(*ą*|*ę*)*t*~	*mal*~	*miejsc*~	*miesiąc*~	*mieszk*~	*miękko*	*milcz*~	*ocen*~
ṽυμa	*λiστ*	*μλaμ*$\hat{\chi}$	*σλaśc*	*φoρτ*	*μaλ*	*μiejστ*	*μoṽsc*	*mieszκ*	*miękcej*	*qit*	*oζeṽ*

opin~	*op*~	*os*(*o*|*ó*)*b*~	*pan*(*ow*|*uj*)~	*pi*(*s*|*sz*|*ś*)**X**	*piecz*(*ą*|*ę*)~	*piosenk*	*praw*
ωpin	*ωp*	*oσoβ*	*κoτρλ*	*πiσ*	*σiλa*	*śpiewać*	*πraβ*

przedmio(*c*|*t*)**X**~	*przekon*~	*qm*(*a*|*o*|*ó*)*w*	*rze*(*cz*|*ką*|*kę*)~	*serc*~	*słab*	*straż*~	*szczęś*	*szedł*	*ślicz*~	*ślub*
θiγ	*βλiφ*	*δeṽη*	*rzekl*	*serdecz*	*sβλaβ*	*φiρ*	*ηaπ*	*chodz*	*βuτ*	*σlub*

uszy$^{\mathbf{X}\in}$(*c*|*ć*|*j*|*k*|*l*|*ł*|*t*)~	*will*~	*zach*~	*zna*~	*zobacz*~	*życz*~	(*iście*|*iść*|*szli*|*szłobym*|*szłom*)
*szy***X**	*βiλλ*	*Ξ*	*ζna*	*widzi*	*ωiσ*	*szedłem*

przyjaci(*e*|*o*|*ó*)(*l*|*ł*)(∅|*c*|*k*)**X**	*ś*(*ś*|*w*)	~$\sharp\hat{\chi}_{m_0}$	~$^{\mathbf{X}\in}$(*k*|*l*)*ach*	~*kich*	~*oce*
πprzyjak	*święta*	∅	**X***a*	*ki*	*oc*

(3) *Do dom*~ → *δoμ*, *pie*(*c*\|*cz*\|*k*) → *βακ*, **givePolish**(∅\|*by*)(∅\|*ś*)(∅\|*cie*\|*m*\|*my*) → *dar*, ~*kacie* → *ka*.

(4) *Replace*

(**P**|∅|*naj*)*ł*~	(∅|*naj*)*gors*(*i*|*z*)~	(∅|*naj*)*gorzej*~	(∅|*naj*)*lepiej*~	(∅|*naj*)*leps*(*i*|*z*)~
λ	*źl*	*źle*	*δobrze*	*δobr*

(∅|*naj*)*mniejs*(*i*|*z*)~		(∅|*naj*)*więks*(*i*|*z*)~			(∅|*za*)*ta*(*niec*|*ńc*)(∅|*z*)~				(*mów*|*powie*)~		(*my*|*umy*)~	
mał		*duż*			*τańc*				*πowie*		*uμy*	
(*wiad*|*wiedz*)~	$^{\mathbf{X}\in}$(*s*|*z*)*ł*	*bra*(*c*|*ć*|*t*)~	*chc*~	*cioc*~	*cioć*~	*ciot*(*c*|*ek*|*k*)~	*dam*~	*darc*~	*dobr*~	*dumien*	*dzieck*~	
wie	**X***jl*	*βbra*	*choc*	*tsioc*	*tsioci*	*tsioci*	*dδaμm*	*tdarc*	*δobr*	*δumien*	*dziec*	

ksi(*ą*|*ę*)*dz*~		*mam*~		*mat*(*c*|*ek*|*k*)~		*myl*~	*naj*($\hat{\chi}$)$_m$$^{\mathbf{X}_1\in}$(*ej*|*sz*)**X**			*nara*$^{\mathbf{X}\in}$(*z*|*ż*)~		*od*~	
księż		*μam*		*μam*		*ṽμy*	($\hat{\chi}$)$_m$**X**$_1$			*ṽara***X**		*odz*	

orzeł(∅|*ek*|*k*)~	*pi*(*ąt*|*ęc*|*ęć*)~	*pies*~	*piln*~	*śn*	*tydzień*~	*wielk*~	*wuj*(*ek*|*k*)~	*zλ*~	*zdum*~	*zmi*~	*znal*~	*dzień*
orł	*5a*	*psa*	*πiln*	*sn*	*tygodni*	*duż*	*wuj*	*źl*	*zδum*	*ζmi*	*znaj*	*dni*

(ojciec|*ojcze)*	*kwiecień*	*oj*	*pić*	*tatu*(*s*|~*ś*)	*uch*((∅|**V**)(∅|*ch*|*m*|*mi*))	*wielce*
ojcach	*kwietni*	*yoj*	*pici*	*ojc*	*uszach*	*duż*

(5) *Do ch*→*sχ, cz*→*č, ł*→*l, ó*→*o, sz*→*š,* ~*deł*→∅, ~*i*→*j.*

(6) *Do wi*(*ad*\|*edz*)→*wie,* **C***i*→**C***ji,* $^{\mathbf{X}\in}$(**P***mow*)~→**X***maw, and call the result* $\hat{\sigma}'$.

(7) *Break down* $\hat{\sigma}' = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}'^{[\min\{\lambda(\hat{\sigma}'),\ell(\hat{\sigma}')\}]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where* $\lambda(\hat{\sigma}') - 1$ *equals either the last position occupied by the string pattern* (**P**\|*bez*(∅\|*e*)\|*njie*\|*pa*\|*st*(∅\|*r*))$\hat{\chi}$~ *in the non-void* $\hat{\sigma}'$*, or* −1 *if* $\hat{\sigma}' = \varnothing$.

(8) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(8.1) *Replace*

$^{\mathbf{X}\in}$(*g*|*k*|*l*)*j*	*ck*	*dzk*	*ńc*	*ow*(∅|*č*]*ji*(*č*|*ec*)|*nji*(∅|*c*|*k*)|*sk*)	(*a*(*c*|*st*|*t*)|*ljiw*|(∅|*aw*|*nji*)*č*(∅|*yk*|~*ycy*)|*njisk*|(*a*(*l*|*r*)*n*(∅|*ji*))	
X	*c*	*d*	*ń*	∅		

(*ač*(∅|*ce*|*k*)|*a*(∅|*ńs*)*k*|*ar*(*ce*|*k*|*z*)|*by*|*dl*|*iciel*|*i*(*st*|*ści*|*w*)|*kacj*|*s*(*k*|*tw*)|*y*(*ciel*|*st*|*ści*|*w*))						
∅						
~*a*(*ček*|*cy*|*rek*)				~$^{\mathbf{X}\in}$(*nt*|*st*)(*ce*|*ek*|*k*)$\hat{\chi}_{m_0}$	~*ij*	~*mję*
∅				**X**	*i*	*mjienj*

(8.2) *Do* *k***V**→**V**, ~(*ce*\|*k*)→∅, ~((*a*\|*i*\|*y*)*sχ*\|*e*\|*ego*\|*em*\|*emu*\|*im*\|*mj*\|*ow*(∅\|*j*)**V**$_{m_0}$\|*om*\|*um*\|*ym*\|**V***ć*)→∅.

The result after these two steps of operations is called $\hat{\sigma}'_2$.

(9) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(10) *Do ć*→*cj, ń*→*nj, ś*→*sj, ź*→*zj.*

Definition 7.4 (Polish protected range)**.** Let $\hat{\sigma}$ be a text string derived from a Polish word, its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \min\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is determined by two non-negative integers $\lambda_1(\hat{\sigma})$ and $\lambda_2(\hat{\sigma})$ specified through the following procedures:

- Look for the string pattern (*na*\|*u*\|*we*\|*wy*)$_{m_0}$**C**$_{m_0}$**V**(*j*\|*š*)$_{m_0}$~ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Look for the pattern (*na*\|*po*\|*prze*\|*przy*\|*roz*\|*s*\|*w*\|*wy*\|*z*)**V**$_{m_0}$**CV**$_{m_0}$**C**~ in the string $\hat{\sigma}$;
- If the pattern above is found, the last position occupied by such a pattern defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = \ell(\hat{\sigma})$. □

Algorithm 7.5 (Polish essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Polish word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$.

(2) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(2.1) *Do* ~**V**$_m$→∅.

(2.2) *Do* $\sim\mathbf{V}_m(m|n(\varnothing|j)|w\check{s}) \to \mathbf{V}_m$, $\sim\mathbf{V}_{m_0}\check{s} \to \mathbf{V}_{m_0}$, $\sim(ąc|ąt|ęcj|ęt) \to \varnothing$.

(2.3) *Do* $\sim\mathbf{V}_m \to \varnothing$.

The result after these three steps of operations is called $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

In Polish (as well as Russian and many other Slavic languages), the imperfect and perfect aspects of the same verb may differ by a prefix. To improve the performance of verb clustering, we need the following extensions to Definition 7.4 and Algorithm 7.5.

Definition 7.6 (Polish verbal protected range)**.** Let $\hat{\sigma}$ be a text string derived from a Polish word, its verbal protected range $\text{VbProtRg}(\hat{\sigma})$ is an integer determined as follows:

- Try to find the string pattern $(do|na|o|odz|po|prze|we|wy)_{m_0}\mathbf{C}_{m_0}\mathbf{V}(j|l)_{m_0}\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\text{VbProtRg}(\hat{\sigma})$; otherwise, set $\text{VbProtRg}(\hat{\sigma}) = 0$. □

Algorithm 7.7 (Polish verbal essential root)**.** *Let* $\hat{\sigma}$ *be the token string associated with a Polish word, then its corresponding verbal essential root* $\text{VbEssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\text{VbProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \text{VbProtRg}(\hat{\sigma})$ *is equal to the verbal protected range of* $\hat{\sigma}$.

(2) *Do* $\sim(\mathbf{V}_{m_0}(l|(\mathbf{V}_{m_0}(\varnothing|c|s|m)(\varnothing|j)(\varnothing|m)))_{m_0})_m \to \varnothing$ *on* $\hat{\sigma}_2$*, and call the result* $\hat{\sigma}'_2$.

(3) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(4) *Do* $\sim^{\hat{\chi}\in}(d|m|n|r|s|w|z)j \to \hat{\chi}$, $\sim c(\varnothing|j) \to t$, $\sim\dot{z} \to z$.

(5) *Replace*

$\underline{(bjierz\|bjior\|br\|bra\|bran\|wez\|wz\|wzji)}$	$\sim dn$	$\underline{l(a\|al\|ej)}$	$\underline{pji(\varnothing\|j\|t)}$	$\sim\dot{z}dz$
bral	*d*	*lan*	*pjil*	*zjdz*

(6) *Replace*

$\underline{(\varnothing\|do\|na\|o\|odz\|po\|roz\|wy\|za)kla(\varnothing\|d\|dd\|dz)}$			
klad			
$\underline{(\varnothing\|w)(i\|id\|idz\|s\chi odz\|šjla\|šjlo\|šjly\|šlji)}$	$\underline{(\varnothing\|w)(jad\|jedz\|jes\chi)}$	$\underline{\mathbf{X}_1\in(do\|ob\|odz\|po\|przy\|wy)(jad\|jedz\|jes\chi\|jezjdz)}$	
$s\chi odz$	$jes\chi$	$\mathbf{X}'_1 jes\chi$	
$\underline{\mathbf{X}_2\in(do\|odz\|odze\|po\|przy\|we\|wy)(j\|jd\|jdz\|š\|s\chi odz\|šed)}$		$\underline{obejrz}$	$\underline{odze(d\|jl\|l\|s\chi odz)}$
$\mathbf{X}''_2 s\chi odz$		*ogląd*	$odzs\chi odz$

where $\mathbf{X}'_1$ *results from doing* $przy \to \pi rzy$ *on* $\mathbf{X}_1$ *and* $\mathbf{X}''_2$ *results from doing* $odze \to odz$, $przy \to \pi rzy$ *on* $\mathbf{X}_2$.

7.1.2 Admissible mutation and approximate clustering

In Polish (as well as Russian and many other Slavic languages), fleeting vowels may appear or disappear in inflected forms of the same word. We need to heuristically detect and remove these fleeting vowels to achieve better clustering results. Meanwhile, in Polish (as well as Russian and many other Slavic languages), one encounters vowel alternations in verb conjugations, as in Spanish or the Germanic languages.

Algorithm 7.8 (Polish vowel blotting)**.** *For a token string* $\hat{\sigma}$*, its blotted form* $\text{BlotFltV}(\hat{\sigma})$ *is constructed as follows:*

(1) *Do* $\sim\mathbf{C}(e|jie|jio)^{\hat{\chi}\in}(c|k|l|n|r|sj|w) \to \mathbf{C}\hat{\chi}$.

(2) *Do* $\sim\hat{\chi}kier \to \hat{\chi}kr$.

(3) *Do* $ą \to ę$, $i(a|o) \to ie$.

In what follows, we will construct a bivariate Boolean-valued function $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ on a "simple heredity test function" in Algorithm 7.9, and a set of "admissible suffix mismatch" rules in Algorithm 7.10.

Algorithm 7.9 (Simple heredity test)**.** *Let $\hat{\alpha}'$ be a string obtained from $\hat{\alpha}$ by the following steps:*

(1) *Break down $\hat{\alpha} = \hat{\alpha}_1\hat{\alpha}_2$ into the concatenation of two strings $\hat{\alpha}_1 = \hat{\alpha}^{[\min\{3,\ell(\hat{\alpha})\}]}$ (see the notation in Definition 3.1) and $\hat{\alpha}_2$, where the length of the first string $\ell(\hat{\alpha}_1) = \min\{3, \ell(\hat{\alpha})\}$ is equal to 3 or the length of $\hat{\alpha}$, whichever is shorter.*

(2) *On $\hat{\alpha}_2$, perform the following substitutions in a sequel:*

(2.1) *Do $\sim\mathbf{C}_m \to \mathbf{X}'$, where $\mathbf{X}'$ is obtained from $\mathbf{C}_m$ by deleting all the occurrences of the letter "j".*

(2.2) *Do $\sim(n|(\mathbf{V}_{m_0}(l|n|\check{s}|t|w)_{m_0})_m) \to \varnothing$.*

The result after these two steps of operations is called $\hat{\alpha}'_2$.

(3) *Construct $\hat{\alpha}' = \mathrm{BlotFltV}(\hat{\alpha}_1\hat{\alpha}'_2)$.*

Define $\hat{\beta}'$ similarly. The Boolean-valued function $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if the lowercase form of $\hat{\alpha}$ contains at least one instance of {ą, a, e, ę, i, l, n, o, ó, r, u, y}* **AND** *at least one of the following six conditions holds:*[89]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\beta} = \hat{\alpha}in$;

(iii) $\hat{\beta} = \hat{\alpha}n$;

(iv) $\hat{\beta} = \hat{\alpha}z$;

(v) $\min\{\ell(\hat{\alpha}'), \ell(\hat{\beta}')\} \geq 3$ **AND** ($\hat{\alpha} = \hat{\beta}'$ **OR** $\hat{\alpha}' = \hat{\beta}$ **OR** $\hat{\alpha}' = \hat{\beta}'$ **OR** $\hat{\beta}' = \hat{\alpha}'z$).

(vi) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{3}$ **AND** $\hat{\beta} = \hat{\alpha}(a|e|o)_m(m|t)$.

In what follows, we define $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}), \mathrm{RootNW}(\hat{\alpha},\hat{\beta}), \mathrm{NW}^*(\hat{\alpha},\hat{\beta})$ and $\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta})$ as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5. Particular to the Polish case, we further define $\mathrm{SuffixNW}'(\hat{\alpha},\hat{\beta})$ by performing the substitutions $(e|jie)\sim \to \varnothing, \sim j \to \varnothing, \sim jc \to c$ to both components of $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})$, in a single sweep. The notation $\mathrm{SuffixSW}'(\hat{\alpha},\hat{\beta})$ is defined similarly. If $\mathrm{SuffixNW}'(\hat{\alpha},\hat{\beta}) = [\hat{\tau}_1, \hat{\tau}_2]$, then $\mathrm{SuffixNW}'(\hat{\beta},\hat{\alpha}) = [\hat{\tau}_2, \hat{\tau}_1]$.

Algorithm 7.10 (Admissible suffix mismatch and vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}), \mathrm{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if* $\mathrm{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing|[a, e]|[a, o]|[ą, ę]|[e, o]$ **AND** $\mathrm{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of {ą, a, e, ę, i, l, n, o, ó, r, u, y}* **AND** *at least one of the following three conditions holds:*

(i) $\mathrm{SuffixNW}'(\hat{\alpha},\hat{\beta}) = [\varnothing, \varnothing]|[\varnothing, b]|[\varnothing, e]|[\varnothing, ie]|[c, k]|[c, t]|[č, k]|[dz, g]|[jc,t]|[s, š]|[sχ, s]|[sχ, š]|[z, ż]|[zd, żdż]|[zg, żdż]$;

(ii) $\mathrm{SuffixNW}'(\hat{\beta},\hat{\alpha}) = [\varnothing, \varnothing]|[\varnothing, b]|[\varnothing, e]|[\varnothing, ie]|[c, k]|[c, t]|[č, k]|[dz, g]|[jc,t]|[s, š]|[sχ, s]|[sχ, š]|[z, ż]|[zd, żdż]|[zg, żdż]$;

(iii) $\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}) = [\varnothing|(\varnothing|j|l)\mathbf{V}_{m_0}(\varnothing|m|sj)(\varnothing|m|cj), \varnothing|(\varnothing|j|l)\mathbf{V}_{m_0}(\varnothing|m|sj)(\varnothing|m|cj)]$.

Similarly, one can evaluate another Boolean-valued function

$$\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

Algorithm 7.11 (Heredity test function)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, define $\hat{\alpha}'$ and $\hat{\beta}'$ as in Algorithm 7.9. The Boolean-valued function* $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta}) =$ **TRUE**;

(ii) $\ell(\mathrm{RootNW}(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}'),\ell(\hat{\beta}')\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}), \mathrm{NW}^*(\hat{\alpha},\hat{\beta})) =$ **TRUE**;

(iii) $\ell(\mathrm{RootSW}(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}'),\ell(\hat{\beta}')\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta})) =$ **TRUE**.

[89] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 7.12 (Polish verb aspect test)**.** *Let* NW($\hat{\alpha}$,β) *be the result of performing Needleman–Wunsch alignment on strings* $\hat{\alpha}$ *and* $\hat{\beta}$*. Let* $\mathbf{X}^*$ *be an arbitrary non-empty string. The Boolean-valued function* VbAspTest($\hat{\alpha}$,$\hat{\beta}$) *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\mathrm{NW}(\hat{\alpha},\hat{\beta}) = \mathbf{X}^*$ (*in other words,* $\hat{\alpha} = \hat{\beta}$*);*

(ii) $\mathrm{NW}(\hat{\alpha},\hat{\beta}) = [\mathbf{P},\mathbf{P}]\mathbf{X}^*$ (*in other words,* $\hat{\alpha}$ *agrees with* $\hat{\beta}$ *up to a pair of word initial mismatching strings, both of which are possible prefixes of Polish verbs*)*;*

(iii) $\mathrm{NW}(\hat{\alpha},\hat{\beta}) = [\mathbf{P},\mathbf{P}]\mathbf{X}^*([a,\, e]|[a,\, o]|[ą,\, ę]|[e,\, o])\mathbf{X}^*$*, where the two occurrences of* $\mathbf{X}^*$ *may or may not represent the same non-empty string.*

Note that Smith–Waterman alignment is not used in this test.

The first two stages in the following approximate clustering algorithm for Polish words are very similar to the English counterpart (Algorithm 4.15), while the third stage is tailored for the verb aspects in Polish.

Algorithm 7.13 (Approximate clustering of Polish words)**.** *The approximate clustering of a list of Polish words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in three stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1, \mathrm{EffSpell}(\hat{\alpha}_1)), \ldots, (\hat{\alpha}_N, \mathrm{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the second component (effective spelling) of each entry. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1)})), \ldots, (\hat{\alpha}_{(N)}, \mathrm{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfies* $\mathrm{HrdTest}(\mathrm{EffSpell}(\hat{\alpha}_{(n)}), \mathrm{EffSpell}(\hat{\alpha}_{(n+1)})) = \mathbf{FALSE}$*, where* $n \in \mathbb{Z} \cap [1, N)$*, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,n_1)}))\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(M,1)})), \ldots, (\hat{\alpha}_{(M,n_M)}, \mathrm{EffSpell}(\hat{\alpha}_{(M,n_M)}))\}\}$*, where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$*.*

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)})), \ldots, (\hat{\alpha}_{(m,n_m)}, \mathrm{EffSpell}(\hat{\alpha}_{(m,n_m)}))\}$ *where* $m \in \mathbb{Z} \cap [1, M]$*, we augment it into a tagged entry* $G_m = (g_m, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)}), \mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})), \mathrm{BlotFltV}(\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\mathrm{BlotFltV}(\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})))$ *(with highest priority),* $\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with medium priority), and* $\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}, \hat{\gamma}'''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)}, \hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\mathrm{SimpHrdTest}(\hat{\gamma}'''_{(m+1)}, \hat{\gamma}'''_{(m)}) = \mathbf{FALSE}$$

AND

$$\mathrm{SimpHrdTest}(\hat{\gamma}'''_{(m)}, \hat{\gamma}'''_{(m+1)}) = \mathbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}''_{(m+1)}, \hat{\gamma}''_{(m)}) = \mathbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \mathbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \mathbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$*, then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Generate a list of word clusters* $\{\check{\Gamma}_1 = (\check{G}_{(1,1)}, \ldots), \ldots, \check{\Gamma}_K = (\check{G}_{(K,1)}, \ldots)\}$ *by discarding all the tags (effective spellings, essential roots, vowel blotted forms) from each* Γ_k*, where* $k \in \mathbb{Z} \cap [1, K]$*.*

(3) *For each word cluster* $\check{\Gamma}_k = (\check{G}_{(k,1)}, \ldots)$*, we augment it into a tagged entry*

$$\mathscr{G}_k \equiv (\check{\Gamma}_k, \check{\Gamma}'_k, \check{\Gamma}''_k) := (\check{\Gamma}_k, \mathrm{EssRoot}(\mathrm{EffSpell}(\check{G}_{(k,1)})), \mathrm{VbEssRoot}(\mathrm{EssRoot}(\mathrm{EffSpell}(\check{G}_{(k,1)})))).$$

Recall that $\hat{\sigma}^{-1}$ *is the reverse of the string* $\hat{\sigma}$ *(Definition 3.1). The list* $\{\mathscr{G}_1 = (\check{\Gamma}_1, \check{\Gamma}'_1, \check{\Gamma}''_1), \ldots, \mathscr{G}_K = (\check{\Gamma}_K, \check{\Gamma}'_K, \check{\Gamma}''_K)\}$ *is sorted sorted alphabetically, with respect to* $(\check{\Gamma}''_k)^{-1}$ *(with higher priority) and* $\check{\Gamma}'_k$ *(with lower priority). If two consecutive neighbors in the alphabetized list* $\{\mathscr{G}_{(1)} = (\check{\Gamma}_{(1)}, \check{\Gamma}'_{(1)}, \check{\Gamma}''_{(1)}), \ldots, \mathscr{G}_{(K)} = (\check{\Gamma}_{(K)}, \check{\Gamma}'_{(K)}, \check{\Gamma}''_{(K)})\}$ *satisfy*

$$\mathrm{VbAspTest}(\hat{\Gamma}''_{(k)}, \hat{\Gamma}''_{(k+1)}) = \mathbf{FALSE}$$

AND

$$\mathrm{VbAspTest}(\hat{\Gamma}'_{(k)}, \hat{\Gamma}'_{(k+1)}) = \mathbf{FALSE}$$

where $k \in \mathbb{Z} \cap [1, K)$*, then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{\mathscr{G}_{(1)}, \ldots, \mathscr{G}_{(K)}\}$ *is divided into separate groups of tagged clusters* $\{\widetilde{\Gamma}_1 = \{\mathscr{G}_{(1,1)}, \ldots, \mathscr{G}_{(1,n_1)}\}, \ldots, \widetilde{\Gamma}_J = \{\mathscr{G}_{(J,1)}, \ldots, \mathscr{G}_{(1,n_J)}\}\}$*, from which the output list of word clusters is generated, after removal of all the tags from each* $\widetilde{\Gamma}_j$*, where* $j \in \mathbb{Z} \cap [1, J]$*.*

Example 7.13.1. We refer to the Polish version of Wiktionary (`https://pl.wiktionary.org/wiki/Kategoria:Gramatyka_j%C4%99zyka_polskiego`), for the divisions of declension and conjugation classes, of nouns and verbs respectively.

Here are some nouns that display the 1st masculine declension patterns:

drani, *drania*, *draniach*, *draniami*, *dranie*, *draniem*, *draniom*, *draniowi*, *draniów*, *draniu*, *drań* — "rascal";

goryl, *goryla*, *gorylach*, *gorylami*, *goryle*, *gorylem*, *goryli*, *gorylom*, *gorylowi*, *gorylu* — "gorilla";

liści, *liścia*, *liściach*, *liście*, *liściem*, *liściom*, *liściowi*, *liściu*, *liść*, *liśćmi* — "leaf".

2nd masculine declension:

owoc, *owocach*, *owocami*, *owoce*, *owocem*, *owocom*, *owocowi*, *owoców*, *owocu* — "fruit";

stróż, *stróża*, *stróżach*, *stróżami*, *stróżem*, *stróżom*, *stróżowi*, *stróżowie*, *stróżu*, *stróży* — "guard";

tasiemca, *tasiemcach*, *tasiemcami*, *tasiemce*, *tasiemcem*, *tasiemcom*, *tasiemcowi*, *tasiemców*, *tasiemcu*, *tasiemiec* — "tapeworm".

3rd masculine declension:

chirurdzy, *chirurg*, *chirurga*, *chirurgach*, *chirurgami*, *chirurgiem*, *chirurgom*, *chirurgowi*, *chirurgów*, *chirurgu* — "surgeon";

kożuch, *kożucha*, *kożuchach*, *kożuchami*, *kożuchem*, *kożuchom*, *kożuchowi*, *kożuchów*, *kożuchu*, *kożuchy* — "sheepskin";

strażacy, *strażak*, *strażaka*, *strażakach*, *strażakami*, *strażakiem*, *strażakom*, *strażakowi*, *strażaków*, *strażaku* — "fireman".

4th masculine declension:

chleb, *chleba*, *chlebach*, *chlebami*, *chlebem*, *chlebie*, *chlebom*, *chlebowi*, *chlebów*, *chleby* — "bread";

doktor, *doktora*, *doktorach*, *doktorami*, *doktorem*, *doktorom*, *doktorowi*, *doktorowie*, *doktorów*, *doktorze*, *doktorzy* — "doctor";

drozd, *drozda*, *drozdach*, *drozdami*, *drozdem*, *drozdom*, *drozdowi*, *drozdów*, *drozdy*, *droździe* — "thrush".

5th masculine declension:

Marsjan, *Marsjanach*, *Marsjanami*, *Marsjanie*, *Marsjanin*, *Marsjanina*, *Marsjaninem*, *Marsjaninie*, *Marsjaninowi*, *Marsjanom*, *Marsjany* — "Martian".

1st neuter declension:

pola, *polach*, *polami*, *pole*, *polem*, *polom*, *polu*, *pól* — "field";

studia, *studiach*, *studiami*, *studiem*, *studio*, *studiom*, *studiów*, *studiu* — "studio".

2nd neuter declension:

płuc, *płuca*, *płucach*, *płucami*, *płucem*, *płuco*, *płucom*, *płucu* — "lung".

3rd neuter declension:

cuda, *cudach*, *cudami*, *cudem*, *cudo*, *cudom*, *cudów*, *cudu*, *cudzie* — "wonder";

palcie, *palt*, *palta*, *paltach*, *paltami*, *paltem*, *palto*, *paltom*, *paltu* — "overcoat".

4th neuter declension:

dziewcząt, *dziewczę*, *dziewczęcia*, *dziewczęciem*, *dziewczęciu*, *dziewczęta*, *dziewczętach*, *dziewczętami*, *dziewczętom* — “girl”.

5th neuter declension:

ramienia, *ramieniem*, *ramieniu*, *ramię*, *ramion*, *ramiona*, *ramionach*, *ramionami*, *ramionom* — “shoulder”.

6th neuter declension:

muzea, *muzeach*, *muzeami*, *muzeom*, *muzeów*, *muzeum* — “museum”.

1st feminine declension:

pani, *paniach*, *paniami*, *panią*, *panie*, *paniom*, *pań* — “lady”;

świni, *świnia*, *świniach*, *świniami*, *świnią*, *świnie*, *świnię*, *świnio*, *świniom*, *świń* — “pig”.

2nd feminine declension:

prac, *praca*, *pracach*, *pracami*, *pracą*, *prace*, *pracę*, *praco*, *pracom*, *pracy* — “job”;

sadza, *sadzach*, *sadzami*, *sadzą*, *sadze*, *sadzę*, *sadzo*, *sadzom*, *sadzy* — “soot”.

3rd feminine declension:

much, *mucha*, *muchach*, *muchami*, *muchą*, *muchę*, *mucho*, *muchom*, *muchy*, *musze* — “bowtie”;

plotce, *plotek*, *plotka*, *plotkach*, *plotkami*, *plotką*, *plotkę*, *plotki*, *plotko*, *plotkom* — “rumour”.

4th feminine declension:

form, *forma*, *formach*, *formami*, *formą*, *formę*, *formie*, *formo*, *formom*, *formy* — “shape”.

5th feminine declension:

brew, *brwi*, *brwiach*, *brwiami*, *brwią*, *brwiom* — “eyebrow”;

kolei, *kolej*, *kolejach*, *kolejami*, *koleją*, *koleje*, *kolejom* — “rail”;

kości, *kościach*, *kością*, *kościom*, *kość*, *kośćmi* — “bone”.

6th feminine declension:

klacz, *klaczach*, *klaczami*, *klaczą*, *klacze*, *klaczom*, *klaczy* — “mare”;

mysz, *myszach*, *myszami*, *myszą*, *myszom*, *myszy* — “mouse”.

As we send these nouns into our clustering algorithm, we receive the following results:

{*brew*, *brwi*, *brwiach*, *brwiami*, *brwią*, *brwiom*}, {*chirurdzy*, *chirurg*, *chirurga*, *chirurgach*, *chirurgami*, *chirurgiem*, *chirurgom*, *chirurgowi*, *chirurgów*, *chirurgu*}, {*chleb*, *chleba*, *chlebach*, *chlebami*, *chlebem*, *chlebie*, *chlebom*, *chlebowi*, *chlebów*, *chleby*}, {*cuda*, *cudach*, *cudami*, *cudem*, *cudo*, *cudom*, *cudów*, *cudu*, *cudzie*}, {*doktor*, *doktora*, *doktorach*, *doktorami*, *doktorem*, *doktorom*, *doktorowi*, *doktorowie*, *doktorów*, *doktorze*, *doktorzy*}, {*drani*, *drania*, *draniach*, *draniami*, *dranie*, *draniem*, *draniom*, *draniowi*, *draniów*, *draniu*, *drań*}, {*drozd*, *drozda*, *drozdach*, *drozdami*, *drozdem*, *drozdom*, *drozdowi*, *drozdów*, *drozdy*, *droździe*}, {*dziewcząt*, *dziewczę*, *dziewczęcia*, *dziewczęciem*, *dziewczęciu*, *dziewczęta*, *dziewczętach*, *dziewczętami*, *dziewczętom*}, {*form*, *forma*, *formach*, *formami*, *formą*, *formę*, *formie*, *formo*, *formom*, *formy*}, {*goryl*, *goryla*, *gorylach*, *gorylami*, *goryle*, *gorylem*, *goryli*, *gorylom*, *gorylowi*, *gorylu*}, {*klacz*, *klaczach*, *klaczami*, *klaczą*, *klacze*, *klaczom*, *klaczy*}, {*kolei*, *kolej*, *kolejach*, *kolejami*, *koleją*, *koleje*, *kolejom*}, {*kości*, *kościach*, *kością*, *kościom*, *kość*, *kośćmi*}, {*kożuch*, *kożucha*, *kożuchach*, *kożuchami*, *kożuchem*, *kożuchom*, *kożuchowi*, *kożuchów*, *kożuchu*, *kożuchy*}, {*liści*, *liścia*, *liściach*, *liście*, *liściem*, *liściom*, *liściowi*, *liściu*, *liść*, *liśćmi*}, {*Marsjan*, *Marsjanach*, *Marsjanami*, *Marsjanie*, *Marsjanin*, *Marsjanina*, *Marsjaninem*, *Marsjaninie*, *Marsjaninowi*, *Marsjanom*, *Marsjany*}, {*much*, *mucha*, *muchach*, *muchami*, *muchą*, *muchę*, *mucho*, *muchom*, *muchy*, *musze*}, {*muzea*, *muzeach*, *muzeami*, *muzeom*, *muzeów*, *muzeum*}, {*mysz*, *myszach*, *myszami*, *myszą*, *myszom*, *myszy*}, {*owoc*, *owocach*, *owocami*, *owoce*, *owocem*, *owocom*,

owocowi, *owoców*, *owocu*}, {*palcie*, *palt*, *palta*, *paltach*, *paltami*, *paltem*, *palto*, *paltom*, *paltu*}, {*pani*, *paniach*, *paniami*, *panią*, *panie*, *paniom*, *pań*}, {*plotce*, *plotek*, *plotka*, *plotkach*, *plotkami*, *plotką*, *plotkę*, *plotki*, *plotko*, *plotkom*}, {*płuc*, *płuca*, *płucach*, *płucami*, *płucem*, *płuco*, *płucom*, *płucu*}, {*pola*, *polach*, *polami*, *pole*, *polem*, *polom*, *polu*, *pól*}, {*prac*, *praca*, *pracach*, *pracami*, *pracą*, *prace*, *pracę*, *praco*, *pracom*, *pracy*}, {*ramienia*, *ramieniem*, *ramieniu*, *ramię*, *ramion*, *ramiona*, *ramionach*, *ramionami*, *ramionom*}, {*sadza*, *sadzach*, *sadzami*, *sadzą*, *sadze*, *sadzę*, *sadzo*, *sadzom*, *sadzy*}, {*strażacy*, *strażak*, *strażaka*, *strażakach*, *strażakami*, *strażakiem*, *strażakom*, *strażakowi*, *strażaków*, *strażaku*}, {*stróż*, *stróża*, *stróżach*, *stróżami*, *stróżem*, *stróżom*, *stróżowi*, *stróżowie*, *stróżu*, *stróży*}, {*studia*, *studiach*, *studiami*, *studiem*, *studio*, *studiom*, *studiów*, *studiu*}, {*świni*, *świnia*, *świniach*, *świniami*, *świnią*, *świnie*, *świnię*, *świnio*, *świniom*, *świń*}, {*tasiemca*, *tasiemcach*, *tasiemcami*, *tasiemce*, *tasiemcem*, *tasiemcom*, *tasiemcowi*, *tasiemców*, *tasiemcu*, *tasiemiec*}.

We pick the following Polish verbs, followed by their English translations (enclosed in quotation marks) and conjugation classes (enclosed in parentheses):

czyta, *czytacie*, *czytać*, *czytaj*, *czytają*, *czytając*, *czytająca*, *czytające*, *czytający*, *czytajcie*, *czytajmy*, *czytali*, *czytaliby*, *czytalibyście*, *czytalibyśmy*, *czytaliście*, *czytaliśmy*, *czytał*, *czytała*, *czytałaby*, *czytałabym*, *czytałabyś*, *czytałam*, *czytałaś*, *czytałby*, *czytałbym*, *czytałbyś*, *czytałem*, *czytałeś*, *czytało*, *czytałoby*, *czytały*, *czytałyby*, *czytałybyście*, *czytałybyśmy*, *czytałyście*, *czytałyśmy*, *czytam*, *czytamy*, *czytana*, *czytane*, *czytani*, *czytanie*, *czytano*, *czytany*, *czytasz*, *przeczyta*, *przeczytacie*, *przeczytać*, *przeczytaj*, *przeczytają*, *przeczytające*, *przeczytający*, *przeczytajcie*, *przeczytajmy*, *przeczytali*, *przeczytaliby*, *przeczytalibyście*, *przeczytalibyśmy*, *przeczytaliście*, *przeczytaliśmy*, *przeczytał*, *przeczytała*, *przeczytałaby*, *przeczytałabym*, *przeczytałabyś*, *przeczytałam*, *przeczytałaś*, *przeczytałby*, *przeczytałbym*, *przeczytałbyś*, *przeczytałem*, *przeczytałeś*, *przeczytało*, *przeczytałoby*, *przeczytały*, *przeczytałyby*, *przeczytałybyście*, *przeczytałybyśmy*, *przeczytałyście*, *przeczytałyśmy*, *przeczytam*, *przeczytamy*, *przeczytana*, *przeczytane*, *przeczytani*, *przeczytanie*, *przeczytano*, *przeczytany*, *przeczytasz*, *przeczytawszy* — “read” (1);

umiał, *umiała*, *umiałaby*, *umiałabym*, *umiałabyś*, *umiałam*, *umiałaś*, *umiałby*, *umiałbym*, *umiałbyś*, *umiałem*, *umiałeś*, *umiało*, *umiałoby*, *umiały*, *umiałyby*, *umiałybyście*, *umiałybyśmy*, *umiałyście*, *umiałyśmy*, *umie*, *umiecie*, *umieć*, *umiej*, *umieją*, *umiejąc*, *umiejąca*, *umiejące*, *umiejący*, *umiejcie*, *umiejmy*, *umieli*, *umieliby*, *umielibyście*, *umielibyśmy*, *umieliście*, *umieliśmy*, *umiem*, *umiemy*, *umiesz* — “can” (2);

taniali, *taniał*, *taniała*, *taniałaby*, *taniałabym*, *taniałabyś*, *taniałam*, *taniałaś*, *taniałby*, *taniałbym*, *taniałbyś*, *taniałe*, *taniałem*, *taniałeś*, *taniało*, *taniałoby*, *taniałobym*, *taniałobyś*, *taniałom*, *taniałoś*, *taniały*, *taniałyby*, *taniałybyście*, *taniałybyśmy*, *taniałyście*, *taniałyśmy*, *taniano*, *tanieć*, *taniej*, *tanieją*, *taniejąc*, *taniejąca*, *taniejące*, *taniejący*, *taniejcie*, *tanieje*, *taniejecie*, *taniejemy*, *taniejesz*, *tanieję*, *taniejmy*, *tanieli*, *tanieliby*, *tanielibyście*, *tanielibyśmy*, *tanieliście*, *tanieliśmy*, *tanienie* — “cheapen” (3);

malować, *malowali*, *malowaliby*, *malowalibyście*, *malowalibyśmy*, *malowaliście*, *malowaliśmy*, *malował*, *malowała*, *malowałaby*, *malowałabym*, *malowałabyś*, *malowałam*, *malowałaś*, *malowałby*, *malowałbym*, *malowałbyś*, *malowałem*, *malowałeś*, *malowało*, *malowałoby*, *malowały*, *malowałyby*, *malowałybyście*, *malowałybyśmy*, *malowałyście*, *malowałyśmy*, *malowana*, *malowane*, *malowani*, *malowanie*, *malowano*, *malowany*, *maluj*, *malują*, *malując*, *malująca*, *malujące*, *malujący*, *malujcie*, *maluje*, *malujecie*, *malujemy*, *malujesz*, *maluję*, *malujmy*, *pomalować*, *pomalowali*, *pomalowaliby*, *pomalowalibyście*, *pomalowalibyśmy*, *pomalowaliście*, *pomalowaliśmy*, *pomalował*, *pomalowała*, *pomalowałaby*, *pomalowałabym*, *pomalowałabyś*, *pomalowałam*, *pomalowałaś*, *pomalowałby*, *pomalowałbym*, *pomalowałbyś*, *pomalowałem*, *pomalowałeś*, *pomalowało*, *pomalowałoby*, *pomalowałobym*, *pomalowałobyś*, *pomalowałom*, *pomalowałoś*, *pomalowały*, *pomalowałyby*, *pomalowałybyście*, *pomalowałybyśmy*, *pomalowałyście*, *pomalowałyśmy*, *pomalowana*, *pomalowane*, *pomalowani*, *pomalowanie*, *pomalowano*, *pomalowany*, *pomalowawszy*, *pomaluj*, *pomalują*, *pomalujcie*, *pomaluje*, *pomalujecie*, *pomalujemy*, *pomalujesz*, *pomaluję*, *pomalujmy* — “portray” (4);

ciągną, *ciągnąc*, *ciągnąca*, *ciągnące*, *ciągnący*, *ciągnąć*, *ciągnął*, *ciągnąłby*, *ciągnąłbym*, *ciągnąłbyś*, *ciągnąłem*, *ciągnąłeś*, *ciągnę*, *ciągnęli*, *ciągnęliby*, *ciągnęlibyście*, *ciągnęlibyśmy*, *ciągnęliście*, *ciągnęliśmy*, *ciągnęła*, *ciągnęłaby*, *ciągnęłabym*, *ciągnęłabyś*, *ciągnęłam*, *ciągnęłaś*, *ciągnęło*, *ciągnęłoby*, *ciągnęły*, *ciągnęłyby*, *ciągnęłybyście*, *ciągnęłybyśmy*, *ciągnęłyście*, *ciągnęłyśmy*, *ciągnie*, *ciągniecie*, *ciągniemy*, *ciągnieni*, *ciągniesz*, *ciągnięci*, *ciągnięcie*, *ciągnięta*, *ciągnięte*, *ciągnięto*, *ciągnięty*, *ciągnij*, *ciągnijcie*, *ciągnijmy*, *ciągniona*, *ciągnione*, *ciągniony* — “pull” (5a);

płyną, *płynąc*, *płynąca*, *płynące*, *płynący*, *płynąć*, *płynął*, *płynąłby*, *płynąłbym*, *płynąłbyś*, *płynąłem*, *płynąłeś*, *płynę*, *płynęli*, *płynęliby*, *płynęlibyście*, *płynęlibyśmy*, *płynęliście*, *płynęliśmy*, *płynęła*, *płynęłaby*, *płynęłabym*, *płynęłabyś*, *płynęłam*, *płynęłaś*, *płynęło*, *płynęłoby*, *płynęły*, *płynęłyby*, *płynęłybyście*, *płynęłybyśmy*, *płynęłyście*, *płynęłyśmy*, *płynie*, *płyniecie*, *płyniemy*, *płyniesz*, *płynięcie*, *płynięto*, *płyń*, *płyńcie*, *płyńmy* — “swim” (5b);

chudli, *chudliby*, *chudlibyście*, *chudlibyśmy*, *chudliście*, *chudliśmy*, *chudł*, *chudła*, *chudłaby*, *chudłabym*, *chudłabyś*, *chudłam*, *chudłaś*, *chudłby*, *chudłbym*, *chudłbyś*, *chudłem*, *chudłeś*, *chudło*, *chudłoby*, *chudły*, *chudłyby*, *chudłybyście*, *chudłybyśmy*, *chudłyście*, *chudłyśmy*, *chudną*, *chudnąc*, *chudnąca*, *chudnące*, *chudnący*, *chudnąć*, *chudnę*, *chudnie*, *chudniecie*, *chudniemy*, *chudniesz*, *chudnij*, *chudnijcie*, *chudnijmy* — "lose weight" (5c);

robi, *robią*, *robiąc*, *robiąca*, *robiące*, *robiący*, *robicie*, *robić*, *robieni*, *robienie*, *robię*, *robili*, *robiliby*, *robilibyście*, *robilibyśmy*, *robiliście*, *robiliśmy*, *robił*, *robiła*, *robiłaby*, *robiłabym*, *robiłabyś*, *robiłam*, *robiłaś*, *robiłby*, *robiłbym*, *robiłbyś*, *robiłem*, *robiłeś*, *robiło*, *robiłoby*, *robiłobym*, *robiłobyś*, *robiłom*, *robiłoś*, *robiły*, *robiłyby*, *robiłybyście*, *robiłybyśmy*, *robiłyście*, *robiłyśmy*, *robimy*, *robiona*, *robione*, *robiono*, *robiony*, *robisz*, *zrobi*, *zrobią*, *zrobicie*, *zrobić*, *zrobieni*, *zrobienie*, *zrobię*, *zrobili*, *zrobiliby*, *zrobilibyście*, *zrobilibyśmy*, *zrobiliście*, *zrobiliśmy*, *zrobił*, *zrobiła*, *zrobiłaby*, *zrobiłabym*, *zrobiłabyś*, *zrobiłam*, *zrobiłaś*, *zrobiłby*, *zrobiłbym*, *zrobiłbyś*, *zrobiłem*, *zrobiłeś*, *zrobiło*, *zrobiłoby*, *zrobiłobym*, *zrobiłobyś*, *zrobiłom*, *zrobiłoś*, *zrobiły*, *zrobiłyby*, *zrobiłybyście*, *zrobiłybyśmy*, *zrobiłyście*, *zrobiłyśmy*, *zrobimy*, *zrobiona*, *zrobione*, *zrobiono*, *zrobiony*, *zrobisz*, *zrobiwszy* — "do" (6a);

uwierz, *uwierzą*, *uwierzcie*, *uwierzenie*, *uwierzę*, *uwierzmy*, *uwierzono*, *uwierzy*, *uwierzycie*, *uwierzyć*, *uwierzyli*, *uwierzyliby*, *uwierzylibyście*, *uwierzylibyśmy*, *uwierzyliście*, *uwierzyliśmy*, *uwierzył*, *uwierzyła*, *uwierzyłaby*, *uwierzyłabym*, *uwierzyłabyś*, *uwierzyłam*, *uwierzyłaś*, *uwierzyłby*, *uwierzyłbym*, *uwierzyłbyś*, *uwierzyłem*, *uwierzyłeś*, *uwierzyło*, *uwierzyłoby*, *uwierzyłobym*, *uwierzyłobyś*, *uwierzyłom*, *uwierzyłoś*, *uwierzyły*, *uwierzyłyby*, *uwierzyłybyście*, *uwierzyłybyśmy*, *uwierzyłyście*, *uwierzyłyśmy*, *uwierzymy*, *uwierzysz*, *uwierzywszy*, *wierz*, *wierzą*, *wierząc*, *wierząca*, *wierzące*, *wierzący*, *wierzcie*, *wierzenie*, *wierzę*, *wierzmy*, *wierzono*, *wierzy*, *wierzycie*, *wierzyć*, *wierzyli*, *wierzyliby*, *wierzylibyście*, *wierzylibyśmy*, *wierzyliście*, *wierzyliśmy*, *wierzył*, *wierzyła*, *wierzyłaby*, *wierzyłabym*, *wierzyłabyś*, *wierzyłam*, *wierzyłaś*, *wierzyłby*, *wierzyłbym*, *wierzyłbyś*, *wierzyłem*, *wierzyłeś*, *wierzyło*, *wierzyłoby*, *wierzyłobym*, *wierzyłobyś*, *wierzyłom*, *wierzyłoś*, *wierzyły*, *wierzyłyby*, *wierzyłybyście*, *wierzyłybyśmy*, *wierzyłyście*, *wierzyłyśmy*, *wierzymy*, *wierzysz* — "believe" (6b);

widzą, *widząc*, *widząca*, *widzące*, *widzący*, *widzenie*, *widzę*, *widzi*, *widział*, *widziała*, *widziałaby*, *widziałabym*, *widziałabyś*, *widziałam*, *widziałaś*, *widziałby*, *widziałbym*, *widziałbyś*, *widziałem*, *widziałeś*, *widziało*, *widziałoby*, *widziałobym*, *widziałobyś*, *widziałom*, *widziałoś*, *widziały*, *widziałyby*, *widziałybyście*, *widziałybyśmy*, *widziałyście*, *widziałyśmy*, *widziana*, *widziane*, *widziani*, *widziano*, *widziany*, *widzicie*, *widzieć*, *widzieli*, *widzieliby*, *widzielibyście*, *widzielibyśmy*, *widzieliście*, *widzieliśmy*, *widzimy*, *widzisz*, *widź*, *widźcie*, *widźmy*, *zobacz*, *zobaczą*, *zobaczcie*, *zobaczenie*, *zobaczę*, *zobaczmy*, *zobaczono*, *zobaczy*, *zobaczycie*, *zobaczyć*, *zobaczyli*, *zobaczyliby*, *zobaczylibyście*, *zobaczylibyśmy*, *zobaczyliście*, *zobaczyliśmy*, *zobaczył*, *zobaczyła*, *zobaczyłaby*, *zobaczyłabym*, *zobaczyłabyś*, *zobaczyłam*, *zobaczyłaś*, *zobaczyłby*, *zobaczyłbym*, *zobaczyłbyś*, *zobaczyłem*, *zobaczyłeś*, *zobaczyło*, *zobaczyłoby*, *zobaczyły*, *zobaczyłyby*, *zobaczyłybyście*, *zobaczyłybyśmy*, *zobaczyłyście*, *zobaczyłyśmy*, *zobaczymy*, *zobaczysz*, *zobaczywszy* — "see" (7a);

leż, *leżał*, *leżała*, *leżałaby*, *leżałabym*, *leżałabyś*, *leżałam*, *leżałaś*, *leżałby*, *leżałbym*, *leżałbyś*, *leżałem*, *leżałeś*, *leżało*, *leżałoby*, *leżałobym*, *leżałobyś*, *leżałom*, *leżałoś*, *leżały*, *leżałyby*, *leżałybyście*, *leżałybyśmy*, *leżałyście*, *leżałyśmy*, *leżano*, *leżą*, *leżąc*, *leżąca*, *leżące*, *leżący*, *leżcie*, *leżeć*, *leżeli*, *leżeliby*, *leżelibyście*, *leżelibyśmy*, *leżeliście*, *leżeliśmy*, *leżenie*, *leżę*, *leżmy*, *leży*, *leżycie*, *leżymy*, *leżysz* — "lie" (7b);

odczyta, *odczytacie*, *odczytać*, *odczytaj*, *odczytają*, *odczytajcie*, *odczytajmy*, *odczytali*, *odczytaliby*, *odczytalibyście*, *odczytalibyśmy*, *odczytaliście*, *odczytaliśmy*, *odczytał*, *odczytała*, *odczytałaby*, *odczytałabym*, *odczytałabyś*, *odczytałam*, *odczytałaś*, *odczytałby*, *odczytałbym*, *odczytałbyś*, *odczytałem*, *odczytałeś*, *odczytało*, *odczytałoby*, *odczytałobym*, *odczytałobyś*, *odczytałom*, *odczytałoś*, *odczytały*, *odczytałyby*, *odczytałybyście*, *odczytałybyśmy*, *odczytałyście*, *odczytałyśmy*, *odczytam*, *odczytamy*, *odczytana*, *odczytane*, *odczytani*, *odczytanie*, *odczytano*, *odczytany*, *odczytasz*, *odczytawszy*, *odczytuj*, *odczytują*, *odczytując*, *odczytująca*, *odczytujące*, *odczytujący*, *odczytujcie*, *odczytuje*, *odczytujecie*, *odczytujemy*, *odczytujesz*, *odczytuję*, *odczytujmy*, *odczytywać*, *odczytywali*, *odczytywaliby*, *odczytywalibyście*, *odczytywalibyśmy*, *odczytywaliście*, *odczytywaliśmy*, *odczytywał*, *odczytywała*, *odczytywałaby*, *odczytywałabym*, *odczytywałabyś*, *odczytywałam*, *odczytywałaś*, *odczytywałby*, *odczytywałbym*, *odczytywałbyś*, *odczytywałem*, *odczytywałeś*, *odczytywało*, *odczytywałoby*, *odczytywałobym*, *odczytywałobyś*, *odczytywałom*, *odczytywałoś*, *odczytywały*, *odczytywałyby*, *odczytywałybyście*, *odczytywałybyśmy*, *odczytywałyście*, *odczytywałyśmy*, *odczytywana*, *odczytywane*, *odczytywani*, *odczytywanie*, *odczytywano*, *odczytywany* — "decipher" (8a);

zyska, *zyskacie*, *zyskać*, *zyskaj*, *zyskają*, *zyskajcie*, *zyskajmy*, *zyskali*, *zyskaliby*, *zyskalibyście*, *zyskalibyśmy*, *zyskaliście*, *zyskaliśmy*, *zyskał*, *zyskała*, *zyskałaby*, *zyskałabym*, *zyskałabyś*, *zyskałam*, *zyskałaś*, *zyskałby*, *zyskałbym*, *zyskałbyś*, *zyskałem*, *zyskałeś*, *zyskało*, *zyskałoby*, *zyskałobym*, *zyskałobyś*, *zyskałom*, *zyskałoś*, *zyskały*, *zyskałyby*, *zyskałybyście*, *zyskałybyśmy*, *zyskałyście*, *zyskałyśmy*, *zyskam*, *zyskamy*, *zyskana*, *zyskane*, *zyskani*, *zyskanie*,

zyskano, *zyskany*, *zyskasz*, *zyskawszy*, *zyskiwać*, *zyskiwali*, *zyskiwaliby*, *zyskiwalibyście*, *zyskiwalibyśmy*, *zyskiwaliście*, *zyskiwaliśmy*, *zyskiwał*, *zyskiwała*, *zyskiwałaby*, *zyskiwałabym*, *zyskiwałabyś*, *zyskiwałam*, *zyskiwałaś*, *zyskiwałby*, *zyskiwałbym*, *zyskiwałbyś*, *zyskiwałem*, *zyskiwałeś*, *zyskiwało*, *zyskiwałoby*, *zyskiwały*, *zyskiwałyby*, *zyskiwałybyście*, *zyskiwałybyśmy*, *zyskiwałyście*, *zyskiwałyśmy*, *zyskiwana*, *zyskiwane*, *zyskiwani*, *zyskiwanie*, *zyskiwano*, *zyskiwany*, *zyskuj*, *zyskują*, *zyskując*, *zyskująca*, *zyskujące*, *zyskujący*, *zyskujcie*, *zyskuje*, *zyskujecie*, *zyskujemy*, *zyskujesz*, *zyskuję*, *zyskujmy* — "gain" (8a);

mazać, *mazali*, *mazaliby*, *mazalibyście*, *mazalibyśmy*, *mazaliście*, *mazaliśmy*, *mazał*, *mazała*, *mazałaby*, *mazałabym*, *mazałabyś*, *mazałam*, *mazałaś*, *mazałby*, *mazałbym*, *mazałbyś*, *mazałem*, *mazałeś*, *mazało*, *mazałoby*, *mazały*, *mazałyby*, *mazałybyście*, *mazałybyśmy*, *mazałyście*, *mazałyśmy*, *mazana*, *mazane*, *mazani*, *mazanie*, *mazano*, *mazany*, *maż*, *mażą*, *mażąc*, *mażąca*, *mażące*, *mażący*, *mażcie*, *maże*, *mażecie*, *mażemy*, *mażesz*, *mażę*, *mażmy* — "smear" (9);

pici, *picie*, *pić*, *pij*, *piją*, *pijąc*, *pijąca*, *pijące*, *pijący*, *pijcie*, *pije*, *pijecie*, *pijemy*, *pijesz*, *piję*, *pijmy*, *pili*, *piliby*, *pilibyście*, *pilibyśmy*, *piliście*, *piliśmy*, *pił*, *piła*, *piłaby*, *piłabym*, *piłabyś*, *piłam*, *piłaś*, *piłby*, *piłbym*, *piłbyś*, *piłem*, *piłeś*, *piło*, *piłoby*, *piłobym*, *piłobyś*, *piłom*, *piłoś*, *piły*, *piłyby*, *piłybyście*, *piłybyśmy*, *piłyście*, *piłyśmy*, *pita*, *pite*, *pito*, *pity* — "drink" (10a);

lać, *lali*, *laliby*, *lalibyście*, *lalibyśmy*, *laliście*, *laliśmy*, *lał*, *lała*, *lałaby*, *lałabym*, *lałabyś*, *lałam*, *lałaś*, *lałby*, *lałbym*, *lałbyś*, *lałem*, *lałeś*, *lało*, *lałoby*, *ały*, *ałyby*, *ałybyście*, *ałybyśmy*, *ałyście*, *ałyśmy*, *lana*, *lane*, *lani*, *lanie*, *lano*, *lany*, *lej*, *leją*, *lejąc*, *lejąca*, *lejące*, *lejący*, *lejcie*, *leje*, *lejecie*, *lejemy*, *lejesz*, *leję*, *lejmy* — "pour" (10b);

bierz, *bierzcie*, *bierze*, *bierzecie*, *bierzemy*, *bierzesz*, *bierzmy*, *biorą*, *biorąc*, *biorąca*, *biorące*, *biorący*, *biorę*, *brać*, *brali*, *braliby*, *bralibyście*, *bralibyśmy*, *braliście*, *braliśmy*, *brał*, *brała*, *brałaby*, *brałabym*, *brałabyś*, *brałam*, *brałaś*, *brałby*, *brałbym*, *brałbyś*, *brałem*, *brałeś*, *brało*, *brałoby*, *brały*, *brałyby*, *brałybyście*, *brałybyśmy*, *brałyście*, *brałyśmy*, *brana*, *brane*, *brani*, *branie*, *brany*, *wezmą*, *wezmę*, *weźmiesz*, *wziąć*, *wziął*, *wziąłby*, *wziąłbym*, *wziąłbyś*, *wziąłem*, *wziąłeś*, *wziąwszy*, *wzięci*, *wzięcie*, *wzięli*, *wzięliby*, *wzięlibyście*, *wzięlibyśmy*, *wzięliście*, *wzięliśmy*, *wzięła*, *wzięłaby*, *wzięłabym*, *wzięłabyś*, *wzięłam*, *wzięłaś*, *wzięło*, *wzięłoby*, *wzięły*, *wzięłyby*, *wzięłybyście*, *wzięłybyśmy*, *wzięłyście*, *wzięłyśmy*, *wzięta*, *wzięte*, *wzięto*, *wzięty* — "take" (10c);

wiezie, *wieziecie*, *wieziemy*, *wiezieni*, *wiezienie*, *wieziesz*, *wieziona*, *wiezione*, *wieziono*, *wieziony*, *wiozą*, *wioząc*, *wiozą*ca, *wiozące*, *wiozący*, *wiozę*, *wiozła*, *wiozłaby*, *wiozłabym*, *wiozłabyś*, *wiozłam*, *wiozłaś*, *wiozło*, *wiozłoby*, *wiozły*, *wiozłyby*, *wiozłybyście*, *wiozłybyśmy*, *wiozłyście*, *wiozłyśmy*, *wiózł*, *wiózłby*, *wiózłbym*, *wiózłbyś*, *wiózłem*, *wiózłeś* — "carry" (11).

These verbs are clustered by our algorithm as follows:

{*bierz*, *bierzcie*, *bierze*, *bierzecie*, *bierzemy*, *bierzesz*, *bierzmy*, *biorą*, *biorąc*, *biorąca*, *biorące*, *biorący*, *biorę*, *brać*, *brali*, *braliby*, *bralibyście*, *bralibyśmy*, *braliście*, *braliśmy*, *brał*, *brała*, *brałaby*, *brałabym*, *brałabyś*, *brałam*, *brałaś*, *brałby*, *brałbym*, *brałbyś*, *brałem*, *brałeś*, *brało*, *brałoby*, *brały*, *brałyby*, *brałybyście*, *brałybyśmy*, *brałyście*, *brałyśmy*, *brana*, *brane*, *brani*, *branie*, *brany*, *wezmą*, *wezmę*, *weźmiesz*, *wziąć*, *wziął*, *wziąłby*, *wziąłbym*, *wziąłbyś*, *wziąłem*, *wziąłeś*, *wziąwszy*, *wzięci*, *wzięcie*, *wzięli*, *wzięliby*, *wzięlibyście*, *wzięlibyśmy*, *wzięliście*, *wzięliśmy*, *wzięła*, *wzięłaby*, *wzięłabym*, *wzięłabyś*, *wzięłam*, *wzięłaś*, *wzięło*, *wzięłoby*, *wzięły*, *wzięłyby*, *wzięłybyście*, *wzięłybyśmy*, *wzięłyście*, *wzięłyśmy*, *wzięta*, *wzięte*, *wzięto*, *wzięty*},

{*chudli*, *chudliby*, *chudlibyście*, *chudlibyśmy*, *chudliście*, *chudliśmy*, *chudł*, *chudła*, *chudłaby*, *chudłabym*, *chudłabyś*, *chudłam*, *chudłaś*, *chudłby*, *chudłbym*, *chudłbyś*, *chudłem*, *chudłeś*, *chudło*, *chudłoby*, *chudły*, *chudłyby*, *chudłybyście*, *chudłybyśmy*, *chudłyście*, *chudłyśmy*, *chudną*, *chudnąc*, *chudnąca*, *chudnące*, *chudnący*, *chudnąć*, *chudnę*, *chudnie*, *chudniecie*, *chudniemy*, *chudniesz*, *chudnij*, *chudnijcie*, *chudnijmy*},

{*ciągną*, *ciągnąc*, *ciągnąca*, *ciągnące*, *ciągnący*, *ciągnąć*, *ciągnął*, *ciągnąłby*, *ciągnąłbym*, *ciągnąłbyś*, *ciągnąłem*, *ciągnąłeś*, *ciągnę*, *ciągnęli*, *ciągnęliby*, *ciągnęlibyście*, *ciągnęlibyśmy*, *ciągnęliście*, *ciągnęliśmy*, *ciągnęła*, *ciągnęłaby*, *ciągnęłabym*, *ciągnęłabyś*, *ciągnęłam*, *ciągnęłaś*, *ciągnęło*, *ciągnęłoby*, *ciągnęły*, *ciągnęłyby*, *ciągnęłybyście*, *ciągnęłybyśmy*, *ciągnęłyście*, *ciągnęłyśmy*, *ciągnie*, *ciągniecie*, *ciągniemy*, *ciągnieni*, *ciągniesz*, *ciągnięci*, *ciągnięcie*, *ciągnięta*, *ciągnięte*, *ciągnięto*, *ciągnięty*, *ciągnij*, *ciągnijcie*, *ciągnijmy*, *ciągniona*, *ciągnione*, *ciągniony*},

{*czyta*, *czytacie*, *czytać*, *czytaj*, *czytają*, *czytając*, *czytająca*, *czytające*, *czytający*, *czytajcie*, *czytajmy*, *czytali*, *czytaliby*, *czytalibyście*, *czytalibyśmy*, *czytaliście*, *czytaliśmy*, *czytał*, *czytała*, *czytałaby*, *czytałabym*, *czytałabyś*,

czytałam, *czytałaś*, *czytałby*, *czytałbym*, *czytałbyś*, *czytałem*, *czytałeś*, *czytało*, *czytałoby*, *czytały*, *czytałyby*, *czytałybyście*, *czytałybyśmy*, *czytałyście*, *czytałyśmy*, *czytam*, *czytamy*, *czytana*, *czytane*, *czytani*, *czytanie*, *czytano*, *czytany*, *czytasz*, *przeczyta*, *przeczytacie*, *przeczytać*, *przeczytaj*, *przeczytają*, *przeczytajcie*, *przeczytajmy*, *przeczytali*, *przeczytaliby*, *przeczytalibyście*, *przeczytalibyśmy*, *przeczytaliście*, *przeczytaliśmy*, *przeczytał*, *przeczytała*, *przeczytałaby*, *przeczytałabym*, *przeczytałabyś*, *przeczytałam*, *przeczytałaś*, *przeczytałby*, *przeczytałbym*, *przeczytałbyś*, *przeczytałem*, *przeczytałeś*, *przeczytało*, *przeczytałoby*, *przeczytały*, *przeczytałyby*, *przeczytałybyście*, *przeczytałybyśmy*, *przeczytałyście*, *przeczytałyśmy*, *przeczytam*, *przeczytamy*, *przeczytana*, *przeczytane*, *przeczytani*, *przeczytanie*, *przeczytano*, *przeczytany*, *przeczytasz*, *przeczytawszy*},

{*lać*, *lali*, *laliby*, *lalibyście*, *lalibyśmy*, *laliście*, *laliśmy*, *lał*, *lała*, *łałaby*, *łałabym*, *łałabyś*, *łałam*, *łałaś*, *łałby*, *łałbym*, *łałbyś*, *łałem*, *łałeś*, *łało*, *łałoby*, *łały*, *łałyby*, *łałybyście*, *łałybyśmy*, *łałyście*, *łałyśmy*, *lana*, *lane*, *lani*, *lanie*, *lano*, *lany*, *lej*, *leją*, *lejąc*, *lejąca*, *lejące*, *lejący*, *lejcie*, *leje*, *lejecie*, *lejemy*, *lejesz*, *leję*, *lejmy*},

{*leż*, *leżał*, *leżała*, *leżałaby*, *leżałabym*, *leżałabyś*, *leżałam*, *leżałaś*, *leżałby*, *leżałbym*, *leżałbyś*, *leżałem*, *leżałeś*, *leżało*, *leżałoby*, *leżałobym*, *leżałobyś*, *leżałom*, *leżałoś*, *leżały*, *leżałyby*, *leżałybyście*, *leżałybyśmy*, *leżałyście*, *leżałyśmy*, *leżano*, *leżą*, *leżąc*, *leżąca*, *leżące*, *leżący*, *leżcie*, *leżeć*, *leżeli*, *leżeliby*, *leżelibyście*, *leżelibyśmy*, *leżeliście*, *leżeliśmy*, *leżenie*, *leżę*, *leżmy*, *leży*, *leżycie*, *leżymy*, *leżysz*},

{*malować*, *malowali*, *malowaliby*, *malowalibyście*, *malowalibyśmy*, *malowaliście*, *malowaliśmy*, *malował*, *malowała*, *malowałaby*, *malowałabym*, *malowałabyś*, *malowałam*, *malowałaś*, *malowałby*, *malowałbym*, *malowałbyś*, *malowałem*, *malowałeś*, *malowało*, *malowałoby*, *malowały*, *malowałyby*, *malowałybyście*, *malowałybyśmy*, *malowałyście*, *malowałyśmy*, *malowana*, *malowane*, *malowani*, *malowanie*, *malowano*, *malowany*, *maluj*, *malują*, *malując*, *malująca*, *malujące*, *malujący*, *malujcie*, *maluje*, *malujecie*, *malujemy*, *malujesz*, *maluję*, *malujmy*},

{*mazać*, *mazali*, *mazaliby*, *mazalibyście*, *mazalibyśmy*, *mazaliście*, *mazaliśmy*, *mazał*, *mazała*, *mazałaby*, *mazałabym*, *mazałabyś*, *mazałam*, *mazałaś*, *mazałby*, *mazałbym*, *mazałbyś*, *mazałem*, *mazałeś*, *mazało*, *mazałoby*, *mazały*, *mazałyby*, *mazałybyście*, *mazałybyśmy*, *mazałyście*, *mazałyśmy*, *mazana*, *mazane*, *mazani*, *mazanie*, *mazano*, *mazany*, *maż*, *mażą*, *mażąc*, *mażąca*, *mażące*, *mażący*, *mażcie*, *maże*, *mażecie*, *mażemy*, *mażesz*, *mażę*, *mażmy*},

{*odczyta*, *odczytacie*, *odczytać*, *odczytaj*, *odczytają*, *odczytajcie*, *odczytajmy*, *odczytali*, *odczytaliby*, *odczytalibyście*, *odczytalibyśmy*, *odczytaliście*, *odczytaliśmy*, *odczytał*, *odczytała*, *odczytałaby*, *odczytałabym*, *odczytałabyś*, *odczytałam*, *odczytałaś*, *odczytałby*, *odczytałbym*, *odczytałbyś*, *odczytałem*, *odczytałeś*, *odczytało*, *odczytałoby*, *odczytałobym*, *odczytałobyś*, *odczytałom*, *odczytałoś*, *odczytały*, *odczytałyby*, *odczytałybyście*, *odczytałybyśmy*, *odczytałyście*, *odczytałyśmy*, *odczytam*, *odczytamy*, *odczytana*, *odczytane*, *odczytani*, *odczytanie*, *odczytano*, *odczytany*, *odczytasz*, *odczytawszy*, *odczytuj*, *odczytują*, *odczytując*, *odczytująca*, *odczytujące*, *odczytujący*, *odczytujcie*, *odczytuje*, *odczytujecie*, *odczytujemy*, *odczytujesz*, *odczytuję*, *odczytujmy*, *odczytywać*, *odczytywali*, *odczytywaliby*, *odczytywalibyście*, *odczytywalibyśmy*, *odczytywaliście*, *odczytywaliśmy*, *odczytywał*, *odczytywała*, *odczytywałaby*, *odczytywałabym*, *odczytywałabyś*, *odczytywałam*, *odczytywałaś*, *odczytywałby*, *odczytywałbym*, *odczytywałbyś*, *odczytywałem*, *odczytywałeś*, *odczytywało*, *odczytywałoby*, *odczytywałobym*, *odczytywałobyś*, *odczytywałom*, *odczytywałoś*, *odczytywały*, *odczytywałyby*, *odczytywałybyście*, *odczytywałybyśmy*, *odczytywałyście*, *odczytywałyśmy*, *odczytywana*, *odczytywane*, *odczytywani*, *odczytywanie*, *odczytywano*, *odczytywany*},

{*pici*, *picie*, *pić*, *pij*, *piją*, *pijąc*, *pijąca*, *pijące*, *pijący*, *pijcie*, *pije*, *pijecie*, *pijemy*, *pijesz*, *piję*, *pijmy*, *pili*, *piliby*, *pilibyście*, *pilibyśmy*, *piliście*, *piliśmy*, *pił*, *piła*, *piłaby*, *piłabym*, *piłabyś*, *piłam*, *piłaś*, *piłby*, *piłbym*, *piłbyś*, *piłem*, *piłeś*, *piło*, *piłoby*, *piłobym*, *piłobyś*, *piłom*, *piłoś*, *piły*, *piłyby*, *piłybyście*, *piłybyśmy*, *piłyście*, *piłyśmy*, *pita*, *pite*, *pito*, *pity*},

{*płyną*, *płynąc*, *płynąca*, *płynące*, *płynący*, *płynąć*, *płynął*, *płynąłby*, *płynąłbym*, *płynąłbyś*, *płynąłem*, *płynąłeś*, *płynę*, *płynęli*, *płynęliby*, *płynęlibyście*, *płynęlibyśmy*, *płynęliście*, *płynęliśmy*, *płynęła*, *płynęłaby*, *płynęłabym*, *płynęłabyś*, *płynęłam*, *płynęłaś*, *płynęło*, *płynęłoby*, *płynęły*, *płynęłyby*, *płynęłybyście*, *płynęłybyśmy*, *płynęłyście*, *płynęłyśmy*, *płynie*, *płyniecie*, *płyniemy*, *płyniesz*, *płynięcie*, *płynięto*, *płyń*, *płyńcie*, *płyńmy*},

{*pomalować*, *pomalowali*, *pomalowaliby*, *pomalowalibyście*, *pomalowalibyśmy*, *pomalowaliście*, *pomalowaliśmy*, *pomalował*, *pomalowała*, *pomalowałaby*, *pomalowałabym*, *pomalowałabyś*, *pomalowałam*, *pomalowałaś*, *pomalowałby*, *pomalowałbym*, *pomalowałbyś*, *pomalowałem*, *pomalowałeś*, *pomalowało*, *pomalowałoby*, *pomalowałobym*, *pomalowałobyś*, *pomalowałom*, *pomalowałoś*, *pomalowały*, *pomalowałyby*, *pomalowałybyście*, *pomalowałybyśmy*, *pomalowałyście*, *pomalowałyśmy*, *pomalowana*, *pomalowane*, *pomalowani*, *pomalowanie*, *pomalowano*, *pomalowany*, *pomalowawszy*, *pomaluj*, *pomalują*, *pomalujcie*, *pomaluje*, *pomalujecie*, *pomalujemy*, *pomalujesz*, *pomaluję*, *pomalujmy*},

{*robi*, *robią*, *robiąc*, *robiąca*, *robiące*, *robiący*, *robicie*, *robić*, *robieni*, *robienie*, *robię*, *robili*, *robiliby*, *robilibyście*, *robilibyśmy*, *robiliście*, *robiliśmy*, *robił*, *robiła*, *robiłaby*, *robiłabym*, *robiłabyś*, *robiłam*, *robiłaś*, *robiłby*, *robiłbym*, *robiłbyś*, *robiłem*, *robiłeś*, *robiło*, *robiłoby*, *robiłobym*, *robiłobyś*, *robiłom*, *robiłoś*, *robiły*, *robiłyby*, *robiłybyście*, *robiłybyśmy*, *robiłyście*, *robiłyśmy*, *robimy*, *robiona*, *robione*, *robiono*, *robiony*, *robisz*, *zrobi*, *zrobią*, *zrobicie*, *zrobić*, *zrobieni*, *zrobienie*, *zrobię*, *zrobili*, *zrobiliby*, *zrobilibyście*, *zrobilibyśmy*, *zrobiliście*, *zrobiliśmy*, *zrobił*, *zrobiła*, *zrobiłaby*, *zrobiłabym*, *zrobiłabyś*, *zrobiłam*, *zrobiłaś*, *zrobiłby*, *zrobiłbym*, *zrobiłbyś*, *zrobiłem*, *zrobiłeś*, *zrobiło*, *zrobiłoby*, *zrobiłobym*, *zrobiłobyś*, *zrobiłom*, *zrobiłoś*, *zrobiły*, *zrobiłyby*, *zrobiłybyście*, *zrobiłybyśmy*, *zrobiłyście*, *zrobiłyśmy*, *zrobimy*, *zrobiona*, *zrobione*, *zrobiono*, *zrobiony*, *zrobisz*, *zrobiwszy*},

{*taniali*, *taniał*, *taniała*, *taniałaby*, *taniałabym*, *taniałabyś*, *taniałam*, *taniałaś*, *taniałby*, *taniałbym*, *taniałbyś*, *taniałe*, *taniałem*, *taniałeś*, *taniało*, *taniałoby*, *taniałobym*, *taniałobyś*, *taniałom*, *taniałoś*, *taniały*, *taniałyby*, *taniałybyście*, *taniałybyśmy*, *taniałyście*, *taniałyśmy*, *taniano*, *tanieć*, *taniej*, *tanieją*, *taniejąc*, *taniejąca*, *taniejące*, *taniejący*, *taniejcie*, *tanieje*, *taniejecie*, *taniejemy*, *taniejesz*, *tanieję*, *taniejmy*, *tanieli*, *tanieliby*, *tanielibyście*, *tanielibyśmy*, *tanieliście*, *tanieliśmy*, *tanienie*},

{*umiał*, *umiała*, *umiałaby*, *umiałabym*, *umiałabyś*, *umiałam*, *umiałaś*, *umiałby*, *umiałbym*, *umiałbyś*, *umiałem*, *umiałeś*, *umiało*, *umiałoby*, *umiały*, *umiałyby*, *umiałybyście*, *umiałybyśmy*, *umiałyście*, *umiałyśmy*, *umie*, *umiecie*, *umieć*, *umiej*, *umieją*, *umiejąc*, *umiejąca*, *umiejące*, *umiejący*, *umiejcie*, *umiejmy*, *umieli*, *umieliby*, *umielibyście*, *umielibyśmy*, *umieliście*, *umieliśmy*, *umiem*, *umiemy*, *umiesz*},

{*uwierz*, *uwierzą*, *uwierzcie*, *uwierzenie*, *uwierzę*, *uwierzmy*, *uwierzono*, *uwierzy*, *uwierzycie*, *uwierzyć*, *uwierzyli*, *uwierzyliby*, *uwierzylibyście*, *uwierzylibyśmy*, *uwierzyliście*, *uwierzyliśmy*, *uwierzył*, *uwierzyła*, *uwierzyłaby*, *uwierzyłabym*, *uwierzyłabyś*, *uwierzyłam*, *uwierzyłaś*, *uwierzyłby*, *uwierzyłbym*, *uwierzyłbyś*, *uwierzyłem*, *uwierzyłeś*, *uwierzyło*, *uwierzyłoby*, *uwierzyłobym*, *uwierzyłobyś*, *uwierzyłom*, *uwierzyłoś*, *uwierzyły*, *uwierzyłyby*, *uwierzyłybyście*, *uwierzyłybyśmy*, *uwierzyłyście*, *uwierzyłyśmy*, *uwierzymy*, *uwierzysz*, *uwierzywszy*, *wierz*, *wierzą*, *wierząc*, *wierząca*, *wierzące*, *wierzący*, *wierzcie*, *wierzenie*, *wierzę*, *wierzmy*, *wierzono*, *wierzy*, *wierzycie*, *wierzyć*, *wierzyli*, *wierzyliby*, *wierzylibyście*, *wierzylibyśmy*, *wierzyliście*, *wierzyliśmy*, *wierzył*, *wierzyła*, *wierzyłaby*, *wierzyłabym*, *wierzyłabyś*, *wierzyłam*, *wierzyłaś*, *wierzyłby*, *wierzyłbym*, *wierzyłbyś*, *wierzyłem*, *wierzyłeś*, *wierzyło*, *wierzyłoby*, *wierzyłobym*, *wierzyłobyś*, *wierzyłom*, *wierzyłoś*, *wierzyły*, *wierzyłyby*, *wierzyłybyście*, *wierzyłybyśmy*, *wierzyłyście*, *wierzyłyśmy*, *wierzymy*, *wierzysz*},

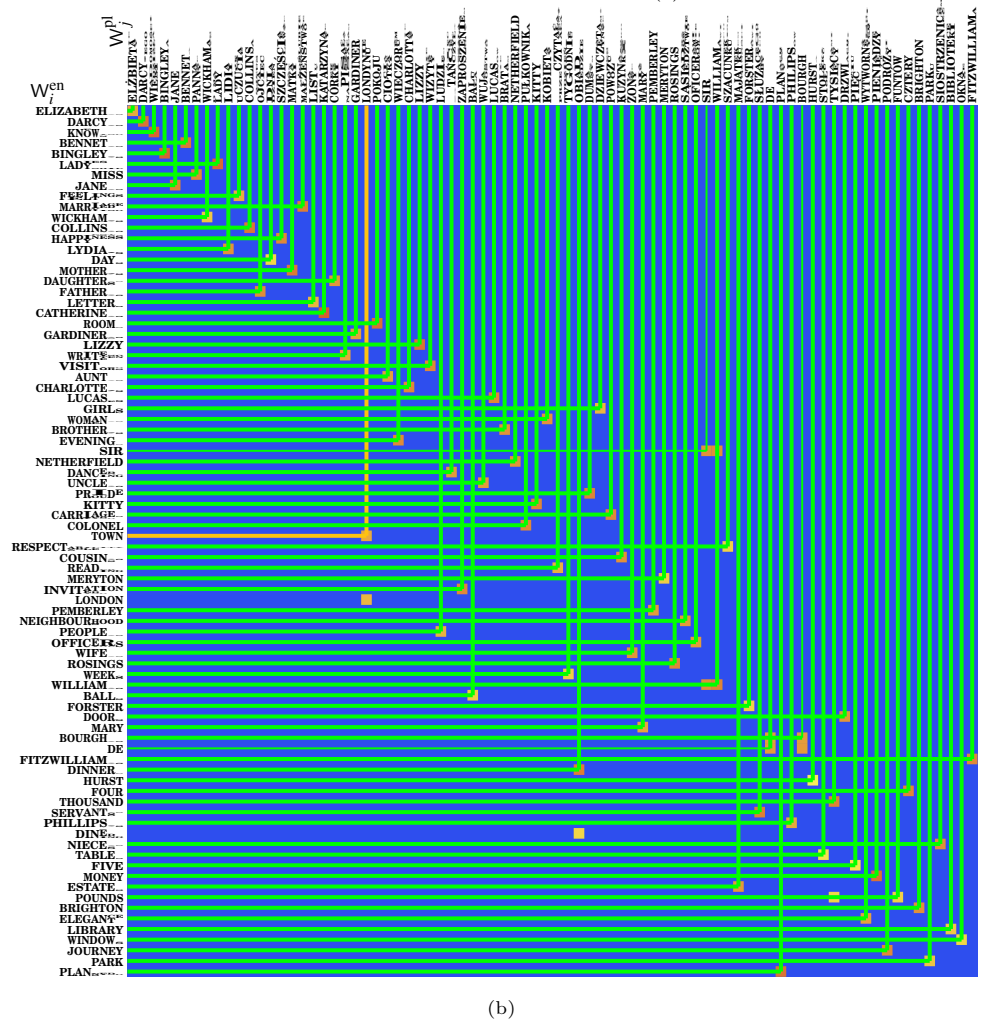
{*widzą*, *widząc*, *widząca*, *widzące*, *widzący*, *widzenie*, *widzę*, *widzi*, *widział*, *widziała*, *widziałaby*, *widziałabym*, *widziałabyś*, *widziałam*, *widziałaś*, *widziałby*, *widziałbym*, *widziałbyś*, *widziałem*, *widziałeś*, *widziało*, *widziałoby*, *widziałobym*, *widziałobyś*, *widziałom*, *widziałoś*, *widziały*, *widziałyby*, *widziałybyście*, *widziałybyśmy*, *widziałyście*, *widziałyśmy*, *widziana*, *widziane*, *widziani*, *widziano*, *widziany*, *widzicie*, *widzieć*, *widzieli*, *widzieliby*, *widzielibyście*, *widzielibyśmy*, *widzieliście*, *widzieliśmy*, *widzimy*, *widzisz*, *widź*, *widźcie*, *widźmy*, *zobacz*, *zobaczą*, *zobaczcie*, *zobaczenie*, *zobaczę*, *zobaczmy*, *zobaczono*, *zobaczy*, *zobaczycie*, *zobaczyć*, *zobaczyli*, *zobaczyliby*, *zobaczylibyście*, *zobaczylibyśmy*, *zobaczyliście*, *zobaczyliśmy*, *zobaczył*, *zobaczyła*, *zobaczyłaby*, *zobaczyłabym*, *zobaczyłabyś*, *zobaczyłam*, *zobaczyłaś*, *zobaczyłby*, *zobaczyłbym*, *zobaczyłbyś*, *zobaczyłem*, *zobaczyłeś*, *zobaczyło*, *zobaczyłoby*, *zobaczyły*, *zobaczyłyby*, *zobaczyłybyście*, *zobaczyłybyśmy*, *zobaczyłyście*, *zobaczyłyśmy*, *zobaczymy*, *zobaczysz*, *zobaczywszy*},

{*wiezie*, *wieziecie*, *wieziemy*, *wiezieni*, *wiezienie*, *wieziesz*, *wieziona*, *wiezione*, *wieziono*, *wieziony*, *wiozą*, *wiozący*, *wiozą*, *wiozące*, *wiozący*, *wiozę*, *wiozła*, *wiozłaby*, *wiozłabym*, *wiozłabyś*, *wiozłam*, *wiozłaś*, *wiozło*, *wiozłoby*, *wiozły*, *wiozłyby*, *wiozłybyście*, *wiozłybyśmy*, *wiozłyście*, *wiozłyśmy*, *wiózł*, *wiózłby*, *wiózłbym*, *wiózłbyś*, *wiózłem*, *wiózłeś*},

{*zyska*, *zyskacie*, *zyskać*, *zyskaj*, *zyskają*, *zyskajcie*, *zyskajmy*, *zyskali*, *zyskaliby*, *zyskalibyście*, *zyskalibyśmy*, *zyskaliście*, *zyskaliśmy*, *zyskał*, *zyskała*, *zyskałaby*, *zyskałabym*, *zyskałabyś*, *zyskałam*, *zyskałaś*, *zyskałby*, *zyskałbym*, *zyskałbyś*, *zyskałem*, *zyskałeś*, *zyskało*, *zyskałoby*, *zyskałobym*, *zyskałobyś*, *zyskałom*, *zyskałoś*, *zyskały*, *zyskałyby*, *zyskałybyście*, *zyskałybyśmy*, *zyskałyście*, *zyskałyśmy*, *zyskam*, *zyskamy*, *zyskana*, *zyskane*, *zyskani*, *zyskanie*, *zyskano*, *zyskany*, *zyskasz*, *zyskawszy*, *zyskiwać*, *zyskiwali*, *zyskiwaliby*, *zyskiwalibyście*, *zyskiwalibyśmy*, *zyskiwaliście*, *zyskiwaliśmy*, *zyskiwał*, *zyskiwała*, *zyskiwałaby*, *zyskiwałabym*, *zyskiwałabyś*, *zyskiwałam*, *zyskiwałaś*, *zyskiwałby*, *zyskiwałbym*, *zyskiwałbyś*, *zyskiwałem*, *zyskiwałeś*, *zyskiwało*, *zyskiwałoby*, *zyskiwały*, *zyskiwałyby*, *zyskiwałybyście*, *zyskiwałybyśmy*, *zyskiwałyście*, *zyskiwałyśmy*, *zyskiwana*, *zyskiwane*, *zyskiwani*, *zyskiwanie*, *zyskiwano*, *zyskiwany*, *zyskuj*, *zyskują*, *zyskując*, *zyskująca*, *zyskujące*, *zyskujący*, *zyskujcie*, *zyskuje*, *zyskujecie*, *zyskujemy*, *zyskujesz*, *zyskuję*, *zyskujmy*}.

Example 7.13.2. In Fig. S10, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text sources).

(a)



(b)

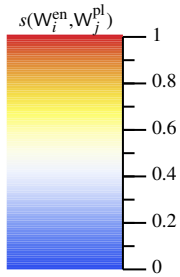
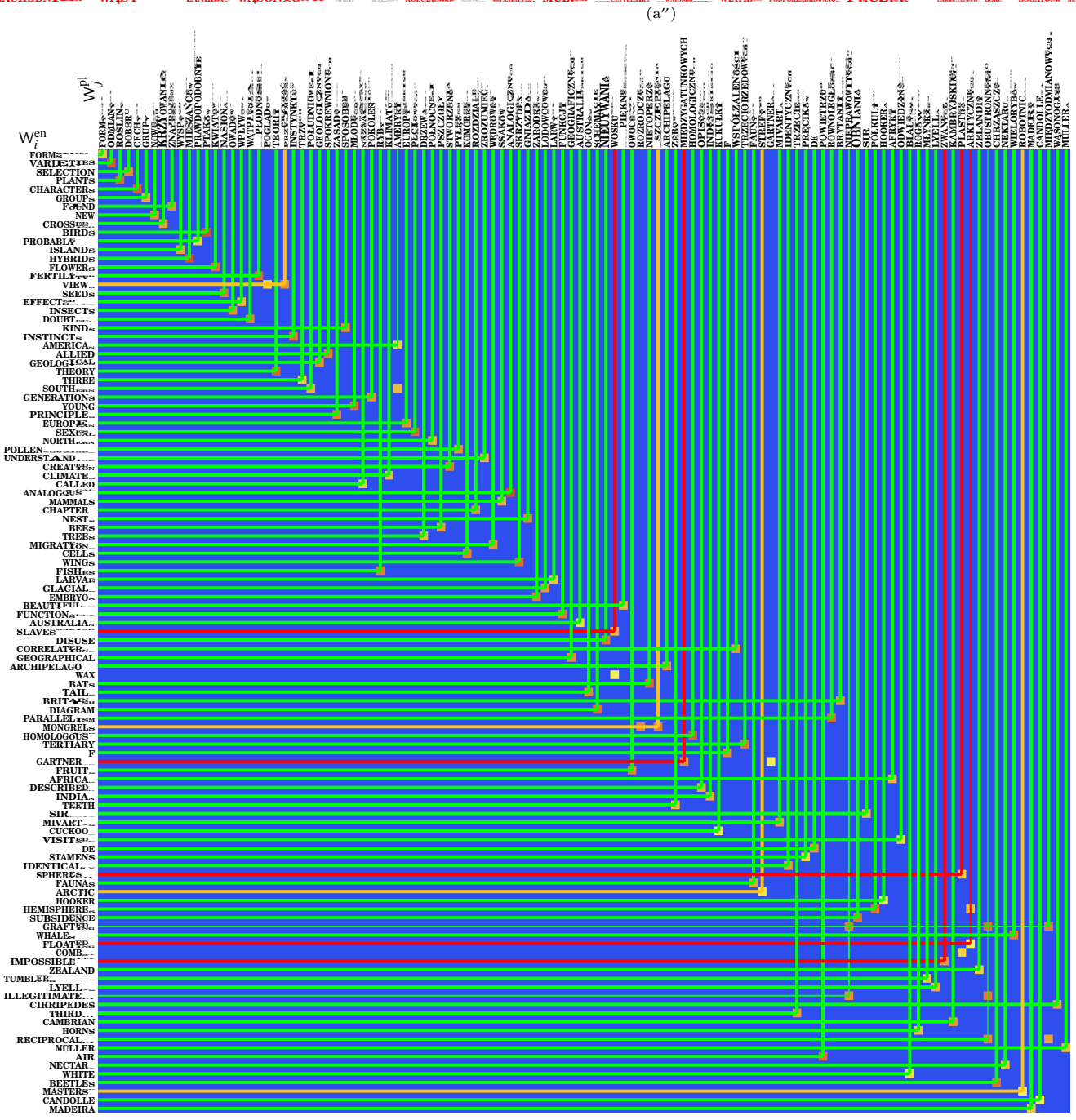
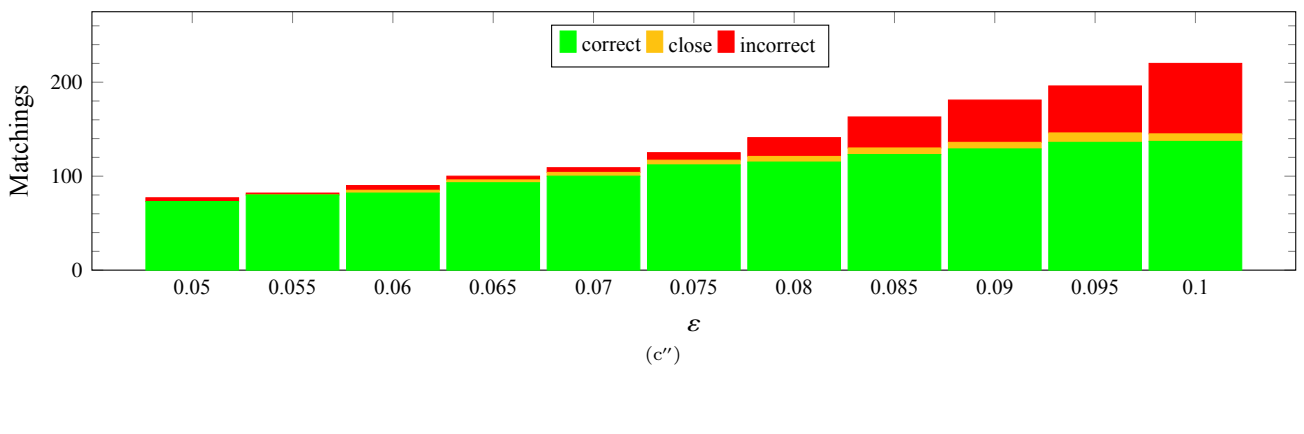


Fig. S10. Text mining in Polish. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Polish version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{pl})$ between selected topics in English and Polish versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *amber*) cross-hair indicates an exact (resp. a close but non-exact) match.

(a″)



(b″)



(c″)

Fig. S10. Text mining in Polish. (*Continued*) (a″) Statistically identified topics ($n_{ii} \geq 20$) in a Polish version of *Origin of Species*, with the same color encoding scheme as Fig. S3. (b″) Semantic similarities $s(W_i^{en}, W_j^{pl})$ between selected topics in English and Polish versions of *Origin of Species*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c″) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

7.2 Modified Porter stemming algorithm for Russian

Our analysis mainly centers on Russian documents in the modern (post-1918) orthography.[90] We also assume that the Russian documents under investigation do not carry stress marks (which only appear in pedagogical contexts) on individual words. We remove the diacritical mark from *ё* in the effective spelling.

In modern Russian, the following present tense conjugations for the Russian verb "to be"

еси, *есмы*, *есмь*, *есте*, *есть*, *суть*

are no longer used as their English counterparts, but they are found in works of 19th-century Russian authors (like Dostoevsky). The word *есть* is still used modern constructions of possessions (such as *У меня есть* ... "I have ..."). We are going to include all the inflected forms of the Russian verb "to be" in our list of stop words.

Definition 7.14 (Russian stop words)**.** If a word belongs to the following list[91]:

а, *б*, *без*, *более*, *больше*, *будем*, *будет*, *будете*, *будешь*, *будто*, *буду*, *будут*, *будучи*, *будь*, *будьте*, *бы*, *был*, *бывав*, *бывавши*, *бывавший*, *бываем*, *бывает*, *бываете*, *бываешь*, *бывай*, *бывайте*, *бывал*, *бывала*, *бывали*, *бывало*, *бывать*, *бываю*, *бывают*, *бывающий*, *бывая*, *бывшая*, *бывшего*, *бывшее*, *бывшей*, *бывшем*, *бывшему*, *бывшею*, *бывши*, *бывшие*, *бывший*, *бывшим*, *бывших*, *бывшую*, *был*, *была*, *были*, *было*, *быть*, *в*, *вам*, *вами*, *вас*, *ваш*, *ваша*, *ваше*, *вашего*, *вашей*, *вашем*, *вашему*, *вашею*, *ваши*, *вашим*, *вашими*, *ваших*, *вашу*, *вдоль*, *вдруг*, *ведь*, *весь*, *весьма*, *взад*, *вместе*, *вместо*, *вне*, *внутри*, *во*, *вовне*, *вовсе*, *возле*, *вокруг*, *вон*, *вопреки*, *вот*, *впереди*, *впрочем*, *все*, *всегда*, *всего*, *всей*, *всем*, *всеми*, *всему*, *всех*, *всею*, *всё*, *вскоре*, *вслед*, *всю*, *вся*, *всяк*, *всяка*, *всякая*, *всяки*, *всякие*, *всякий*, *всяким*, *всякими*, *всяких*, *всяко*, *всякого*, *всякое*, *всякой*, *всяком*, *всякому*, *всякою*, *всякую*, *вы*, *где*, *да*, *давай*, *давайте*, *даже*, *дай*, *дайте*, *делав*, *делавшая*, *делавшего*, *делавшее*, *делавшей*, *делавшем*, *делавшему*, *делавшею*, *делавши*, *делавшие*, *делавший*, *делавшим*, *делавшими*, *делавших*, *делавшую*, *делаем*, *делаема*, *делаемая*, *делаемо*, *делаемого*, *делаемое*, *делаемой*, *делаемом*, *делаемому*, *делаемою*, *делаемую*, *делаемы*, *делаемые*, *делаемый*, *делаемым*, *делаемыми*, *делаемых*, *делает*, *делаете*, *делаешь*, *делай*, *делайте*, *делал*, *делала*, *делали*, *делало*, *делан*, *делана*, *деланная*, *деланного*, *деланное*, *деланной*, *деланном*, *деланному*, *деланною*, *деланную*, *деланные*, *деланный*, *деланным*, *деланными*, *деланных*, *делано*, *деланы*, *делать*, *делаю*, *делают*, *делающая*, *делающего*, *делающее*, *делающей*, *делающем*, *делающему*, *делающею*, *делающие*, *делающий*, *делающим*, *делающими*, *делающих*, *делающую*, *делая*, *для*, *до*, *должен*, *должна*, *должно*, *должны*, *другая*, *другие*, *другим*, *другими*, *других*, *другого*, *другое*, *другой*, *другом*, *другому*, *другою*, *другую*, *е*, *его*, *едва*, *ее*, *её*, *ей*, *еле*, *ему*, *еси*, *если*, *есмы*, *есмь*, *есте*, *есть*, *еще*, *ещё*, *ею*, *ж*, *же*, *за*, *затем*, *зато*, *зачем*, *здесь*, *и*, *ибо*, *из*, *изо*, *или*, *им*, *именно*, *ими*, *иначе*, *иная*, *иногда*, *иного*, *иное*, *иной*, *ином*, *иному*, *иною*, *иную*, *иные*, *иным*, *иными*, *иных*, *их*, *к*, *ка*, *каждая*, *каждого*, *каждое*, *каждой*, *каждом*, *каждому*, *каждою*, *каждую*, *каждые*, *каждый*, *каждым*, *каждыми*, *каждых*, *как*, *какая*, *какие*, *каким*, *какими*, *каких*, *каков*, *какова*, *каковая*, *каково*, *какового*, *каковое*, *каковой*, *каковом*, *каковому*, *каковою*, *каковую*, *каковы*, *каковые*, *каковым*, *каковыми*, *каковых*, *какого*, *какое*, *какой*, *каком*, *какому*, *какою*, *какую*, *кем*, *ко*, *когда*, *кого*, *кое*, *коего*, *коей*, *коем*, *коему*, *коею*, *кои*, *коим*, *коими*, *коих*, *кой*, *ком*, *кому*, *конечно*, *которая*, *которого*, *которое*, *которой*, *котором*, *которому*, *которою*, *которую*, *которые*, *который*, *которым*, *которыми*, *которых*, *кою*, *коя*, *кроме*, *кто*, *куда*, *ли*, *либо*, *лишь*, *любая*, *любого*, *любое*, *любой*, *любом*, *любому*, *любою*, *любую*, *любые*, *любым*, *любыми*, *любых*, *м*, *между*, *менее*, *меня*, *мне*, *мног*, *многа*, *многая*, *многи*, *многие*, *многий*, *многим*, *многими*, *многих*, *много*, *многого*, *многое*, *многой*, *многом*, *многому*, *многою*, *многую*, *мной*, *мною*, *мог*, *моги*, *могите*, *могла*, *могли*, *могло*, *могу*, *могут*, *могущая*, *могущего*, *могущее*, *могущей*, *могущем*, *могущему*, *могущею*, *могущие*, *могущий*, *могущим*, *могущими*, *могущих*, *могущую*, *могшая*, *могшего*, *могшее*, *могшей*, *могшем*, *могшему*, *могшею*, *могшие*, *могший*, *могшим*, *могшими*, *могших*, *могшую*, *мое*, *моего*, *моей*, *моем*, *моему*, *моею*, *моё*, *можем*, *может*, *можете*, *можешь*, *можно*, *мои*, *моим*, *моими*, *моих*, *мой*, *мою*, *моя*, *мы*, *на*, *навстречу*, *над*, *надо*, *назад*, *наиболее*, *наконец*, *нам*, *нами*, *напролет*, *напролёт*, *напротив*, *нас*, *насколько*, *настолько*, *наш*, *наша*, *наше*, *нашего*, *нашей*, *нашем*, *нашему*, *нашею*, *наши*, *нашим*, *нашими*, *наших*, *нашу*, *не*, *негде*, *него*, *нее*, *неё*, *ней*, *некем*, *некогда*, *некого*, *некому*, *некоторая*, *некоторого*, *некоторое*, *некоторой*, *некотором*, *некоторому*, *некоторою*, *некоторую*, *некоторые*, *некоторый*, *некоторым*, *некоторыми*, *некоторых*, *некуда*, *нельзя*, *нем*, *немног*, *немнога*, *немногая*, *немноги*, *немногие*, *немногий*, *немногим*, *немногими*, *немногих*, *немного*, *немногого*, *немногое*, *немногой*, *немногом*, *немногому*, *немногою*, *немногую*, *нему*, *нередко*, *нескольким*, *несколькими*, *нескольких*, *несколько*, *несмотря*, *нет*, *неужели*, *нечего*, *нечем*, *нечему*, *нею*, *нём*,

[90] When 19th-century Russian literature is printed and read today, the spellings therein are thoroughly modernized.

[91] Our list of Russian stop words is based on `http://snowball.tartarus.org/algorithms/russian/stop.txt`, with extensive modifications to roughly match their counterparts in English. In particular, we have included all the inflected forms of *быть* "to be", *каждый* "every", *какой* "what kind", *кто* "who", *чей* "whose", *что* "what", and all the possessive pronouns declined in six cases and three genders. Some archaic, poetic or dialectal forms are also included.

ни, *нибудь*, *нигде*, *никак*, *никакая*, *никакие*, *никаким*, *никакими*, *никаких*, *никакого*, *никакое*, *никакой*, *никаком*, *никакому*, *никакою*, *никакую*, *никем*, *никогда*, *никого*, *никому*, *никто*, *никуда*, *ним*, *ними*, *ниоткуда*, *нисколько*, *них*, *ничего*, *ничей*, *ничем*, *ничему*, *ничто*, *ничьего*, *ничьей*, *ничьему*, *ничьею*, *ничьё*, *ничьём*, *ничьи*, *ничьим*, *ничьими*, *ничьих*, *ничью*, *ничья*, *но*, *ну*, *ныне*, *нэи*, *о*, *об*, *оба*, *обе*, *обеим*, *обеими*, *обеих*, *обо*, *обоего*, *обоим*, *обоими*, *обоих*, *один*, *одна*, *однако*, *одни*, *одним*, *одними*, *одних*, *одно*, *одного*, *одной*, *одном*, *одному*, *одною*, *одну*, *около*, *он*, *она*, *они*, *оно*, *опять*, *от*, *откуда*, *отнюдь*, *ото*, *отсюда*, *оттого*, *оттуда*, *отчего*, *очень*, *перед*, *по*, *под*, *подле*, *пожалуй*, *позади*, *пока*, *поныне*, *поскольку*, *после*, *посреди*, *потом*, *потому*, *почем*, *почему*, *почём*, *почти*, *поэтому*, *прежде*, *при*, *притом*, *про*, *против*, *пускай*, *пусть*, *раз*, *разве*, *с*, *сам*, *сама*, *самая*, *сами*, *самим*, *самими*, *самих*, *само*, *самого*, *самое*, *самоё*, *самой*, *самом*, *самому*, *самою*, *саму*, *самую*, *самые*, *самый*, *самым*, *самыми*, *самых*, *свое*, *своего*, *своей*, *своем*, *своему*, *своею*, *своё*, *свои*, *своим*, *своими*, *своих*, *свой*, *свою*, *своя*, *сделав*, *сделавшая*, *сделавшего*, *сделавшее*, *сделавшей*, *сделавшем*, *сделавшему*, *сделавшею*, *сделавши*, *сделавшие*, *сделавший*, *сделавшим*, *сделавшими*, *сделавших*, *сделавшую*, *сделаем*, *сделает*, *сделаете*, *сделаешь*, *сделай*, *сделайте*, *сделал*, *сделала*, *сделали*, *сделало*, *сделан*, *сделана*, *сделанная*, *сделанного*, *сделанное*, *сделанной*, *сделанном*, *сделанному*, *сделанною*, *сделанную*, *сделанные*, *сделанный*, *сделанным*, *сделанными*, *сделанных*, *сделано*, *сделаны*, *сделать*, *сделаю*, *сделают*, *се*, *себе*, *себя*, *сего*, *сей*, *сейчас*, *сем*, *сему*, *сею*, *сё*, *сём*, *сие*, *сии*, *сим*, *сими*, *сих*, *сию*, *сия*, *сквозь*, *сколь*, *скольким*, *сколькими*, *скольких*, *сколько*, *скоро*, *слишком*, *словно*, *снова*, *со*, *собой*, *собою*, *совсем*, *согласно*, *спустя*, *среди*, *столь*, *стольким*, *столькими*, *стольких*, *столько*, *суть*, *сущая*, *сущего*, *сущее*, *сущей*, *сущем*, *сущему*, *сущею*, *сущие*, *сущий*, *сущим*, *сущими*, *сущих*, *сущую*, *сюда*, *та*, *так*, *такая*, *также*, *таки*, *такие*, *таким*, *такими*, *таких*, *таков*, *такова*, *таково*, *таковы*, *такого*, *такое*, *такой*, *таком*, *такому*, *такою*, *такую*, *там*, *твое*, *твоего*, *твоей*, *твоем*, *твоему*, *твоею*, *твоё*, *твои*, *твоим*, *твоими*, *твоих*, *твой*, *твою*, *твоя*, *те*, *тебе*, *тебя*, *тем*, *теми*, *теперь*, *тех*, *то*, *тобой*, *тобою*, *тогда*, *того*, *тоже*, *той*, *только*, *том*, *тому*, *тот*, *точно*, *тою*, *три*, *ту*, *туда*, *тут*, *ты*, *у*, *уж*, *уже*, *хоть*, *хотя*, *часто*, *чего*, *чей*, *чем*, *чему*, *через*, *чём*, *что*, *чтоб*, *чтобы*, *чуть*, *чье*, *чьего*, *чьей*, *чьем*, *чьему*, *чьею*, *чьё*, *чьём*, *чьи*, *чьим*, *чьими*, *чьих*, *чью*, *чья*, *эи*, *эта*, *эти*, *этим*, *этими*, *этих*, *это*, *этого*, *этой*, *этом*, *этому*, *этот*, *этою*, *эту*, *эты*, *я*,

then we consider it a Russian stop word. All the Russian stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

The following Russian verb has highly irregular conjugations:

дав, *давав*, *дававшая*, *дававшего*, *дававшее*, *дававшей*, *дававшем*, *дававшему*, *дававшею*, *дававши*, *дававшие*, *дававший*, *дававшим*, *дававшими*, *дававших*, *дававшую*, *даваем*, *даваема*, *даваемая*, *даваемо*, *даваемого*, *даваемое*, *даваемой*, *даваемом*, *даваемому*, *даваемою*, *даваемую*, *даваемы*, *даваемые*, *даваемый*, *даваемым*, *даваемыми*, *даваемых*, *давай*, *давайте*, *давал*, *давала*, *давали*, *давало*, *давать*, *давая*, *давшая*, *давшего*, *давшее*, *давшей*, *давшем*, *давшему*, *давшею*, *давши*, *давшие*, *давший*, *давшим*, *давшими*, *давших*, *давшую*, *дадим*, *дадите*, *дадут*, *даем*, *дает*, *даете*, *даешь*, *даём*, *даёт*, *даёте*, *даёшь*, *дай*, *дайте*, *дал*, *дала*, *дали*, *дало*, *дам*, *дан*, *дана*, *данная*, *данного*, *данное*, *данной*, *данном*, *данному*, *данною*, *данную*, *данные*, *данный*, *данным*, *данными*, *данных*, *дано*, *даны*, *даст*, *дать*, *дашь*, *даю*, *дают*, *дающая*, *дающего*, *дающее*, *дающей*, *дающем*, *дающему*, *дающею*, *дающие*, *дающий*, *дающим*, *дающими*, *дающих*, *дающую* — “give”.

We are not going to include these conjugated forms as stop words, so as to be consistent with other languages under our consideration. Nevertheless, we will define a string pattern **giveRussian** using the list above, to facilitate the clustering of Russian content words.

For sorting purposes (*Mathematica* v11.0 does not seem to support multi-level sorting of Russian words), we will need to transliterate the Russian effective spelling and essential root into the Latin script. The transliteration is not meant to be phonetically accurate, but only serves as a convenient one-to-one mapping. The “alphabetic order” in our clustering algorithm for Russian words will refer to the Latin transliterations, rather than the native order according to the Russian alphabet.

Algorithm 7.15 (Transliteration of Russian Letters)**.** *A text string $\hat{\sigma}$ derived from a Russian word is transliterated into Latin-based script through the following replacements:*

а	*б*	*в*	*г*	*д*	*е*	*ж*	*з*	*и*	*й*	*к*	*л*	*м*	*н*	*о*	*п*	*р*	*с*	*т*	*у*	*ф*	*х*	*ц*	*ч*	*ш*	*щ*	*ъ*	*ы*	*ь*	*э*	*ю*	*я*
a	*b*	*v*	*g*	*d*	*e*	*j*	*z*	*i*	*ï*	*k*	*l*	*m*	*n*	*o*	*p*	*r*	*s*	*t*	*u*	*f*	*h*	*c*	*q*	*w*	*W*	*æ*	*y*	*œ*	*é*	*û*	*â*

The result of such a transliteration is denoted by RuLat($\hat{\sigma}$).

7.2.1 Effective spelling and essential root

It is assumed that all Russian words are converted to lowercase[92] before going through any of the procedures below.

Definition 7.16 (Russian Effective Vowels and Verb Prefixes)**.** Hereafter in §7.2, the symbol $\mathbf{V}^*$ stands for any member from the list {*а*, *е*, *и*, *о*, *у*, *ы*, *ь*, *ю*, *я*}, the so-called Russian effective vowels.[93] In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}^*_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Russian vowel extensions.

Dual to the notations above, the symbol $\mathbf{C}^*$ stands for any character that does not belong to the list {*а*, *е*, *и*, *о*, *у*, *ы*, *ь*, *ю*, *я*}, and $\mathbf{C}^*_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list.

The symbol $\mathbf{P} = (\varnothing|в|вз|во|воз|вс|вы|до|за|из|ис|на|о|об|обо|от|ото|пере|по|под|подо|при|раз|разо|рас|с|со|у)$ represents the possible prefix of a Russian verb.

We further define $\mathbf{P}^* = (взъ|возъ|всъ|въ|изъ|исъ|объ|отъ|подъ|разъ|расъ|съ|\mathbf{P})$ to account for a possible insertion of the Russian hard sign ъ after a prefix ending in a consonant. □

Algorithm 7.17 (Russian effective spelling)**.** *For a Russian word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in sequential steps:*

(1) *If* $\hat{\sigma}$ *contains* (*а*|*е*|*ё*|*и*|*о*|*у*|*ы*|*ь*|*э*|*ю*|*я*)*, then leave it as is; otherwise, replace it by* $\xi\hat{\sigma}\xi$*.*

(2) *Replace*[94]

(∅|по)сид~		(∅|у)с(ев|евш|ел|ял)~				$\mathbf{X}^{\in}$(∅|по|у)чт~		
σыт		σыт				**X**чест		
гр$\mathbf{X}^{\in}$(а|е|и|о)$\hat{\chi}$~	жел~	завтрак~	зву(к|ч)~	изум~	исп$\mathbf{X}^{\in}$(о|ы)~	каза~	камин~	кокет~
γρ**X**$\hat{\chi}'$	жел	βρεκφστ	σοωṽδ	изуμ	**X**σξπ	кажущий	χεαρθ	qокет
колен~	корач	коро(тк|ток|ч)		крат(к|ок|ч)	кри~	кро(т|тк|ток|щ)		
колṽен	σηορτ	σηορτ		σηορτ	κρρι	μылδ		
кров(∅|е)л~	кроват~	крыл~	л(е|ё)г(∅|к|ч)$\hat{\chi}$		од(∅|и)н~			
ροωφ	ββεδδ	κρылλ	лайγτ$\hat{\chi}$		*1*			
остан(а|о)в~	отча~	рав(∅|е)н	разговар~	разум~	са(ди|дя|жу)~	солнц~		
остаṽов	отча	ραβн	говор	ραζуμ	σыти	σолε		
ств(ова|у)**X**	стен~	сторон~	т(е|ё|ь)м~	тарел~	уш(а|е|ек|и|к|н)**X**~	хо(т|ч)		
сть	ωаλλ	σайδн	τεμδк	тарел	уωр	χот		
холл~	холм~	чис(∅|е)л	<u>денни</u>	<u>сент</u>	<u>сесть</u>	<u>стара(∅|я)</u>	<u>ух(а|е|о|ом|у)</u>	
χолλ	холμ	ξисл	δеṽṽи	σанкт	σыт	старый	уωр	

where $\hat{\chi}'$ *results from doing* $в \to β$, $м \to μ$, $н \to ṽ$, $ц \to τσ$, $ч \to ξ$ *on* $\hat{\chi}$.

(3) *Replace*

(∅|по)стара~		(еже|каждо)дневн**X**~		γраṽи(ц|ч)~		б(а|о)лт		внима~		
τρηа		день		гранич		болт		ттṽа		
военн~	вол	дам**V**~	женат~	завтра(∅|шн)~		кров~	март~	миллион		павлин
войн	ββолλ	δδαμ**V**	μар	τμορω		кροβ	μρατσ	мξмон		πикк
пал(е|ь)ц	план~	погод	преданн~	предел~	пред~	пт(а|и)**X**~	сад~	свад~	скал	
φиṽγ	плаṽ	ωеθ	πрат	λлиμ	πр	βирδ	γарδ	μар	рокη	
соба(к|ч)	специал		страни	стрем	сударын	сумм~	тон(∅|о)(к|ч|ьш)~			
δоγ	спеξл		φремд	риζ	мадам	σумμ	θиṽ			
убе(∅|ж)д	угод~	улиц	хва(т|ч)~	част$\mathbf{X}^{\in}$(е|и|н|ь|я)~			южX~	<u>леди</u>	<u>па</u>	<u>увы</u>
βиṽк	гṽод	τрит	χβат	част**X**			юг	лледδ	ππасσ	αлασ

[92] As of v11.0, *Mathematica* does not produce desired case conversions of Russian letters upon invoking the `ToLowerCase` or `ToUpperCase` operation. Therefore, case conversions for all the Russian letters need to be hand-coded if one implements our algorithm in *Mathematica*.

[93] We note that although э is indeed a vowel letter, it never appears in the ending of a Russian word. The soft sign ь was historically a short vowel, and it still occurs often in modern Russian word endings.

[94] Words like *кровать* "bed" and *кровли* "roofs" appear to carry commonly seen verb suffixes *-ать* and *-ли*, but are actually not related to *кровь* "blood". We have created exceptional effective spellings to separate these words (and their inflections) into three classes. However, we note that our algorithm does not distinguish *кров* "shelter" (and its inflections) from any of the inflected forms of *кровь*. Furthermore, *кровь* "blood", *кровоточить* "to bleed" and *кровотечение* "bleeding" are split into three different clusters by our current method.

(4) *Replace*

(∅|вз)волн~	(∅|вы)т(е|ё)(к|ч)(∅|ш)~	(∅|по)жен$^{\mathbf{X}\in}$(и|ю|я)~	(∅|об|с)новŷ~	(∅|по)счаст~
ωаβ	τек	μар	ṽовŷ	σчаст

(∅|на|по|про)чит~	(∅|по)чувств	(в|воз|по)люб~	(по|про)яв~	**P***giveRussian(∅|сь|ся)~	**P***ник
čит	φул	люб	яв	**P***ρгывр	**P***νиγ

бенн	бо(к|ч)	брак(о|у)ŷ	вез	велик~	вид	влия~	втор~	выжима~	выжм~	гад	горьк~	гост~	гул
βеṽṽ	σиоб	рефσŷ	γвар	болеш	βиб	ĩṽфля	2о	выжа	выжа	γаб	γорьк	γост	γуллл

дал(ек|ёк|ьн)	дв(а|е|у(м|мя|х))	девуш(∅|е)к~	девят~	десят	ди(к|ч)~	див$_{\hat{y}\notin}$(ш)	добр~	долин~	дол~
φар	2е2	дев	9я	10е	дик	биβŷ	довр	долин	доп

дум	жизн~	заб	замуж	зл~	зна(л|н|ю)~	июл~	июн~	карет	кла(в|л|ст)(∅|ш)~	клон	красн~
бум	жит	ζаб	мар	ζл	зна	γюл	γюṽ	γарет	клад	γлоṽ	ργеб

кон(∅|е)ц~	лид~	мил~	мир~	молв~	мягк~	недел~	ноч~	обед	ок**V***(∅|м)~
коṽζ	лид	мил	мир	мов	мягч	ωек	ṽохт	овδ	очи

отве(т|ч)~	отлив~	п(е(в|вш|л|т)|о(ем|ет|ешь|ём|ёт|ёшь|ю|ют|ющ))
аṽσω	отлив	песн

парк~	перв~	пес(∅|е)н~	пеш	письменн~	подлин~	подол~
парк	1е	песн	пеξ	писат	подлиṽ	подол

поз$^{\mathbf{X}\in}$(в|о)~	пойм~	покл~	поко	полков~	поч(е|ё)(∅|с)т~	поч**V***~	пр**V**$^{*}_{m}$м
з**X**	поня	покл	ρум	полков	ηоṽр	ч**V***	пр**V**$^{*}_{m}$μ

пр**V**$^{*}_{m}$т	пр$^{\mathbf{X}\in}$(и|о)лив~	приним~	приятел~	прол(е|ё)~	прони(к|ц)~
пр**V**$^{*}_{m}$τ	пр**X**лиβ	приня	приятел	ле	проṽик

разлу(к|ч)~	свет~	сл(у|ы)ш	смел~	собы~	сц~	счит	тан(∅|е)ц~	титул~
σпе	σвет	лыσ	σмел	σобη	σζ	цчит	танζ	титул

треб	трет~	трон~	уи~	цел	человек	шарк	янв~	<u>боже</u>
треβ	3и	трог	Ωи	ξел	люд	šарγ	jяṽ	бог

<u>брат(∅|а|е|ом|у|ьев|ья|ьям|ьями|ьях)</u>	<u>лет(∅|а(∅|м|ми|х))</u>
βбратβ	год

<u>мыл(**V***|∅)</u>	<u>окон</u>	<u>писем</u>	<u>полях</u>	<u>сем(и|~ь)</u>	<u>сн(ах|ов|у|ы)</u>
мыть	окно	письмо	поля	7е	сон

(5) *Replace*

(по|с)ложˉ	**P***еду~	**P***(буд|бы)	**P***(йд|йт)~	**P***ле(т|ч)~	брак~	звол	знатн~	зна(к|ч)~
клад	**P***еха	**P***βе	**P***ид	флет	мар	алω	ṽовл	μнаṽ

мер(и)~	плать~	пож~	пре~	прив~	прил~	сем(ей|ь)~	чет	<u>**P***еде(м|т|те|шь)</u>
μмер	бплать	ж	πρε	β	л	φам	кеп	**P***еха

(6) *Replace*

<u>(дочь|доч(∅|а|е)к)</u>	(∅|наи)(высок|высш)~	(∅|наи)(худш|хуж)~
дочер	ввысок	плох

(∅|наи)младш~	(∅|наи)низш~	(∅|наи|по)лучш~	(∅|с)дел~	(тет|тёт)(о|уш|уше)к~
ммолод	низк	хорош	бдел	тетя

C$^{*}_{m}$ост~	бал~	бл~	глян	де(й|я)~	джо~	дом~	друз~	дяд(е|ь|юш|юше)к~
C$^{*}_{m}$озт	ббал	жжбл	гляд	бгдея	сджо	бомчм	друг	дядя

е(з|х)~	жб	забет	игр~	исьм	крут~	ле~	м**V***л~	мин~	наде~
съе	ж	за	ыгр	исзз	круч	лле	мм**V***л	мчин	ннаде

наиχ̂$_{m}$ш**X**	нест	обр~	па~	пол$\hat{\chi}\in$(а|е|и|о|ы|ь|ю|я)~	полн~	пра	раст	род~	ска(ж|з)
χ̂$_{m}$ш	нес	ччобр	ппа	полχ̂	бполн	бпра	рос	ррод	говор

слав~	*слов~*	*собств*	*спор~*	*стол~*	*тз*	*юнош~*	*даст*	*де*	*дорого*	*отцах*
зслав	*цслав*	*зопств*	*цспор*	*цстол*	*цц*	*юност*	*дал*	*дпде*	*дорог*	*отец*

пап($\mathbf{V}^*$|∅)(∅|*й*|*ми*|*х*)	*пони*	(*мать*|*мамоч*(∅|*а*|*е*)*к*~|*мам*($\mathbf{V}^*$|∅)(∅|*й*|*ми*|*х*))
отец	*ппони*	*матер*

(7) *Replace*

(∅|*по*|*с*|*со*)*жм~*	$\mathbf{C}^*_{m_0}$*гн~*	*да~*	*е*(*е*|*й*)*к*	*зв*	*особ~*	*певши*	*пон~*	*риск~*	*сл~*
жал	$\mathbf{C}^*_{m_0}$*гюн*	*чдача*	*ееч*	*зов*	*ззоб*	*по*	*н*	*рисч*	*жсл*

соб~	$^{\mathbf{X}_1}$(∅|*вы*|*по*|*у*)*м*$^{\mathbf{X}_2\in}$(*о*|*ы*)~	*день*	*лоб*	*мягок*	*рот*	*училищ***X**	(*живш*|~*жить*)	*~ести*	*~ец*
зоб	$\mathbf{X}_1$*мюж*$\mathbf{X}_2$	*дни*	*лба*	*мягч*	*рта*	*скул*	*жил*	*ел*	*~ьца*

and call the result $\hat{\sigma}'$.

(8) *Break down* $\hat{\sigma}' = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}'^{[\min\{\lambda(\hat{\sigma}'),\ell(\hat{\sigma}')\}]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where* $\lambda(\hat{\sigma}') - 1$ *equals either the last position occupied by the string pattern* (**P***|*без*|*безо*|*бес*|*не*|*ппа*|*стр*)$\hat{\chi}$~ *in the non-void* $\hat{\sigma}'$*, or* -1 *if* $\hat{\sigma}' = \varnothing$.

(9) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(4.1) *Do* (*ак*|*вич*\|*лив*|*чек*|*чив*|*чик*|*як*|(*тел*|*чк*)**X**) → ∅, (*ар*|*ач*|*ник*|*ниц*|*ств*|*яр*) → ∅, $^{\mathbf{X}\in}$(*е*|*ь*)*б* → **X**, (*кац*|(∅|*ц*)*ир*(*ов*|*у*)) → ∅, ~(*сь*|*ся*|*ться*) → ∅, ~*мя* → *мен*, ~*ш***V***(*е*|*и*|*й*\|*м*|*х*|*ю*|*я*) → *ш*.

(4.2) *Do* ~((*а*|*о*|*я*)*м*|(*а*|*и*|*ы*|*я*)*х*|**V****ев*|(*е*|*о*)*го*|(*е*|*о*)*му*|*й*|(*й*)$_{m_0}$*ме*\|*ми*|*ов*|**V****т*(∅|*ь*)|*шь*) → ∅, (*ас*|*а*|*ис*|*ос*)*т* → ∅, ~(*ец*|*к*) → ∅, (*к*|*ц*)**V***(∅|*х*) → **V***, ~(**V****т*|(*й*)$_{m_0}$*ме*) → ∅.

The result after these two steps of operations is called $\hat{\sigma}'_2$.

(10) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(11) *Do* (*ё*|*Ё*) → *е, ммальч***V*****X**~ → *ммальч*.

Definition 7.18 (Russian protected range)**.** Let $\hat{\sigma}$ be a text string derived from a Russian word, its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \min\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is determined by two non-negative integers $\lambda_1(\hat{\sigma})$ and $\lambda_2(\hat{\sigma})$ specified through the following procedures:

- Look for the string pattern (*вы*|*на*|*у*)$_{m_0}$**C**$^*_{m_0}$**V***(*й*)$_{m_0}$~ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Look for the pattern (∅|*вз*|*вы*|*на*|*по*|*про*|*раз*|*рас*|*с*)**V**$^*_{m_0}$**C*****V**$^*_{m_0}$**C***~ in the string $\hat{\sigma}$;
- If the pattern above is found, the last position occupied by such a pattern defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = \ell(\hat{\sigma})$. □

Algorithm 7.19 (Russian essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Russian word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Break down* $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ *is equal to the protected range of* $\hat{\sigma}$.

(2) *Do* *ь* → *е on* $\hat{\sigma}_1$*, and call the result* $\hat{\sigma}'_1$.

(3) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(3.1) *Do* ~**V**$^*_m$ → ∅*;*

(3.2) *Do* ~**V**$^*_{m_0}$(*вш*|*й*(∅|*ш*)|*м*|*нн*)$_m$ → **V**$^*_{m_0}$*;*

(3.3) *Do* ~**V**$^*_m$ → ∅.

The result after these three steps of operations is called $\hat{\sigma}'_2$.

(4) *Concatenate* $\hat{\sigma}'_1$ *and* $\hat{\sigma}'_2$*. If* $\ell(\hat{\sigma}'_1\hat{\sigma}'_2) = 3$*, do* ~(*а*|*е*|*и*|*й*|*о*|*у*|*ы*|*ь*|*ю*|*я*) → ∅ *on* $\hat{\sigma}'_1\hat{\sigma}'_2$*, and call the result* $\hat{\sigma}'$*; otherwise, define* $\hat{\sigma}' = \hat{\sigma}'_1\hat{\sigma}'_2$.

(5) *Do* ~(*й*|*ь*)$\hat{\chi}$(**V****в*)$_{m_0}$ → *е*$\hat{\chi}$ *on* $\hat{\sigma}'$.

(6) *Replace*

дн~	*((∅|по|с|со)жа(∅|в|м)|сж)*		*(реб|ребен|ребят)*		$\mathbf{X}^{\in}$*(б|в|л|н|ш)е*
ддень	*жал*		*дет*		**X***и*
краш	*усл*	*чдачад*	*~вел*	*~жив*	*~пе (в|л|м)*
красив	*услыш*	*чдач*	*вед*	*жил*	*по*

To improve the performance of verb clustering in Russian, we need the following extension to Algorithm 7.19.

Algorithm 7.20 (Russian verbal essential root)**.** *Let $\hat{\sigma}$ be the token string associated with a Russian word, then its corresponding verbal essential root* VbEssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Break down $\hat{\sigma} = \hat{\sigma}_1\hat{\sigma}_2$ into the concatenation of two strings $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$ (see the notation in Definition 3.1) and $\hat{\sigma}_2$, where the length of the first string $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma})$ is equal to the protected range of $\hat{\sigma}$.*

(2) *Do $\sim(\mathbf{V}_{m_0}(в|л|н|\mathbf{V}_m m)_{m_0})_m \to \varnothing$ on $\hat{\sigma}_2$, and call the result $\hat{\sigma}'_2$.*

(3) *Concatenate $\hat{\sigma}_1$ and $\hat{\sigma}'_2$.*

(4) *Do прид → приид.*

(5) *Do $\mathbf{P}^*$(езд|езж)~ → $\mathbf{P}^*$ех, (бр|вз|воз)(∅|ем|ьм) → бер, $\mathbf{P}^*$(ид|хож|ш|шед) → $\mathbf{P}^*$ход.*

7.2.2 Admissible mutation and approximate clustering

In Russian (as well as Polish and many other Slavic languages), fleeting vowels may appear or disappear in inflected forms of the same word. We need to heuristically detect and remove these fleeting vowels to achieve better clustering results. Meanwhile, in Russian (as well as Polish and many other Slavic languages), one encounters vowel alternations in verb conjugations, as in Spanish or the Germanic languages.

Algorithm 7.21 (Russian vowel blotting)**.** *For a token string $\hat{\sigma}$, its blotted form* BlotFltV($\hat{\sigma}$) *is constructed as follows:*

(1) *Do $\sim\mathbf{C}^*(е|о)^{\hat{\chi}\in}(к|н|л|ц)((\mathbf{V}^*_{m_0}(в|л|н|м)_{m_0})_m|ш) \to \mathbf{C}^*\hat{\chi}$.*

(2) *Do $\sim^{\hat{\chi}\in}(в|м)ер((\mathbf{V}^*_{m_0}(в|л|н|м)_{m_0})_m|ш) \to \hat{\chi}р$.*

(3) *Do е → и, о → а.*

In what follows, we will construct a bivariate Boolean-valued function HrdTest($\hat{\alpha}$,$\hat{\beta}$) on a "simple heredity test function" in Algorithm 7.22, and a set of "admissible suffix mismatch" rules in Algorithm 7.23.

Algorithm 7.22 (Simple heredity test)**.** *Let $\hat{\alpha}'$ be a string obtained from $\hat{\alpha}$ by the following steps:*

(1) *Break down $\hat{\alpha} = \hat{\alpha}_1\hat{\alpha}_2$ into the concatenation of two strings $\hat{\alpha}_1 = \hat{\alpha}^{[\min\{3,\ell(\hat{\alpha})\}]}$ (see the notation in Definition 3.1) and $\hat{\alpha}_2$, where the length of the first string $\ell(\hat{\alpha}_1) = \min\{3, \ell(\hat{\alpha})\}$ is equal to 3 or the length of $\hat{\alpha}$, whichever is shorter.*

(2) *Do $\sim((\mathbf{V}^*_{m_0}(в|л|н|м|ц)_{m_0})_m|ш)$ on $\hat{\alpha}_2$, and call the result $\hat{\alpha}'_2$.*

(3) *Construct $\hat{\alpha}' = \hat{\alpha}_1\hat{\alpha}'_2$.*

Define $\hat{\beta}'$ similarly. The Boolean-valued function SimpHrdTest($\hat{\alpha}$,$\hat{\beta}$) *returns* **TRUE** *if the lowercase form of $\hat{\alpha}$ contains at least one instance of {а, е, и, о, у, ы, ь, ю, я, лб, рт, сн}* **AND** *at least one of the following five conditions holds:*[95]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\beta} = \hat{\alpha}ин$;

(iii) $\hat{\beta} = \hat{\alpha}ш$;

(iv) $\min\{\ell(\hat{\alpha}'), \ell(\hat{\beta}')\} \geq 3$ **AND** ($\hat{\alpha} = \hat{\beta}'$ **OR** $\hat{\alpha}' = \hat{\beta}$ **OR** $\hat{\alpha}' = \hat{\beta}'$).

(v) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{3}$ **AND** $\hat{\beta} = \hat{\alpha}(а|е|о|я)_m(в|м|т)_m$.

In what follows, we define SuffixNW($\hat{\alpha}$,$\hat{\beta}$), RootNW($\hat{\alpha}$,$\hat{\beta}$), NW*($\hat{\alpha}$,$\hat{\beta}$) and SuffixSW($\hat{\alpha}$,$\hat{\beta}$), SuffixSW($\hat{\alpha}$,$\hat{\beta}$), SW*($\hat{\alpha}$,$\hat{\beta}$) as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5. Particular to the Russian case, we further define SuffixNW′($\hat{\alpha}$,$\hat{\beta}$) by performing the substitutions $\sim((\mathbf{V}^*_{m_0}(в|л|н|\mathbf{V}^*_m m)_{m_0})_m|ш) \to \varnothing$ to both components of SuffixNW($\hat{\alpha}$,$\hat{\beta}$). The notation SuffixSW′($\hat{\alpha}$,$\hat{\beta}$) is defined similarly. If SuffixNW′($\hat{\alpha}$,$\hat{\beta}$) = [$\hat{\tau}_1$, $\hat{\tau}_2$], then SuffixNW′($\hat{\beta}$,$\hat{\alpha}$) = [$\hat{\tau}_2$, $\hat{\tau}_1$].

[95] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 7.23 (Admissible suffix mismatch and vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if* $\text{NW}(\hat{\alpha},\hat{\beta}) = \varnothing|[\varnothing,\ u]||[a,\ o]||[e,\ u]||[u,\ \varnothing]$ **AND** $(\text{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* {*a, е, и, о, у, ы, ь, ю, я*} **OR** $\Omega(\text{RootNW}(\hat{\alpha},\hat{\beta})) = (к|н|т|ц|р))$ **AND** *at least one of the following three conditions holds:*

(i) $\text{SuffixNW}'(\hat{\alpha},\hat{\beta}) = [г,\ ж]||[г,\ з]||[д,\ ж]||[ж,\ з]||[ж,\ ч]||[к,\ ч]||[с,\ ш]||[т,\ ч]||[х,\ ш]||[ек,\ ч]||[ок,\ ч]||[ск,\ щ]||[ст,\ д]||[ст,\ щ]$*;*

(ii) $\text{SuffixNW}'(\hat{\beta},\hat{\alpha}) = [г,\ ж]||[г,\ з]||[д,\ ж]||[ж,\ з]||[ж,\ ч]||[к,\ ч]||[с,\ ш]||[т,\ ч]||[х,\ ш]||[ек,\ ч]||[ок,\ ч]||[ск,\ щ]||[ст,\ д]||[ст,\ щ]$*;*

(iii) $\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [\varnothing|(\mathbf{V}_m(\varnothing|в|т|н|т|щ)),\varnothing|(\mathbf{V}_m(\varnothing|в|т|н|т|щ))]$.

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmMut}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW*.*

Algorithm 7.24 (Heredity test function)**.** *The structure of the Russian heredity test function* $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ *is identical to the Polish version (Algorithm 7.11), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW*,* RootSW*,* SuffixSW*,* SW *and the strings $\hat{\alpha}'$,$\hat{\beta}'$ must follow the Russian rules stated above.*

Algorithm 7.25 (Russian verb aspect test)**.** *Let* $\text{NW}(\hat{\alpha},\beta)$ *be the result of performing Needleman–Wunsch alignment on strings $\hat{\alpha}$ and $\hat{\beta}$. Let $\mathbf{X}^*$ be an arbitrary non-empty string. The Boolean-valued function* $\text{VbAspTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following four conditions holds:*

(i) $\text{NW}(\hat{\alpha},\hat{\beta}) = \mathbf{X}^*$ (*in other words, $\hat{\alpha} = \hat{\beta}$);*

(ii) $\text{NW}(\hat{\alpha},\hat{\beta}) = [\mathbf{P}^*,\mathbf{P}^*]\mathbf{X}^*$ (*in other words, $\hat{\alpha}$ agrees with $\hat{\beta}$ up to a pair of word initial mismatching strings, both of which are possible prefixes of Russian verbs*)*;*

(iii) $\text{NW}(\hat{\alpha},\hat{\beta}) = [\mathbf{P}^*,\mathbf{P}^*]\mathbf{X}^*([\varnothing,\ u]||[a,\ o]||[e,\ u]||[u,\ \varnothing]||[o,\ a])\mathbf{X}^*$*, where the two occurrences of* $\mathbf{X}^*$ *may or may not represent the same non-empty string;*

(iv) $\text{NW}(\hat{\alpha},\hat{\beta}) = [\mathbf{P}^*,\mathbf{P}^*]\mathbf{X}^*[\hat{\tau}_1,\hat{\tau}_2]$ *so that either* $[\hat{\tau}'_1,\hat{\tau}'_2]$ *or* $[\hat{\tau}'_2,\hat{\tau}'_1]$ *matches* $[г,\ ж]||[г,\ з]||[д,\ ж]||[ж,\ з]||[ж,\ ч]||[к,\ ч]||[с,\ ш]||[т,\ ч]||[х,\ ш]||[ек,\ ч]||[ок,\ ч]||[ск,\ щ]||[ст,\ д]||[ст,\ щ]$*. Here,* $\hat{\tau}'_1$ *results from doing* $\sim((\mathbf{V}^*_{m_0}(в|т|н|\mathbf{V}^*_m m)_{m_0})_m|ш) \rightarrow \varnothing$ *on* $\hat{\tau}_1$*, and* $\hat{\tau}'_2$ *is similarly defined.*

Note that Smith–Waterman alignment is not used in this test.

The following approximate clustering algorithm for Russian words are very similar to the Polish counterpart (Algorithm 7.13), except that Latin transliteration (Algorithm 7.15) is used before sorting the tokens. For clarity, we state the algorithm in full.

Algorithm 7.26 (Approximate clustering of Russian words)**.** *The approximate clustering of a list of English words* $\{\hat{\alpha}_1,\ \ldots,\ \hat{\alpha}_N\}$ *is completed in three stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1,\text{EffSpell}(\hat{\alpha}_1)),\ \ldots,\ (\hat{\alpha}_N,\text{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the Latin transliteration of second component (Algorithm 7.15 applied to the effective spelling) of each entry. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)},\text{EffSpell}(\hat{\alpha}_{(1)})),\ \ldots,\ (\hat{\alpha}_{(N)},\text{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfies* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n)}),\text{EffSpell}(\hat{\alpha}_{(n+1)})) =$ **FALSE***, where* $n \in \mathbb{Z}\cap[1,N)$*, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)},\text{EffSpell}(\hat{\alpha}_{(1,1)})),\ \ldots,\ (\hat{\alpha}_{(1,n_1)},\text{EffSpell}(\hat{\alpha}_{(1,n_1)}))\},\ \ldots,$ $g_M = \{(\hat{\alpha}_{(M,1)},\text{EffSpell}(\hat{\alpha}_{(M,1)})),\ \ldots,\ (\hat{\alpha}_{(M,n_M)},\text{EffSpell}(\hat{\alpha}_{(M,n_M)}))\}\}$*, where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$*.*

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)},\text{EffSpell}(\hat{\alpha}_{(m,1)})),\ \ldots,\ (\hat{\alpha}_{(m,n_m)},\text{EffSpell}(\hat{\alpha}_{(m,n_m)}))\}$ *where* $m \in \mathbb{Z}\cap[1,M]$*, we augment it into a tagged entry* $G_m = (g_m,\text{EffSpell}(\hat{\alpha}_{(m,1)}),\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})),\text{BlotFltV}(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))))$*. The list* $\{G_1,\ \ldots,\ G_M\}$ *is sorted alphabetically, with respect to* $\text{RuLat}(\text{BlotFltV}(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))))$ *(with highest priority),* $\text{RuLat}(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})))$ *(with medium priority), and* $\text{RuLat}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)},\hat{\gamma}'_{(1)},\hat{\gamma}''_{(1)},\hat{\gamma}'''_{(1)}),\ \ldots,\ G_{(M)} = (g_{(M)},\hat{\gamma}'_{(M)},\hat{\gamma}''_{(M)},\hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\text{SimpHrdTest}(\hat{\gamma}'''_{(m+1)},\hat{\gamma}'''_{(m)}) = \textbf{FALSE}$$

AND

$$\text{SimpHrdTest}(\hat{\gamma}'''_{(m)}, \hat{\gamma}'''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}''_{(m+1)}, \hat{\gamma}''_{(m)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, then add a demarcation line between these two entries. In this manner, the alphabetized list $\{G_{(1)}, \ldots, G_{(M)}\}$ is divided into separate groups of tagged clusters $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$. Generate a list of word clusters $\{\check{\Gamma}_1 = (\check{G}_{(1,1)}, \ldots), \ldots, \check{\Gamma}_K = (\check{G}_{(K,1)}, \ldots)\}$ by discarding all the tags (effective spellings, essential roots, vowel blotted forms) from each Γ_k, where $k \in \mathbb{Z} \cap [1, K]$.

(3) *For each word cluster $\check{\Gamma}_k = (\check{G}_{(k,1)}, \ldots)$, we augment it into a tagged entry*

$$\mathcal{G}_k \equiv (\check{\Gamma}_k, \check{\Gamma}'_k, \check{\Gamma}''_k) := (\check{\Gamma}_k, \text{EssRoot}(\text{EffSpell}(\check{G}_{(k,1)})), \text{VbEssRoot}(\text{EssRoot}(\text{EffSpell}(\check{G}_{(k,1)})))).$$

Recall that $\hat{\sigma}^{-1}$ is the reverse of the string $\hat{\sigma}$ (Definition 3.1). The list $\{\mathcal{G}_1 = (\check{\Gamma}_1, \check{\Gamma}'_1, \check{\Gamma}''_1), \ldots, \mathcal{G}_K = (\check{\Gamma}_K, \check{\Gamma}'_K, \check{\Gamma}''_K)\}$ is sorted sorted alphabetically, with respect to $(\text{RuLat}(\check{\Gamma}''_k))^{-1}$ (with higher priority) and $\text{RuLat}(\check{\Gamma}'_k)$ (with lower priority). If two consecutive neighbors in the alphabetized list $\{\mathcal{G}_{(1)} = (\check{\Gamma}_{(1)}, \check{\Gamma}'_{(1)}, \check{\Gamma}''_{(1)}), \ldots, \mathcal{G}_{(K)} = (\check{\Gamma}_{(K)}, \check{\Gamma}'_{(K)}, \check{\Gamma}''_{(K)})\}$ satisfy

$$\text{VbAspTest}(\hat{\Gamma}''_{(k)}, \hat{\Gamma}''_{(k+1)}) = \textbf{FALSE}$$

AND

$$\text{VbAspTest}(\hat{\Gamma}'_{(k)}, \hat{\Gamma}'_{(k+1)}) = \textbf{FALSE}$$

where $k \in \mathbb{Z} \cap [1, K)$, then add a demarcation line between these two entries. In this manner, the alphabetized list $\{\mathcal{G}_{(1)}, \ldots, \mathcal{G}_{(K)}\}$ is divided into separate groups of tagged clusters $\{\widetilde{\Gamma}_1 = \{\mathcal{G}_{(1,1)}, \ldots, \mathcal{G}_{(1,n_1)}\}, \ldots, \widetilde{\Gamma}_J = \{\mathcal{G}_{(J,1)}, \ldots, \mathcal{G}_{(1,n_J)}\}\}$, from which the output list of word clusters is generated, after removal of all the tags from each $\widetilde{\Gamma}_j$, where $j \in \mathbb{Z} \cap [1, J]$.

Example 7.26.1. Most Russian nouns distinguish six cases (nominative, genitive, dative, accusative, instrumental, prepositional) in singular and plural. A limited number of nouns have additional forms in three more cases (vocative, locative and partitive).

The sample nouns below are chosen according to the following web pages:

`https://en.wiktionary.org/wiki/Appendix:Russian_nouns#Declension_paradigms`

`https://en.wiktionary.org/wiki/Template:ru-noun-table#Basic_examples`

апреле, *апрелей*, *апрелем*, *апрели*, *апрель*, *апрелю*, *апреля*, *апрелям*, *апрелями*, *апрелях* — “April”;

блюд, *блюда*, *блюдам*, *блюдами*, *блюдах*, *блюде*, *блюдо*, *блюдом*, *блюду* — “dish”;

бог, *бога*, *богам*, *богами*, *богах*, *боге*, *боги*, *богов*, *богом*, *богу*, *боже* — “god”;

бояр, *боярам*, *боярами*, *боярах*, *бояре*, *боярин*, *боярина*, *боярине*, *боярином*, *боярину* — “boyar”;

вид, *вида*, *видам*, *видами*, *видах*, *виде*, *видов*, *видом*, *виду*, *виды* — “appearance”;

волчищ, *волчища*, *волчищам*, *волчищами*, *волчищах*, *волчище*, *волчищей*, *волчищем*, *волчищи*, *волчищу* — “maturewolf”;

воротец, *воротца*, *воротцам*, *воротцами*, *воротцах* — “wicket”;

времена, *временам*, *временами*, *временах*, *временем*, *времени*, *времён*, *время* — “time”;

газет, *газета*, *газетам*, *газетами*, *газетах*, *газете*, *газетой*, *газетою*, *газету*, *газеты* — “newspaper”;

герое, *героев*, *героем*, *герои*, *герой*, *герою*, *героя*, *героям*, *героями*, *героях* — “hero”;

гетеросексуальности, *гетеросексуальность*, *гетеросексуальностью* — “heterosexuality”;

голов, *голова*, *головам*, *головами*, *головах*, *голове*, *головой*, *головою*, *голову*, *головы* — “head”;

дверей, *двери*, *дверь*, *дверьми*, *дверью*, *дверям*, *дверями*, *дверях* — “door”;

дев, *дева*, *девам*, *девами*, *девах*, *деве*, *дево*, *девой*, *девою*, *деву*, *девы* — “maiden”;

день, *дне*, *дней*, *днём*, *дни*, *дню*, *дня*, *дням*, *днями*, *днях* — “day”;

детей, *дети*, *детьми*, *детям*, *детях*, *ребёнка*, *ребёнке*, *ребёнком*, *ребёнку*, *ребёнок*, *ребят*, *ребята*, *ребятам*, *ребятами*, *ребятах* — “child”;

друг, *друга*, *друге*, *другом*, *другу*, *друже*, *друзей*, *друзья*, *друзьям*, *друзьями*, *друзьях* — “friend”;

железа, *железам*, *железами*, *железах*, *железе*, *железой*, *железою*, *железу*, *железы*, *желёз* — “gland”;

жилищ, *жилища*, *жилищам*, *жилищами*, *жилищах*, *жилище*, *жилищем*, *жилищу* — “dwelling”;

житие, *житием*, *житии*, *житий*, *житию*, *жития*, *житиям*, *житиями*, *житиях* — “life”;

журнал, *журнала*, *журналам*, *журналами*, *журналах*, *журнале*, *журналов*, *журналом*, *журналу*, *журналы* — “magazine”;

искр, *искра*, *искрам*, *искрами*, *искрах*, *искре*, *искрой*, *искрою*, *искру*, *искры* — “spark”;

историей, *историею*, *истории*, *историй*, *историю*, *история*, *историям*, *историями*, *историях* — “history”;

камень, *каменьев*, *каменья*, *каменьям*, *каменьями*, *каменьях*, *камне*, *камней*, *камнем*, *камни*, *камню*, *камня*, *камням*, *камнями*, *камнях* — “stone”;

карандаш, *карандаша*, *карандашам*, *карандашами*, *карандашах*, *карандаше*, *карандашей*, *карандаши*, *карандашом*, *карандашу* — “pencil”;

книг, *книга*, *книгам*, *книгами*, *книгах*, *книге*, *книги*, *книгой*, *книгою*, *книгу* — “book”;

колен, *колена*, *коленам*, *коленами*, *коленах*, *колене*, *колено*, *коленом*, *колену* — “tribe”;

колена, *колене*, *коленей*, *колени*, *колено*, *коленом*, *колену*, *коленям*, *коленями*, *коленях* — “knee”;

колена, *колене*, *колено*, *коленом*, *колену*, *коленьев*, *коленья*, *коленьям*, *коленьями*, *коленьях* — “joint”;

копеек, *копейка*, *копейкам*, *копейками*, *копейках*, *копейке*, *копейки*, *копейкой*, *копейкою*, *копейку* — “kopek”;

лёд, *льда*, *льдам*, *льдами*, *льдах*, *льде*, *льдов*, *льдом*, *льду*, *льды* — “ice”;

лоскут, *лоскута*, *лоскутам*, *лоскутами*, *лоскутах*, *лоскуте*, *лоскутов*, *лоскутом*, *лоскуту*, *лоскуты*, *лоскутьев*, *лоскутья*, *лоскутьям*, *лоскутьями*, *лоскутьях* — “rag”;

лошадей, *лошади*, *лошадь*, *лошадьми*, *лошадью*, *лошадям*, *лошадями*, *лошадях* — “horse”;

мальчат, *мальчата*, *мальчатам*, *мальчатами*, *мальчатах*, *мальчонка*, *мальчонкам*, *мальчонками*, *мальчонках*, *мальчонке*, *мальчонки*, *мальчонков*, *мальчонком*, *мальчонку*, *мальчонок* — “littleboy”;

море, *морей*, *морем*, *морю*, *моря*, *морям*, *морями*, *морях* — “sea”;

мост, *моста*, *мостам*, *мостами*, *мостах*, *мосте*, *мостов*, *мостом*, *мосту*, *мосты* — “bridge”;

муж, *мужа*, *муже*, *мужей*, *мужем*, *мужу*, *мужья*, *мужьям*, *мужьями*, *мужьях* — “husband”;

неделе, *неделей*, *неделею*, *недели*, *недель*, *неделю*, *неделя*, *неделям*, *неделями*, *неделях* — “week”;

нож, *ножа*, *ножам*, *ножами*, *ножах*, *ноже*, *ножей*, *ножи*, *ножом*, *ножу* — “knife”;

няне, *няней*, *нянею*, *няни*, *нянь*, *няню*, *няня*, *няням*, *нянями*, *нянях* — “nurse”;

озерца, *озерце*, *озерцо*, *озерцом*, *озерцу*, *озёрец*, *озёрца*, *озёрцам*, *озёрцами*, *озёрцах* — “lake”;

окна, *окнам*, *окнами*, *окнах*, *окне*, *окно*, *окном*, *окну*, *окон* — “window”;

ореол, *ореола*, *ореолам*, *ореолами*, *ореолах*, *ореоле*, *ореолов*, *ореолом*, *ореолу*, *ореолы* — “halo”;

отец, *отца*, *отцам*, *отцами*, *отцах*, *отце*, *отцов*, *отцом*, *отцу*, *отцы* — “father”;

палец, *пальца*, *пальцам*, *пальцами*, *пальцах*, *пальце*, *пальцев*, *пальцем*, *пальцу*, *пальцы* — “finger”;

писем, *письма*, *письмам*, *письмами*, *письмах*, *письме*, *письмо*, *письмом*, *письму* — “letter”;

площадей, *площади*, *площадь*, *площадью*, *площадям*, *площадями*, *площадях* — “square”;

повод, *повода*, *поводе*, *поводом*, *поводу*, *поводьев*, *поводья*, *поводьям*, *поводьями*, *поводьях* — “rein”;

поле, *полей*, *полем*, *полю*, *поля*, *полям*, *полями*, *полях* — “field”;

пух, *пуха*, *пухе*, *пухом*, *пуху* — “feather”;

работ, *работа*, *работам*, *работами*, *работах*, *работе*, *работой*, *работою*, *работу*, *работы* — “task”;

рот, *рта*, *ртам*, *ртами*, *ртах*, *рте*, *ртов*, *ртом*, *рту*, *рты* — “mouth”;

сапожек, *сапожка*, *сапожкам*, *сапожками*, *сапожках*, *сапожке*, *сапожки*, *сапожков*, *сапожком*, *сапожку*, *сапожок* — “boot”;

сестёр, *сестра*, *сестре*, *сестрой*, *сестрою*, *сестру*, *сестры*, *сёстрам*, *сёстрами*, *сёстрах*, *сёстры* — “sister”;

словаре, *словарей*, *словарём*, *словари*, *словарь*, *словарю*, *словаря*, *словарям*, *словарями*, *словарях* — “dictionary”;

случае, *случаев*, *случаем*, *случаи*, *случай*, *случаю*, *случая*, *случаям*, *случаями*, *случаях* — “event”;

сна, *снам*, *снами*, *снах*, *сне*, *снов*, *сном*, *сну*, *сны*, *сон* — “dream”;

снег, *снега*, *снегам*, *снегами*, *снегах*, *снеге*, *снегов*, *снегом*, *снегу* — “snow”;

стол, *стола*, *столам*, *столами*, *столах*, *столе*, *столов*, *столом*, *столу*, *столы* — “table”;

сын, *сына*, *сынам*, *сынами*, *сынах*, *сыне*, *сынов*, *сыновей*, *сыновья*, *сыновьям*, *сыновьями*, *сыновьях*, *сыном*, *сыну*, *сыны* — “son”;

товарищ, *товарища*, *товарищам*, *товарищами*, *товарищах*, *товарище*, *товарищей*, *товарищем*, *товарищи*, *товарищу* — “comrade”;

урок, *урока*, *урокам*, *уроками*, *уроках*, *уроке*, *уроки*, *уроков*, *уроком*, *уроку* — “lesson”;

учащегося, *учащемся*, *учащемуся*, *учащиеся*, *учащийся*, *учащимися*, *учащимся*, *учащихся* — “pupil”;

учение, *учением*, *учении*, *учений*, *учению*, *учения*, *учениям*, *учениями*, *учениях* — “learning”;

училищ, *училища*, *училищам*, *училищами*, *училищах*, *училище*, *училищем*, *училищу* — “college”;

учителе, *учителей*, *учителем*, *учители*, *учитель*, *учителю*, *учителя*, *учителям*, *учителями*, *учителях* — “teacher”;

феномен, *феномена*, *феноменам*, *феноменами*, *феноменах*, *феномене*, *феноменов*, *феноменом*, *феномену*, *феномены* — “phenomenon”;

хвала, *хвалам*, *хвалами*, *хвалах*, *хвале*, *хвалой*, *хвалою*, *хвалу*, *хвалы* — “praise”;

чае, *чаем*, *чаёв*, *чаи*, *чай*, *чаю*, *чая*, *чаям*, *чаями*, *чаях* — “tea”.

It is worth noting that, unlike Polish, the consonants of the Russian noun stem mostly remain stable in all declined forms. The sample nouns given above are clustered by our algorithm as follows:

{*апреле*, *апрелей*, *апрелем*, *апрели*, *апрель*, *апрелю*, *апреля*, *апрелям*, *апрелями*, *апрелях*},

{*блюд*, *блюда*, *блюдам*, *блюдами*, *блюдах*, *блюде*, *блюдо*, *блюдом*, *блюду*},

{*бог*, *бога*, *богам*, *богами*, *богах*, *боге*, *боги*, *богов*, *богом*, *богу*, *боже*},

{*бояр*, *боярам*, *боярами*, *боярах*, *бояре*, *боярин*, *боярина*, *боярине*, *боярином*, *боярину*},

{*вид*, *вида*, *видам*, *видами*, *видах*, *виде*, *видов*, *видом*, *виду*, *виды*},

{*волчищ*, *волчища*, *волчищам*, *волчищами*, *волчищах*, *волчище*, *волчищей*, *волчищем*, *волчищи*, *волчищу*},

{*воротец*, *воротца*, *воротцам*, *воротцами*, *воротцах*},

{*времена*, *временам*, *временами*, *временах*, *временем*, *времени*, *времён*, *время*},

{*газет*, *газета*, *газетам*, *газетами*, *газетах*, *газете*, *газетой*, *газетою*, *газету*, *газеты*},

{*герое*, *героев*, *героем*, *герои*, *герой*, *герою*, *героя*, *героям*, *героями*, *героях*},

{*гетеросексуальности*, *гетеросексуальность*, *гетеросексуальностью*},

{*голов*, *голова*, *головам*, *головами*, *головах*, *голове*, *головой*, *головою*, *голову*, *головы*},

{*дверей*, *двери*, *дверь*, *дверьми*, *дверью*, *дверям*, *дверями*, *дверях*},

{*дев*, *дева*, *девам*, *девами*, *девах*, *деве*, *дево*, *девой*, *девою*, *деву*, *девы*},

{*день*, *дне*, *дней*, *днём*, *дни*, *дню*, *дня*, *дням*, *днями*, *днях*},

{*детей*, *дети*, *детьми*, *детям*, *детях*, *ребёнка*, *ребёнке*, *ребёнком*, *ребёнку*, *ребёнок*, *ребят*, *ребята*, *ребятам*, *ребятами*, *ребятах*},

{*друг*, *друга*, *друге*, *другом*, *другу*, *друже*, *друзей*, *друзья*, *друзьям*, *друзьями*, *друзьях*},

{*железа*, *железам*, *железами*, *железах*, *железе*, *железой*, *железою*, *железу*, *железы*, *желёз*},

{*жилищ*, *жилища*, *жилищам*, *жилищами*, *жилищах*, *жилище*, *жилищем*, *жилищу*},

{*житие*, *житием*, *житии*, *житий*, *житию*, *жития*, *житиям*, *житиями*, *житиях*},

{*журнал*, *журнала*, *журналам*, *журналами*, *журналах*, *журнале*, *журналов*, *журналом*, *журналу*, *журналы*},

{*искр*, *искра*, *искрам*, *искрами*, *искрах*, *искре*, *искрой*, *искрою*, *искру*, *искры*},

{*историей*, *историею*, *истории*, *историй*, *историю*, *история*, *историям*, *историями*, *историях*},

{*камень*, *каменьев*, *каменья*, *каменьям*, *каменьями*, *каменьях*, *камне*, *камней*, *камнем*, *камни*, *камню*, *камня*, *камням*, *камнями*, *камнях*},

{*карандаш*, *карандаша*, *карандашам*, *карандашами*, *карандашах*, *карандаше*, *карандашей*, *карандаши*, *карандашом*, *карандашу*},

{*книг*, *книга*, *книгам*, *книгами*, *книгах*, *книге*, *книги*, *книгой*, *книгою*, *книгу*},

{*колен*, *колена*, *коленам*, *коленами*, *коленах*, *колене*, *коленей*, *колени*, *колено*, *коленом*, *колену*, *коленьев*, *коленья*, *коленьям*, *коленьями*, *коленьях*, *коленям*, *коленями*, *коленях*},

{*копеек*, *копейка*, *копейкам*, *копейками*, *копейках*, *копейке*, *копейки*, *копейкой*, *копейкою*, *копейку*},

{*лёд*, *льда*, *льдам*, *льдами*, *льдах*, *льде*, *льдов*, *льдом*, *льду*, *льды*},

{*лоскут*, *лоскута*, *лоскутам*, *лоскутами*, *лоскутах*, *лоскуте*, *лоскутов*, *лоскутом*, *лоскуту*, *лоскуты*, *лоскутьев*, *лоскутья*, *лоскутьям*, *лоскутьями*, *лоскутьях*},

{*лошадей*, *лошади*, *лошадь*, *лошадьми*, *лошадью*, *лошадям*, *лошадями*, *лошадях*},

{*мальчат*, *мальчата*, *мальчатам*, *мальчатами*, *мальчатах*, *мальчонка*, *мальчонкам*, *мальчонками*, *мальчонках*, *мальчонке*, *мальчонки*, *мальчонков*, *мальчонком*, *мальчонку*, *мальчонок*},

{*море*, *морей*, *морем*, *морю*, *моря*, *морям*, *морями*, *морях*},

{*мост*, *моста*, *мостам*, *мостами*, *мостах*, *мосте*, *мостов*, *мостом*, *мосту*, *мосты*},

{*муж*, *мужа*, *муже*, *мужей*, *мужем*, *мужу*, *мужья*, *мужьям*, *мужьями*, *мужьях*},

{*неделе*, *недель*, *неделю*, *недели*, *недель*, *неделю*, *неделя*, *неделям*, *неделями*, *неделях*},

{*нож*, *ножа*, *ножам*, *ножами*, *ножах*, *ноже*, *ножей*, *ножи*, *ножом*, *ножу*},

{*няне*, *няней*, *нянею*, *няни*, *нянь*, *няню*, *няня*, *няням*, *нянями*, *нянях*},

{*озерца*, *озерце*, *озерцо*, *озерцом*, *озерцу*, *озёрец*, *озёрца*, *озёрцам*, *озёрцами*, *озёрцах*},

{*окна*, *окнам*, *окнами*, *окнах*, *окне*, *окно*, *окном*, *окну*, *окон*},

{*ореол*, *ореола*, *ореолам*, *ореолами*, *ореолах*, *ореоле*, *ореолов*, *ореолом*, *ореолу*, *ореолы*},

{*отец*, *отца*, *отцам*, *отцами*, *отцах*, *отце*, *отцов*, *отцом*, *отцу*, *отцы*},

{*палец*, *пальца*, *пальцам*, *пальцами*, *пальцах*, *пальце*, *пальцев*, *пальцем*, *пальцу*, *пальцы*},

{*писем*, *письма*, *письмам*, *письмами*, *письмах*, *письме*, *письмо*, *письмом*, *письму*},

{*площадей*, *площади*, *площадь*, *площадью*, *площадям*, *площадями*, *площадях*},

{*повод*, *повода*, *поводе*, *поводом*, *поводу*, *поводьев*, *поводья*, *поводьям*, *поводьями*, *поводьях*},

{*поле*, *полей*, *полем*, *полю*, *поля*, *полям*, *полями*, *полях*},

{*пух*, *пуха*, *пухе*, *пухом*, *пуху*},

{*работ*, *работа*, *работам*, *работами*, *работах*, *работе*, *работой*, *работою*, *работу*, *работы*},

{*рот*, *рта*, *ртам*, *ртами*, *ртах*, *рте*, *ртов*, *ртом*, *рту*, *рты*},

{*сапожек*, *сапожка*, *сапожкам*, *сапожками*, *сапожках*, *сапожке*, *сапожки*, *сапожков*, *сапожком*, *сапожку*, *сапожок*},

{*сестёр*, *сестра*, *сестре*, *сестрой*, *сестрою*, *сестру*, *сестры*, *сёстрам*, *сёстрами*, *сёстрах*, *сёстры*},

{*словаре*, *словарей*, *словарём*, *словари*, *словарь*, *словарю*, *словаря*, *словарям*, *словарями*, *словарях*},

{*случае*, *случаев*, *случаем*, *случаи*, *случай*, *случаю*, *случая*, *случаям*, *случаями*, *случаях*},

{*сна*, *снам*, *снами*, *снах*, *сне*, *снов*, *сном*, *сну*, *сны*, *сон*},

{*снег*, *снега*, *снегам*, *снегами*, *снегах*, *снеге*, *снегов*, *снегом*, *снегу*},

{*стол*, *стола*, *столам*, *столами*, *столах*, *столе*, *столов*, *столом*, *столу*, *столы*},

{*сын*, *сына*, *сынам*, *сынами*, *сынах*, *сыне*, *сынов*, *сыновей*, *сыновья*, *сыновьям*, *сыновьями*, *сыновьях*, *сыном*, *сыну*, *сыны*},

{*товарищ*, *товарища*, *товарищам*, *товарищами*, *товарищах*, *товарище*, *товарищей*, *товарищем*, *товарищи*, *товарищу*},

{*урок*, *урока*, *урокам*, *уроками*, *уроках*, *уроке*, *уроки*, *уроков*, *уроком*, *уроку*},

{*учащегося*, *учащемся*, *учащемуся*, *учащиеся*, *учащийся*, *учащимися*, *учащимся*, *учащихся*, *учителе*, *учителей*, *учителем*, *учители*, *учитель*, *учителю*, *учителя*, *учителям*, *учителями*, *учителях*},

{*учение*, *учением*, *учении*, *учений*, *учению*, *учения*, *учениям*, *учениями*, *учениях*},

{*училищ*, *училища*, *училищам*, *училищами*, *училищах*, *училище*, *училищем*, *училищу*},

{*феномен*, *феномена*, *феноменам*, *феноменами*, *феноменах*, *феномене*, *феноменов*, *феноменом*, *феномену*, *феномены*},

{*хвала*, *хвалам*, *хвалами*, *хвалах*, *хвале*, *хвалой*, *хвалою*, *хвалу*, *хвалы*},

{*чае*, *чаем*, *чаёв*, *чаи*, *чай*, *чаю*, *чая*, *чаям*, *чаями*, *чаях*},

As we send the following sample adjectives (including participles of certain verbs)

большая, *большего*, *большее*, *большей*, *большем*, *большему*, *большею*, *большие*, *больший*, *большим*, *большими*, *больших*, *большого*, *большое*, *большой*, *большом*, *большому*, *большою*, *большую*, *велик*, *велика*, *велики*, *велико*, *наибольшая*, *наибольшего*, *наибольшее*, *наибольшей*, *наибольшем*, *наибольшему*, *наибольшею*, *наибольшие*, *наибольший*, *наибольшим*, *наибольшими*, *наибольших*, *наибольшую* — “big”;

делавшая, *делавшего*, *делавшее*, *делавшей*, *делавшем*, *делавшему*, *делавшею*, *делавшие*, *делавший*, *делавшим*, *делавшими*, *делавших*, *делавшую* — “do (past active participles)”;

делающая, *делающего*, *делающее*, *делающей*, *делающем*, *делающему*, *делающею*, *делающие*, *делающий*, *делающим*, *делающими*, *делающих*, *делающую* — “do (present active participles)”;

дорог, *дорога*, *дорогая*, *дороги*, *дорогие*, *дорогим*, *дорогими*, *дорогих*, *дорого*, *дорогого*, *дорогое*, *дорогой*, *дорогом*, *дорогому*, *дорогою*, *дорогую*, *дороже* — “expensive”;

красив, *красива*, *красивая*, *красивее*, *красивейший*, *красиво*, *красивого*, *красивое*, *красивой*, *красивом*, *красивому*, *красивою*, *красивую*, *красивы*, *красивые*, *красивый*, *красивым*, *красивыми*, *красивых*, *краше* — “beautiful”;

крут, *крута*, *крутая*, *крутейший*, *круто*, *крутого*, *крутое*, *крутой*, *крутом*, *крутому*, *крутою*, *крутую*, *круты*, *крутые*, *крутым*, *крутыми*, *крутых*, *круче* — “steep”;

любим, *любима*, *любимая*, *любимо*, *любимого*, *любимое*, *любимой*, *любимом*, *любимому*, *любимою*, *любимую*, *любимы*, *любимые*, *любимый*, *любимым*, *любимыми*, *любимых* — “love (present passive participles)”;

молод, *молода*, *молодая*, *молодо*, *молодого*, *молодое*, *молодой*, *молодом*, *молодому*, *молодою*, *молодую*, *молоды*, *молодые*, *молодым*, *молодыми*, *молодых*, *моложе* — “young”;

мягка, *мягкая*, *мягки*, *мягкие*, *мягкий*, *мягким*, *мягкими*, *мягких*, *мягко*, *мягкого*, *мягкое*, *мягкой*, *мягком*, *мягкому*, *мягкою*, *мягкую*, *мягок*, *мягчайший*, *мягче*, *наимягчайший* — “soft”;

решена, *решено*, *решены*, *решён*, *решённая*, *решённого*, *решённое*, *решённой*, *решённом*, *решённому*, *решённою*, *решённую*, *решённые*, *решённый*, *решённым*, *решёнными*, *решённых* — “decide (past passive perfect participles)”;

тих, *тиха*, *тихая*, *тихи*, *тихие*, *тихий*, *тихим*, *тихими*, *тихих*, *тихо*, *тихого*, *тихое*, *тихой*, *тихом*, *тихому*, *тихою*, *тихую*, *тишайший*, *тише* — “silent”;

част, *часта*, *частая*, *часто*, *частого*, *частое*, *частой*, *частом*, *частому*, *частою*, *частую*, *часты*, *частые*, *частый*, *частым*, *частыми*, *частых*, *чаще* — “frequent”.

to our algorithm, we receive the following results:

{*большая*, *большего*, *большее*, *большей*, *большем*, *большему*, *большею*, *большие*, *больший*, *большим*, *большими*, *больших*, *большого*, *большое*, *большой*, *большом*, *большому*, *большою*, *большую*, *велик*, *велика*, *велики*, *велико*, *наибольшая*, *наибольшего*, *наибольшее*, *наибольшей*, *наибольшем*, *наибольшему*, *наибольшею*, *наибольшие*, *наибольший*, *наибольшим*, *наибольшими*, *наибольших*, *наибольшую*},

{*делавшая*, *делавшего*, *делавшее*, *делавшей*, *делавшем*, *делавшему*, *делавшею*, *делавшие*, *делавший*, *делавшим*, *делавшими*, *делавших*, *делавшую*, *делающая*, *делающего*, *делающее*, *делающей*, *делающем*, *делающему*, *делающею*, *делающие*, *делающий*, *делающим*, *делающими*, *делающих*, *делающую*},

{*дорог*, *дорога*, *дорогая*, *дороги*, *дорогие*, *дорогим*, *дорогими*, *дорогих*, *дорого*, *дорогого*, *дорогое*, *дорогой*, *дорогом*, *дорогому*, *дорогою*, *дорогую*, *дороже*},

{*красив*, *красива*, *красивая*, *красивее*, *красивейший*, *красиво*, *красивого*, *красивое*, *красивой*, *красивом*, *красивому*, *красивою*, *красивую*, *красивы*, *красивые*, *красивый*, *красивым*, *красивыми*, *красивых*, *краше*},

{*крут*, *крута*, *крутая*, *крутейший*, *круто*, *крутого*, *крутое*, *крутой*, *крутом*, *крутому*, *крутою*, *крутую*, *круты*, *крутые*, *крутым*, *крутыми*, *крутых*, *круче*},

{*любим*, *любима*, *любимая*, *любимо*, *любимого*, *любимое*, *любимой*, *любимом*, *любимому*, *любимою*, *любимую*, *любимы*, *любимые*, *любимый*, *любимым*, *любимыми*, *любимых*},

{*молод*, *молода*, *молодая*, *молодо*, *молодого*, *молодое*, *молодой*, *молодом*, *молодому*, *молодою*, *молодую*, *молоды*, *молодые*, *молодым*, *молодыми*, *молодых*, *моложе*},

{*мягка*, *мягкая*, *мягки*, *мягкие*, *мягкий*, *мягким*, *мягкими*, *мягких*, *мягко*, *мягкого*, *мягкое*, *мягкой*, *мягком*, *мягкому*, *мягкою*, *мягкую*, *мягок*, *мягчайший*, *мягче*, *наимягчайший*},

{*решена*, *решено*, *решены*, *решён*, *решённая*, *решённого*, *решённое*, *решённой*, *решённом*, *решённому*, *решённою*, *решённую*, *решённые*, *решённый*, *решённым*, *решёнными*, *решённых*},

{*тих*, *тиха*, *тихая*, *тихи*, *тихие*, *тихий*, *тихим*, *тихими*, *тихих*, *тихо*, *тихого*, *тихое*, *тихой*, *тихом*, *тихому*, *тихою*, *тихую*, *тишайший*, *тише*},

{*част*, *часта*, *частая*, *часто*, *частого*, *частое*, *частой*, *частом*, *частому*, *частою*, *частую*, *часты*, *частые*, *частый*, *частым*, *частыми*, *частых*, *чаще*},

Example 7.26.2. Like Polish, a great majority of Russian verbs come in imperfective/perfective pairs. Consonant alternations still find their way into the Russian verb stems (more precisely, the end of stems) in conjugated forms. In the following, we list representative verbs from all the 16 Zaliznyak classes (`https://en.wiktionary.org/wiki/Appendix:Russian_verbs`), together with their imperfective/perfective counterparts that are formed in various ways (we note that the two aspects of the same verb may not belong to the same Zaliznyak class). These verb conjugations may also involve consonant changes in the stem, or other types of irregularities. The Zaliznyak class number, which will be shown parenthetically after the English translation for each word family below, describes at least one aspect (imperfective or perfective) of the verb in question.

делав, *делавши*, *делавший*, *делаем*, *делаемый*, *делает*, *делаете*, *делаешь*, *делай*, *делайте*, *делал*, *делала*, *делали*, *делало*, *деланный*, *делать*, *делаю*, *делают*, *делающий*, *делая*, *сделав*, *сделавши*, *сделавший*, *сделаем*, *сделает*, *сделаете*, *сделаешь*, *сделай*, *сделайте*, *сделал*, *сделала*, *сделали*, *сделало*, *сделанный*, *сделать*, *сделаю*, *сделают* — “do” (1);

нарисовав, *нарисовавши*, *нарисовавший*, *нарисовал*, *нарисовала*, *нарисовали*, *нарисовало*, *нарисованный*, *нарисовать*, *нарисуем*, *нарисует*, *нарисуете*, *нарисуешь*, *нарисуй*, *нарисуйте*, *нарисую*, *нарисуют*, *рисовав*, *рисовавши*, *рисовавший*, *рисовал*, *рисовала*, *рисовали*, *рисовало*, *рисованный*, *рисовать*, *рисуем*, *рисуемый*, *рисует*, *рисуете*, *рисуешь*, *рисуй*, *рисуйте*, *рисую*, *рисуют*, *рисующий*, *рисуя* — “draw” (2a);

блевав, *блевавши*, *блевавший*, *блевал*, *блевала*, *блевали*, *блевало*, *блеванём*, *блеванёт*, *блеванёте*, *блеванёшь*, *блевани*, *блеваните*, *блевану*, *блеванув*, *блеванувши*, *блеванувший*, *блеванул*, *блеванула*, *блеванули*, *блевануло*, *блеванут*, *блевануть*, *блевать*, *блёванный*, *блюём*, *блюёт*, *блюёте*, *блюёшь*, *блюй*, *блюйте*, *блюю*, *блюют*, *блюющий*, *блюя* — “vomit” (2b);

гиб, *гибла*, *гибли*, *гибло*, *гибнем*, *гибнет*, *гибнете*, *гибнешь*, *гибни*, *гибните*, *гибну*, *гибнув*, *гибнувши*, *гибнувший*, *гибнул*, *гибнут*, *гибнуть*, *гибнущий*, *погиб*, *погибла*, *погибли*, *погибло*, *погибнем*, *погибнет*, *погибнете*, *погибнешь*, *погибни*, *погибните*, *погибну*, *погибнут*, *погибнуть*, *погибши*, *погибший* — “die” (3a);

рискнём, *рискнёт*, *рискнёте*, *рискнёшь*, *рискни*, *рискните*, *рискну*, *рискнув*, *рискнувши*, *рискнувший*, *рискнул*, *рискнула*, *рискнули*, *рискнуло*, *рискнут*, *рискнуть*, *рисковав*, *рисковавши*, *рисковавший*, *рисковал*, *рисковала*, *рисковали*, *рисковало*, *рисковать*, *рискуем*, *рискует*, *рискуете*, *рискуешь*, *рискуй*, *рискуйте*, *рискую*, *рискуют*, *рискующий*, *рискуя* — “risk” (3b);

взглядывав, *взглядывавши*, *взглядывавший*, *взглядываем*, *взглядываемый*, *взглядывает*, *взглядываете*, *взглядываешь*, *взглядывай*, *взглядывайте*, *взглядывал*, *взглядывала*, *взглядывали*, *взглядывало*, *взглядывать*, *взглядываю*, *взглядывают*, *взглядывающий*, *взглядывая*, *взглянем*, *взглянет*, *взглянете*, *взглянешь*, *взгляни*, *взгляните*, *взгляну*, *взглянув*, *взглянувши*, *взглянувший*, *взглянул*, *взглянула*, *взглянули*, *взглянуло*, *взглянут*, *взглянуть*, *глядев*, *глядевши*, *глядевший*, *глядел*, *глядела*, *глядели*, *глядело*, *глядеть*, *гляди*, *глядим*, *глядит*, *глядите*, *глядишь*, *глядя*, *глядят*, *глядящий*, *гляжу* — “glance” (3c);

жаленный, *жалив*, *жаливши*, *жаливший*, *жалил*, *жалила*, *жалили*, *жалило*, *жалим*, *жалимый*, *жалит*, *жалите*, *жалить*, *жалишь*, *жаль*, *жальте*, *жалю*, *жаля*, *жалят*, *жалящий*, *ужаленный*, *ужалив*, *ужаливши*, *ужаливший*, *ужалил*, *ужалила*, *ужалили*, *ужалило*, *ужалим*, *ужалит*, *ужалите*, *ужалить*, *ужалишь*, *ужаль*, *ужальте*, *ужалю*, *ужалят* — “sting” (4a);

пощади, *пощадив*, *пощадивши*, *пощадивший*, *пощадил*, *пощадила*, *пощадили*, *пощадило*, *пощадим*, *пощадит*, *пощадите*, *пощадить*, *пощадишь*, *пощадят*, *пощажённый*, *пощажу*, *щади*, *щадив*, *щадивши*, *щадивший*, *щадил*, *щадила*, *щадили*, *щадило*, *щадим*, *щадимый*, *щадит*, *щадите*, *щадить*, *щадишь*, *щадя*, *щадят*, *щадящий*, *щажённый*, *щажу* — “spare” (4b);

люби, *любив*, *любивши*, *любивший*, *любил*, *любила*, *любили*, *любило*, *любим*, *любимый*, *любит*, *любите*, *любить*, *любишь*, *люблeнный*, *люблю*, *любя*, *любят*, *любящий*, *полюби*, *полюбив*, *полюбивши*, *полюбивший*, *полюбил*, *полюбила*, *полюбили*, *полюбило*, *полюбим*, *полюбит*, *полюбите*, *полюбить*, *полюбишь*, *полюбленный*, *полюблю*, *полюбят* — “love” (4c);

слыша, *слышав*, *слышавши*, *слышавший*, *слышал*, *слышала*, *слышали*, *слышало*, *слышанный*, *слышат*, *слышать*, *слышащий*, *слышим*, *слышимый*, *слышит*, *слышите*, *слышишь*, *слышу*, *услышав*, *услышавши*, *услышавший*, *услышал*, *услышала*, *услышали*, *услышало*, *услышанный*, *услышат*, *услышать*, *услышим*, *услышит*, *услышите*, *услышишь*, *услышу*, *услышь*, *услышьте* — “hear” (5a);

бренча, *бренчав*, *бренчавши*, *бренчавший*, *бренчал*, *бренчала*, *бренчали*, *бренчало*, *бренчат*, *бренчать*, *бренчащий*, *бренчи*, *бренчим*, *бренчит*, *бренчите*, *бренчишь*, *бренчу* — “jingle” (5b);

изгнав, *изгнавши*, *изгнавший*, *изгнал*, *изгнала*, *изгнали*, *изгнало*, *изгнанный*, *изгнать*, *изгони*, *изгоним*, *изгонит*, *изгоните*, *изгонишь*, *изгоню*, *изгоняв*, *изгонявши*, *изгонявший*, *изгоняем*, *изгоняемый*, *изгоняет*, *изгоняете*, *изгоняешь*, *изгоняй*, *изгоняйте*, *изгонял*, *изгоняла*, *изгоняли*, *изгоняло*, *изгонят*, *изгонять*, *изгоняю*, *изгоняют*, *изгоняющий*, *изгоняя* — “banish” (5c);

веем, *веемый*, *веет*, *веете*, *веешь*, *вей*, *вейте*, *вею*, *веют*, *веющий*, *вея*, *веяв*, *веявши*, *веявший*, *веял*, *веяла*, *веяли*, *веяло*, *веянный*, *веять* — “flutter” (6a);

взывав, *взывавши*, *взывавший*, *взываем*, *взывает*, *взываете*, *взываешь*, *взывай*, *взывайте*, *взывал*, *взывала*, *взывали*, *взывало*, *взывать*, *взываю*, *взывают*, *взывающий*, *взывая*, *воззвав*, *воззвавши*, *воззвавший*, *воззвал*, *воззвала*, *воззвали*, *воззвало*, *воззванный*, *воззвать*, *воззовём*, *воззовёт*, *воззовёте*, *воззовёшь*, *воззови*, *воззовите*, *воззову*, *воззовут* — “appeal” (6b);

взыскав, *взыскавши*, *взыскавший*, *взыскал*, *взыскала*, *взыскали*, *взыскало*, *взысканный*, *взыскать*, *взыскивав*, *взыскивавши*, *взыскивавший*, *взыскиваем*, *взыскиваемый*, *взыскивает*, *взыскиваете*, *взыскиваешь*, *взыскивай*, *взыскивайте*, *взыскивал*, *взыскивала*, *взыскивали*, *взыскивало*, *взыскивать*, *взыскиваю*, *взыскивают*, *взыскивающий*, *взыскивая*, *взыщем*, *взыщет*, *взыщете*, *взыщешь*, *взыщи*, *взыщите*, *взыщу*, *взыщут* — “recover” (6c);

влез, *влезав*, *влезавши*, *влезавший*, *влезаем*, *влезает*, *влезаете*, *влезаешь*, *влезай*, *влезайте*, *влезал*, *влезала*, *влезали*, *влезало*, *влезать*, *влезаю*, *влезают*, *влезающий*, *влезая*, *влезем*, *влезет*, *влезете*, *влезешь*, *влезла*, *влезли*, *влезло*, *влезть*, *влезу*, *влезут*, *влезши*, *влезший*, *влезь*, *влезьте* — “meddle” (7a);

ведём, *ведённый*, *ведёт*, *ведёте*, *ведёшь*, *веди*, *ведите*, *ведомый*, *веду*, *ведут*, *ведущий*, *ведши*, *ведший*, *ведя*, *вела*, *вели*, *вело*, *вести*, *поведём*, *поведённый*, *поведёт*, *поведёте*, *поведёшь*, *поведи*, *поведите*, *поведу*, *поведут*, *поведши*, *поведший*, *поведя*, *повела*, *повели*, *повело*, *повести* — “lead” (7b);

вытек, *вытекав*, *вытекавши*, *вытекавший*, *вытекаем*, *вытекает*, *вытекаете*, *вытекаешь*, *вытекай*, *вытекайте*, *вытекал*, *вытекала*, *вытекали*, *вытекало*, *вытекать*, *вытекаю*, *вытекают*, *вытекающий*, *вытекая*, *вытеки*, *вытеките*, *вытекла*, *вытекли*, *вытекло*, *вытеку*, *вытекут*, *вытекши*, *вытекший*, *вытечем*, *вытечет*, *вытечете*, *вытечешь*, *вытечь*, *теки*, *теките*, *текла*, *текли*, *текло*, *теку*, *текут*, *текущий*, *течём*, *течёт*, *течёте*, *течёшь*, *течь*, *тёк*, *тёкши*, *тёкший* — “leak” (8a);

береги, *берегите*, *берегла*, *берегли*, *берегло*, *берегу*, *берегут*, *берегущий*, *бережём*, *бережённый*, *бережёт*, *бережёте*, *бережёшь*, *беречь*, *берёг*, *берёгши*, *берёгший*, *сбереги*, *сберегите*, *сберегла*, *сберегли*, *сберегло*, *сберегу*, *сберегут*, *сбережём*, *сбережённый*, *сбережёт*, *сбережёте*, *сбережёшь*, *сберечь*, *сберёг*, *сберёгши*, *сберёгший*, *убереги*, *уберегите*, *уберегла*, *уберегли*, *уберегло*, *уберегу*, *уберегут*, *убережём*, *убережённый*, *убережёт*, *убережёте*, *убережёшь*, *уберечь*, *уберёг*, *уберёгши*, *уберёгший* — “guard” (8b);

вымер, *вымерев*, *вымереть*, *вымерла*, *вымерли*, *вымерло*, *вымерши*, *вымерший*, *вымирав*, *вымиравши*, *вымиравший*, *вымираем*, *вымирает*, *вымираете*, *вымираешь*, *вымирай*, *вымирайте*, *вымирал*, *вымирала*, *вымирали*, *вымирало*, *вымирать*, *вымираю*, *вымирают*, *вымирающий*, *вымирая*, *вымрем*, *вымрет*, *вымрете*, *вымрешь*, *вымри*, *вымрите*, *вымру*, *вымрут* — “becomeextinct” (9a);

потерев, *потереть*, *потёр*, *потёрла*, *потёрли*, *потёрло*, *потёртый*, *потёрши*, *потёрший*, *потрём*, *потрёт*, *потрёте*, *потрёшь*, *потри*, *потрите*, *потру*, *потрут*, *тереть*, *тёр*, *тёрла*, *тёрли*, *тёрло*, *тёртый*, *тёрши*, *тёрший*, *трём*, *трёт*, *трёте*, *трёшь*, *три*, *трите*, *тру*, *трут*, *трущий* — “rub” (9b);

выпалывав, *выпалывавши*, *выпалывавший*, *выпалываем*, *выпалываемый*, *выпалывает*, *выпалываете*, *выпалываешь*, *выпалывай*, *выпалывайте*, *выпалывал*, *выпалывала*, *выпалывали*, *выпалывало*, *выпалывать*, *выпалываю*, *выпалывают*, *выпалывающий*, *выпалывая*, *выполем*, *выполет*, *выполете*, *выполешь*, *выполи*, *выполите*, *выполов*, *выполовши*, *выполовший*, *выполол*, *выполола*, *выпололи*, *выпололо*, *выполотый*, *выполоть*, *выполю*, *выполют* — “weed” (10a);

вкалывав, *вкалывавши*, *вкалывавший*, *вкалываем*, *вкалываемый*, *вкалывает*, *вкалываете*, *вкалываешь*, *вкалывай*, *вкалывайте*, *вкалывал*, *вкалывала*, *вкалывали*, *вкалывало*, *вкалывать*, *вкалываю*, *вкалывают*, *вкалывающий*, *вкалывая*, *вколем*, *вколет*, *вколете*, *вколешь*, *вколи*, *вколите*, *вколов*, *вколовши*, *вколовший*, *вколол*, *вколола*, *вкололи*, *вкололо*, *вколотый*, *вколоть*, *вколю*, *вколют* — “stick” (10c);

вылейся, *вылейтесь*, *выливавшийся*, *выливавшись*, *выливаемся*, *выливаетесь*, *выливается*, *выливаешься*, *выливайся*, *выливайтесь*, *выливалась*, *выливались*, *выливалось*, *выливался*, *выливаться*, *выливаюсь*, *выливаются*, *выливающийся*, *выливаясь*, *вылившийся*, *вылившись*, *вылилась*, *вылились*, *вылилось*, *вылился*, *вылиться*, *выльемся*, *выльетесь*, *выльется*, *выльешься*, *выльюсь*, *выльются* — “spill over” (11a);

добей, *добейте*, *добив*, *добивав*, *добивавши*, *добивавший*, *добиваем*, *добиваемый*, *добивает*, *добиваете*, *добиваешь*, *добивай*, *добивайте*, *добивал*, *добивала*, *добивали*, *добивало*, *добивать*, *добиваю*, *добивают*, *добивающий*, *добивая*, *добивши*, *добивший*, *добил*, *добила*, *добили*, *добило*, *добитый*, *добить*, *добьём*, *добьёт*, *добьёте*, *добьёшь*, *добью*, *добьют* — “crush” (11b);

вымоем, *вымоет*, *вымоете*, *вымоешь*, *вымой*, *вымойте*, *вымою*, *вымоют*, *вымыв*, *вымывши*, *вымывший*, *вымыл*, *вымыла*, *вымыли*, *вымыло*, *вымытый*, *вымыть*, *моем*, *моемый*, *моет*, *моете*, *моешь*, *мой*, *мойте*, *мою*, *моют*, *моющий*, *моя*, *мыв*, *мывши*, *мывший*, *мыл*, *мыла*, *мыли*, *мыло*, *мытый*, *мыть*, *помоем*, *помоет*, *помоете*, *помоешь*, *помой*, *помойте*, *помою*, *помоют*, *помыв*, *помывши*, *помывший*, *помыл*, *помыла*, *помыли*, *помыло*, *помытый*, *помыть*, *умоем*, *умоет*, *умоете*, *умоешь*, *умой*, *умойте*, *умою*, *умоют*, *умыв*, *умывши*, *умывший*, *умыл*, *умыла*, *умыли*, *умыло*, *умытый*, *умыть* — “wash” (12a);

гнив, *гнивши*, *гнивший*, *гниём*, *гниёт*, *гниёте*, *гниёшь*, *гнил*, *гнила*, *гнили*, *гнило*, *гнить*, *гнию*, *гниют*, *гнющий*, *сгнив*, *сгнивши*, *сгнивший*, *сгниём*, *сгниёт*, *сгниёте*, *сгниёшь*, *сгнил*, *сгнила*, *сгнили*, *сгнило*, *сгнить*, *сгнию*, *сгниют* — “putrefy” (12b);

дав, *давав*, *дававши*, *дававший*, *даваемый*, *давай*, *давайте*, *давал*, *давала*, *давали*, *давало*, *давать*, *давая*, *давши*, *давший*, *дадим*, *дадите*, *дадут*, *даём*, *даёт*, *даёте*, *даёшь*, *дай*, *дайте*, *дал*, *дала*, *дали*, *дало*, *дам*, *данный*, *даст*, *дать*, *дашь*, *даю*, *дают*, *дающий* — “give” (13);

выжав, *выжавши*, *выжавший*, *выжал*, *выжала*, *выжали*, *выжало*, *выжатый*, *выжать*, *выжимав*, *выжимавши*, *выжимавший*, *выжимаем*, *выжимаемый*, *выжимает*, *выжимаете*, *выжимаешь*, *выжимай*, *выжимайте*, *выжимал*, *выжимала*, *выжимали*, *выжимало*, *выжимать*, *выжимаю*, *выжимают*, *выжимающий*, *выжимая*, *выжмем*, *выжмет*, *выжмете*, *выжмешь*, *выжми*, *выжмите*, *выжму*, *выжмут* — “wring” (14a);

жав, *жавши*, *жавший*, *жал*, *жала*, *жали*, *жало*, *жатый*, *жать*, *жмём*, *жмёт*, *жмёте*, *жмёшь*, *жми*, *жмите*, *жму*, *жмут*, *жмущий*, *жмя*, *пожав*, *пожавши*, *пожавший*, *пожал*, *пожала*, *пожали*, *пожало*, *пожатый*, *пожать*, *пожмём*, *пожмёт*, *пожмёте*, *пожмёшь*, *пожми*, *пожмите*, *пожму*, *пожмут*, *сжав*, *сжавши*, *сжавший*, *сжал*, *сжала*, *сжали*, *сжало*, *сжатый*, *сжать*, *сожмём*, *сожмёт*, *сожмёте*, *сожмёшь*, *сожми*, *сожмите*, *сожму*, *сожмут* — “squeeze” (14b);

обнимав, *обнимавши*, *обнимавший*, *обнимаем*, *обнимаемый*, *обнимает*, *обнимаете*, *обнимаешь*, *обнимай*, *обнимайте*, *обнимал*, *обнимала*, *обнимали*, *обнимало*, *обнимать*, *обнимаю*, *обнимают*, *обнимающий*, *обнимая*, *обнимем*, *обнимет*, *обнимете*, *обнимешь*, *обними*, *обнимите*, *обниму*, *обнимут*, *обняв*, *обнявши*, *обнявший*, *обнял*, *обняла*, *обняли*, *обняло*, *обнятый*, *обнять* — "hug" (14c);

встав, *вставав*, *встававши*, *встававший*, *вставай*, *вставайте*, *вставал*, *вставала*, *вставали*, *вставало*, *вставать*, *вставая*, *вставши*, *вставший*, *встаём*, *встаёт*, *встаёте*, *встаёшь*, *встал*, *встала*, *встали*, *встало*, *встанем*, *встанет*, *встанете*, *встанешь*, *встану*, *встанут*, *встань*, *встаньте*, *встать*, *встаю*, *встают*, *встающий* — "rise" (15);

выжив, *выживав*, *выживавши*, *выживавший*, *выживаем*, *выживает*, *выживаете*, *выживаешь*, *выживай*, *выживайте*, *выживал*, *выживала*, *выживали*, *выживало*, *выживать*, *выживаю*, *выживают*, *выживающий*, *выживая*, *выживем*, *выживет*, *выживете*, *выживешь*, *выживи*, *выживите*, *выживу*, *выживут*, *выживши*, *выживший*, *выжил*, *выжила*, *выжили*, *выжило*, *выжить* — "survive" (16a);

жив, *живём*, *живёт*, *живёте*, *живёшь*, *живи*, *живите*, *живу*, *живут*, *живущий*, *живши*, *живший*, *живя*, *жил*, *жила*, *жили*, *жило*, *жить*, *пожив*, *поживём*, *поживёт*, *поживёте*, *поживёшь*, *пожили*, *поживите*, *поживу*, *поживут*, *поживши*, *поживший*, *пожил*, *пожила*, *пожили*, *пожило*, *пожить* — "live" (16b).

These verbs are clustered by our algorithms as follows:

{*береги*, *берегите*, *берегла*, *берегли*, *берегло*, *берегу*, *берегут*, *берегущий*, *бережём*, *бережённый*, *бережёт*, *бережёте*, *бережёшь*, *беречь*, *берёг*, *берёгши*, *берёгший*, *сбереги*, *сберегите*, *сберегла*, *сберегли*, *сберегло*, *сберегу*, *сберегут*, *сбережём*, *сбережённый*, *сбережёт*, *сбережёте*, *сбережёшь*, *сберечь*, *сберёг*, *сберёгши*, *сберёгший*, *убереги*, *уберегите*, *уберегла*, *уберегли*, *уберегло*, *уберегу*, *уберегут*, *убережём*, *убережённый*, *убережёт*, *убережёте*, *убережёшь*, *уберечь*, *уберёг*, *уберёгши*, *уберёгший*},

{*блевав*, *блевавши*, *блевавший*, *блевал*, *блевала*, *блевали*, *блевало*, *блеванём*, *блеванёт*, *блеванёте*, *блеванёшь*, *блевани*, *блеваните*, *блевану*, *блеванув*, *блеванувши*, *блеванувший*, *блеванул*, *блеванула*, *блеванули*, *блевануло*, *блеванут*, *блевануть*, *блевать*, *блёванный*, *блюём*, *блюёт*, *блюёте*, *блюёшь*, *блюй*, *блюйте*, *блюю*, *блюют*, *блюющий*, *блюя*},

{*бренча*, *бренчав*, *бренчавши*, *бренчавший*, *бренчал*, *бренчала*, *бренчали*, *бренчало*, *бренчат*, *бренчать*, *бренчащий*, *бренчи*, *бренчим*, *бренчит*, *бренчите*, *бренчишь*, *бренчу*},

{*ведём*, *ведённый*, *ведёт*, *ведёте*, *ведёшь*, *веди*, *ведите*, *ведомый*, *ведши*, *ведший*, *ведя*, *вела*, *вели*, *вело*, *вести*, *поведём*, *поведённый*, *поведёт*, *поведёте*, *поведёшь*, *поведи*, *поведите*, *поведу*, *поведут*, *поведши*, *поведший*, *поведя*, *повела*, *повели*, *повело*, *повести*},

{*веду*, *ведут*, *ведущий*},

{*веем*, *веемый*, *веете*, *веешь*, *вей*, *вейте*, *вею*, *веют*, *веющий*, *вея*, *веяв*, *веявши*, *веявший*, *веял*, *веяла*, *веяли*, *веяло*, *веянный*, *веять*},

{*веет*},

{*взглядывав*, *взглядывавши*, *взглядывавший*, *взглядываем*, *взглядываемый*, *взглядывает*, *взглядываете*, *взглядываешь*, *взглядывай*, *взглядывайте*, *взглядывал*, *взглядывала*, *взглядывали*, *взглядывало*, *взглядывать*, *взглядываю*, *взглядывают*, *взглядывающий*, *взглядывая*, *взглянем*, *взглянет*, *взглянете*, *взглянешь*, *взгляни*, *взгляните*, *взгляну*, *взглянув*, *взглянувши*, *взглянувший*, *взглянул*, *взглянула*, *взглянули*, *взглянуло*, *взглянут*, *взглянуть*, *глядев*, *глядевши*, *глядевший*, *глядел*, *глядела*, *глядели*, *глядело*, *глядеть*, *гляди*, *глядим*, *глядит*, *глядите*, *глядишь*, *глядя*, *глядят*, *глядящий*, *гляжу*},

{*взывав*, *взывавши*, *взывавший*, *взываем*, *взывает*, *взываете*, *взываешь*, *взывай*, *взывайте*, *взывал*, *взывала*, *взывали*, *взывало*, *взывать*, *взываю*, *взывают*, *взывающий*, *взывая*, *взыщем*, *взыщет*, *взыщете*, *взыщешь*, *взыщи*, *взыщите*, *взыщу*, *взыщут*},

{*взыскав*, *взыскавши*, *взыскавший*, *взысканный*, *взыскать*, *взыскивав*, *взыскивавши*, *взыскивавший*, *взыскиваем*, *взыскиваемый*, *взыскивает*, *взыскиваете*, *взыскиваешь*, *взыскивай*, *взыскивайте*, *взыскивал*, *взыскивала*, *взыскивали*, *взыскивало*, *взыскивать*, *взыскиваю*, *взыскивают*, *взыскивающий*, *взыскивая*},

{*взыскал*, *взыскала*, *взыскали*, *взыскало*},

{вкалывав, вкалывавши, вкалывавший, вкалываем, вкалываемый, вкалывает, вкалываете, вкалываешь, вкалывай, вкалывайте, вкалывал, вкалывала, вкалывали, вкалывало, вкалывать, вкалываю, вкалывают, вкалывающий, вкалывая, вколем, вколет, вколете, вколешь, вколи, вколите, вколов, вколовши, вколовший, вколол, вколола, вкололи, вкололо, вколотый, вколоть, вколю, вколют},

{влез, влезав, влезавши, влезавший, влезаем, влезает, влезаете, влезаешь, влезай, влезайте, влезал, влезала, влезали, влезало, влезать, влезаю, влезают, влезающий, влезая, влезем, влезет, влезете, влезешь, влезла, влезли, влезло, влезть, влезу, влезут, влезши, влезший, влезь, влезьте},

{воззвав, воззвавши, воззвавший, воззвал, воззвала, воззвали, воззвало, воззванный, воззвать, воззовём, воззовёт, воззовёте, воззовёшь, воззови, воззовите, воззову, воззовут},

{встав, вставав, встававши, встававший, вставай, вставайте, вставал, вставала, вставали, вставало, вставать, вставая, вставши, вставший, встаём, встаёт, встаёте, встаёшь, встал, встала, встали, встало, встанем, встанет, встанете, встанешь, встану, встанут, встань, встаньте, встать, встаю, встают, встающий},

{выжав, выжавши, выжавший, выжал, выжала, выжали, выжало, выжатый, выжать, выжив, выживав, выживавши, выживавший, выживаем, выживает, выживаете, выживаешь, выживай, выживайте, выживал, выживала, выживали, выживало, выживать, выживаю, выживают, выживающий, выживая, выживем, выживет, выживете, выживешь, выживи, выживите, выживу, выживут, выживши, выживший, выжил, выжила, выжили, выжило, выжимав, выжимавши, выжимавший, выжимаем, выжимаемый, выжимает, выжимаете, выжимаешь, выжимай, выжимайте, выжимал, выжимала, выжимали, выжимало, выжимать, выжимаю, выжимают, выжимающий, выжимая, выжить, выжмем, выжмет, выжмете, выжмешь, выжми, выжмите, выжму, выжмут},

{вылейся, вылейтесь, выливавшийся, выливавшись, выливаемся, выливаетесь, выливается, выливаешься, выливайся, выливайтесь, выливалась, выливались, выливалось, выливался, выливаться, выливаюсь, выливаются, выливающийся, выливаясь, вылившийся, вылившись, вылилась, вылились, вылилось, вылился, вылиться, выльемся, выльетесь, выльется, выльешься, выльюсь, выльются},

{вымер, вымерев, вымереть, вымерла, вымерли, вымерло, вымерши, вымерший, вымирав, вымиравши, вымиравший, вымираем, вымирает, вымираете, вымираешь, вымирай, вымирайте, вымирал, вымирала, вымирали, вымирало, вымирать, вымираю, вымирают, вымирающий, вымирая, вымрем, вымрет, вымрете, вымрешь, вымри, вымрите, вымру, вымрут},

{вымоем, вымоет, вымоете, вымоешь, вымой, вымойте, вымою, вымоют, вымыв, вымывши, вымывший, вымыл, вымыла, вымыли, вымыло, вымытый, вымыть, моем, моемый, моет, моете, моешь, мой, мойте, мою, моют, моющий, моя, мыв, мывши, мывший, мыл, мыла, мыли, мыло, мытый, мыть, помоем, помоет, помоете, помоешь, помой, помойте, помою, помоют, помыв, помывши, помывший, помыл, помыла, помыли, помыло, помытый, помыть, умоем, умоет, умоете, умоешь, умой, умойте, умою, умоют, умыв, умывши, умывший, умыл, умыла, умыли, умыло, умытый, умыть},

{выпалывав, выпалывавши, выпалывавший, выпалываем, выпалываемый, выпалывает, выпалываете, выпалываешь, выпалывай, выпалывайте, выпалывал, выпалывала, выпалывали, выпалывало, выпалывать, выпалываю, выпалывают, выпалывающий, выпалывая, выполем, выполет, выполете, выполешь, выполи, выполите, выполов, выполовши, выполовший, выполол, выполола, выпололи, выпололо, выполотый, выполоть, выполю, выполют},

{вытек, вытекав, вытекавши, вытекавший, вытекаем, вытекает, вытекаете, вытекаешь, вытекай, вытекайте, вытекал, вытекала, вытекали, вытекало, вытекать, вытекаю, вытекают, вытекающий, вытекая, вытеки, вытеките, вытекла, вытекли, вытекло, вытеку, вытекут, вытекши, вытекший, вытечем, вытечет, вытечете, вытечешь, вытечь, теки, теките, текла, текли, текло, теку, текут, текущий, течём, течёт, течёте, течёшь, течь, тёк, тёкши, тёкший},

{гиб, гибла, гибли, гибло, гибнем, гибнет, гибнете, гибнешь, гибни, гибните, гибну, гибнув, гибнувши, гибнувший, гибнул, гибнут, гибнуть, гибнущий, погиб, погибла, погибли, погибло, погибнем, погибнет, погибнете, погибнешь, погибни, погибните, погибну, погибнут, погибнуть, погибши, погибший},

{гнив, гнивши, гнивший, гниём, гниёт, гниёте, гниёшь, гнил, гнила, гнили, гнило, гнить, гнию, гниют, гнющий, сгнив, сгнивши, сгнивший, сгниём, сгниёт, сгниёте, сгниёшь, сгнил, сгнила, сгнили, сгнило, сгнить, сгнию, сгниют},

{*дав*, *давав*, *дававши*, *дававший*, *даваемый*, *давай*, *давайте*, *давал*, *давала*, *давали*, *давало*, *давать*, *давая*, *давши*, *давший*, *дадим*, *дадите*, *дадут*, *даём*, *даёт*, *даёте*, *даёшь*, *дай*, *дайте*, *дал*, *дала*, *дали*, *дало*, *дам*, *данный*, *даст*, *дать*, *дашь*, *даю*, *дают*, *дающий*},

{*делав*, *делавши*, *делавший*, *делаем*, *делаемый*, *делает*, *делаете*, *делаешь*, *делай*, *делайте*, *делал*, *делала*, *делали*, *делало*, *деланный*, *делать*, *делаю*, *делают*, *делающий*, *делая*, *сделав*, *сделавши*, *сделавший*, *сделаем*, *сделает*, *сделаете*, *сделаешь*, *сделай*, *сделайте*, *сделал*, *сделала*, *сделали*, *сделало*, *сделанный*, *сделать*, *сделаю*, *сделают*},

{*добей*, *добейте*, *добив*, *добивав*, *добивавши*, *добивавший*, *добиваем*, *добиваемый*, *добивает*, *добиваете*, *добиваешь*, *добивай*, *добивайте*, *добивал*, *добивала*, *добивали*, *добивало*, *добивать*, *добиваю*, *добивают*, *добивающий*, *добивая*, *добивши*, *добивший*, *добил*, *добила*, *добили*, *добило*, *добитый*, *добить*, *добьём*, *добьёт*, *добьёте*, *добьёшь*, *добью*, *добьют*},

{*жав*, *жавши*, *жавший*, *жал*, *жала*, *жаленный*, *жали*, *жалив*, *жалил*, *жалила*, *жалили*, *жалило*, *жалим*, *жалимый*, *жалит*, *жалите*, *жалить*, *жалишь*, *жало*, *жаль*, *жальте*, *жалю*, *жаля*, *жалят*, *жалящий*, *жатый*, *жать*, *жмём*, *жмёт*, *жмёте*, *жмёшь*, *жми*, *жмите*, *жму*, *жмут*, *жмущий*, *жмя*, *пожав*, *пожавши*, *пожавший*, *пожал*, *пожала*, *пожали*, *пожало*, *пожатый*, *пожать*, *пожмём*, *пожмёт*, *пожмёте*, *пожмёшь*, *пожми*, *пожмите*, *пожму*, *пожмут*, *сжав*, *сжавши*, *сжавший*, *сжал*, *сжала*, *сжали*, *сжало*, *сжатый*, *сжать*, *сожмём*, *сожмёт*, *сожмёте*, *сожмёшь*, *сожми*, *сожмите*, *сожму*, *сожмут*},

{*жаливши*, *жаливший*, *ужаливши*, *ужаливший*},

{*жив*, *живём*, *живёт*, *живёте*, *живёшь*, *живи*, *живите*, *живу*, *живут*, *живущий*, *живши*, *живший*, *живя*, *жил*, *жила*, *жили*, *жило*, *жить*, *пожив*, *поживём*, *поживёт*, *поживёте*, *поживёшь*, *поживи*, *поживите*, *поживу*, *поживут*, *поживши*, *поживший*, *пожил*, *пожила*, *пожили*, *пожило*, *пожить*},

{*изгнав*, *изгнавши*, *изгнавший*, *изгнал*, *изгнала*, *изгнали*, *изгнало*, *изгнанный*, *изгнать*, *изгони*, *изгоним*, *изгонит*, *изгоните*, *изгонишь*, *изгоню*, *изгоняв*, *изгонявши*, *изгонявший*, *изгоняем*, *изгоняемый*, *изгоняет*, *изгоняете*, *изгоняешь*, *изгоняй*, *изгоняйте*, *изгонял*, *изгоняла*, *изгоняли*, *изгоняло*, *изгонят*, *изгонять*, *изгоняю*, *изгоняют*, *изгоняющий*, *изгоняя*},

{*люби*, *любив*, *любивши*, *любивший*, *любил*, *любила*, *любили*, *любило*, *любим*, *любимый*, *любит*, *любите*, *любить*, *любишь*, *любленный*, *люблю*, *любя*, *любят*, *любящий*, *полюби*, *полюбив*, *полюбивши*, *полюбивший*, *полюбил*, *полюбила*, *полюбили*, *полюбило*, *полюбим*, *полюбит*, *полюбите*, *полюбить*, *полюбишь*, *полюбленный*, *полюблю*, *полюбят*},

{*нарисовав*, *нарисовавши*, *нарисовавший*, *нарисовал*, *нарисовала*, *нарисовали*, *нарисовало*, *нарисованный*, *нарисовать*, *нарисуем*, *нарисует*, *нарисуете*, *нарисуешь*, *нарисуй*, *нарисуйте*, *нарисую*, *нарисуют*, *рисовав*, *рисовавши*, *рисовавший*, *рисовал*, *рисовала*, *рисовали*, *рисовало*, *рисованный*, *рисовать*, *рисуем*, *рисуемый*, *рисует*, *рисуете*, *рисуешь*, *рисуй*, *рисуйте*, *рисую*, *рисуют*, *рисующий*, *рисуя*},

{*обнимав*, *обнимавши*, *обнимавший*, *обнимаем*, *обнимаемый*, *обнимает*, *обнимаете*, *обнимаешь*, *обнимай*, *обнимайте*, *обнимал*, *обнимала*, *обнимали*, *обнимало*, *обнимать*, *обнимаю*, *обнимают*, *обнимающий*, *обнимая*, *обнимем*, *обнимет*, *обнимете*, *обнимешь*, *обними*, *обнимите*, *обниму*, *обнимут*, *обняв*, *обнявши*, *обнявший*, *обнял*, *обняла*, *обняли*, *обняло*, *обнятый*, *обнять*},

{*потерев*, *потереть*, *потёр*, *потёрла*, *потёрли*, *потёрло*, *потёртый*, *потёрши*, *потёрший*, *потрём*, *потрёт*, *потрёте*, *потрёшь*, *потри*, *потрите*, *потру*, *потрут*, *тереть*, *тёр*, *тёрла*, *тёрли*, *тёрло*, *тёртый*, *тёрши*, *тёрший*},

{*пощади*, *пощадив*, *пощадивши*, *пощадивший*, *пощадил*, *пощадила*, *пощадили*, *пощадило*, *пощадим*, *пощадит*, *пощадите*, *пощадить*, *пощадишь*, *пощадят*, *пощажённый*, *пощажу*, *щади*, *щадив*, *щадивши*, *щадивший*, *щадил*, *щадила*, *щадили*, *щадило*, *щадим*, *щадимый*, *щадит*, *щадите*, *щадить*, *щадишь*, *щадя*, *щадят*, *щадящий*, *щажённый*, *щажу*},

{*рискнём*, *рискнёт*, *рискнёте*, *рискнёшь*, *рискни*, *рискните*, *рискну*, *рискнув*, *рискнувши*, *рискнувший*, *рискнул*, *рискнула*, *рискнули*, *рискнуло*, *рискнут*, *рискнуть*, *рисковав*, *рисковавши*, *рисковавший*, *рисковал*, *рисковала*, *рисковали*, *рисковало*, *рисковать*, *рискуем*, *рискует*, *рискуете*, *рискуешь*, *рискуй*, *рискуйте*, *рискую*, *рискуют*, *рискующий*, *рискуя*},

{*слыша*, *слышав*, *слышавши*, *слышавший*, *слышал*, *слышала*, *слышали*, *слышало*, *слышанный*, *слышат*, *слышать*, *слышащий*, *слышим*, *слышимый*, *слышит*, *слышите*, *слышишь*, *слышу*, *услышав*, *услышавши*, *услышавший*, *услышал*, *услышала*, *услышали*, *услышало*, *услышанный*, *услышат*, *услышать*, *услышим*, *услышит*, *услышите*, *услышишь*, *услышу*, *услышь*, *услышьте*},

{*трём*, *трёт*, *трёте*, *трёшь*, *три*, *трите*, *тру*, *трут*, *трущий*},

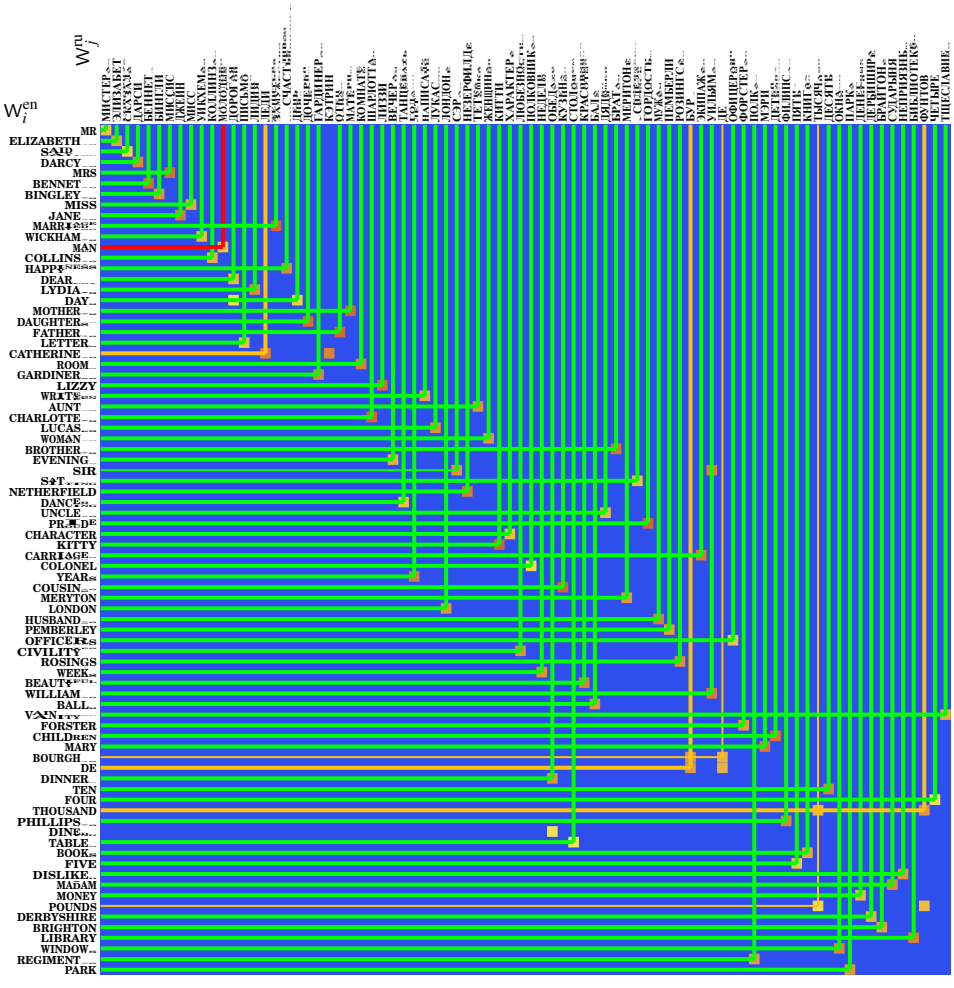
{*ужаленный*, *ужалив*, *ужалил*, *ужалила*, *ужалили*, *ужалило*, *ужалим*, *ужалит*, *ужалите*, *ужалить*, *ужалишь*, *ужаль*, *ужальте*, *ужалю*, *ужалят*},

Please beware that the verb clustering example above contains a few notable errors, partly due to interference between verbs with similar-looking stems.

Example 7.26.3. In Fig. S11, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text sources).

In Russian, “bat” (a kind of flying mammal) is called *летучая мышь* (literally “flying mouse”), so we consider *bat* an exact match to *летучая* “flying” in Fig. S11b″.

(a)



(b)

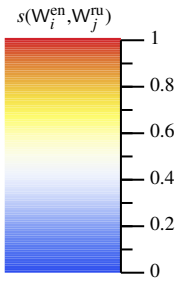


Fig. S11. Text mining in Russian. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Russian version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(\mathrm{W}_i^{\mathrm{en}}, \mathrm{W}_j^{\mathrm{ru}})$ between selected topics in English and Russian versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.

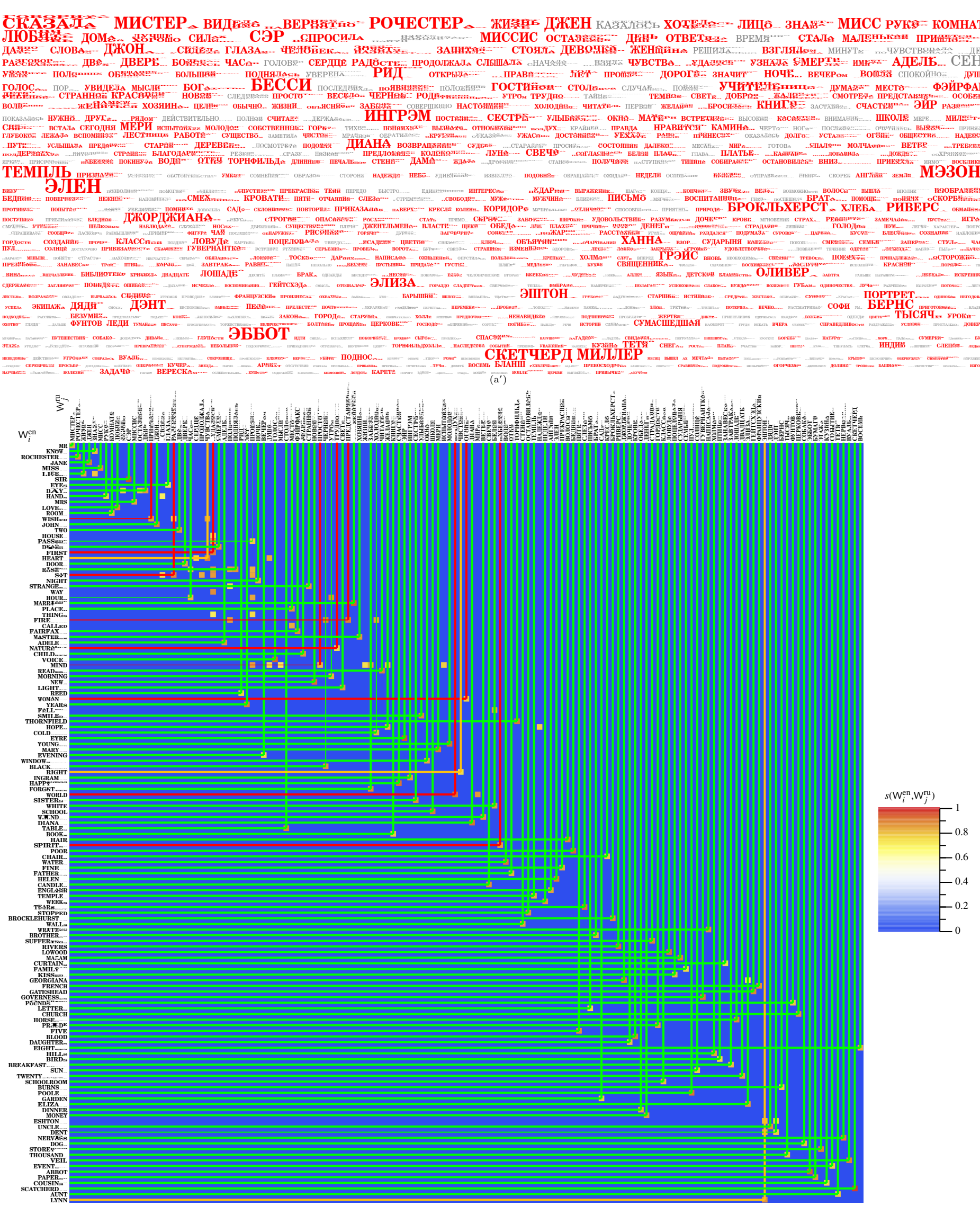


Fig. S11. Text mining in Russian. (*Continued*) (a′) Statistically identified topics ($n_{ii} \geq 20$) in a Russian version of *Jane Eyre*, with the same color encoding scheme as Fig. S3. (b′) Semantic similarities $s(W_i^{en}, W_j^{ru})$ between selected topics in English and Russian versions of *Jane Eyre*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c′) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

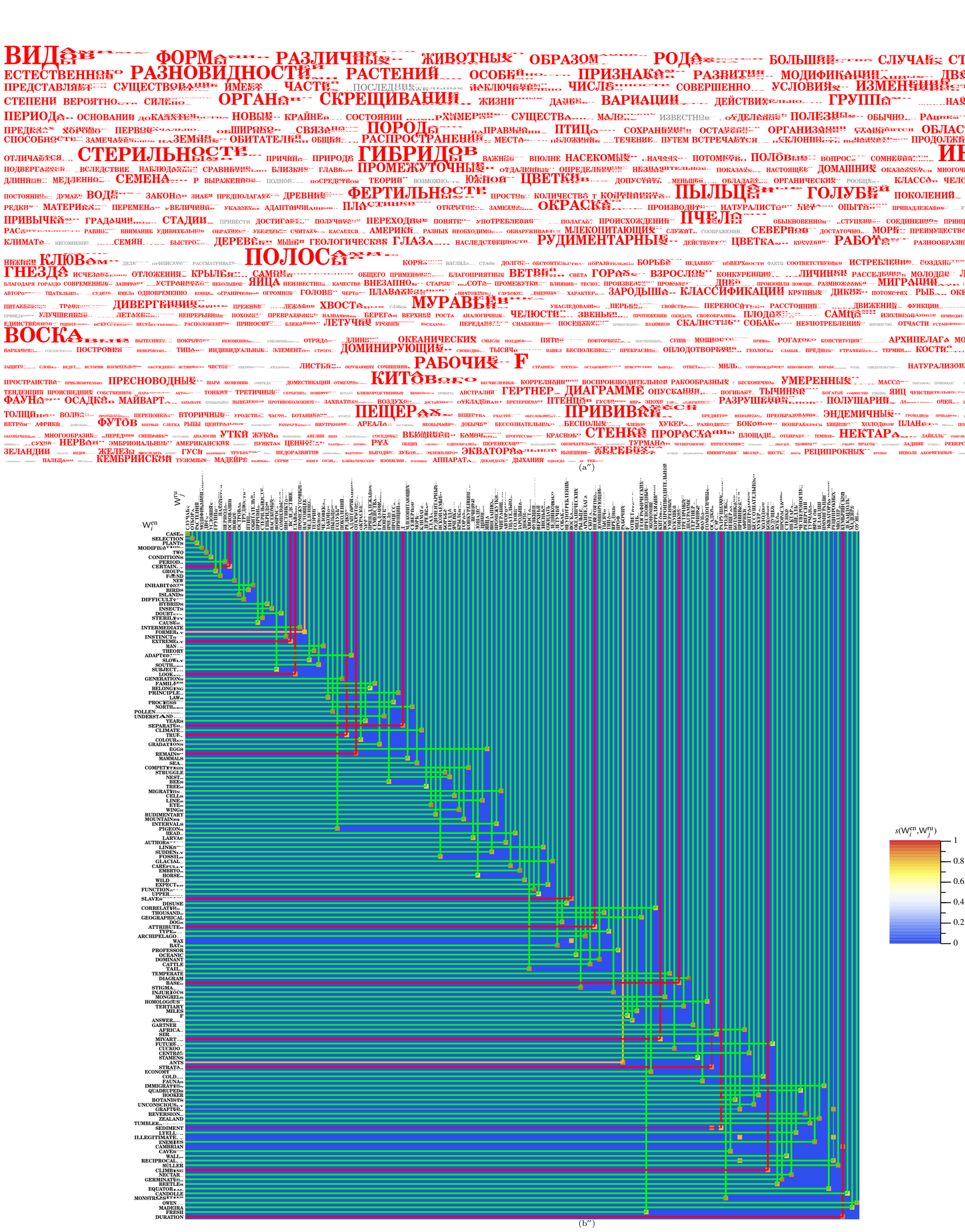


Fig. S11. Text mining in Russian. (*Continued*) (a″) Statistically identified topics ($n_{ii} \geq 20$) in a Russian version of *Origin of Species*, with the same color encoding scheme as Fig. S3. (b″) Semantic similarities $s(\mathrm{W}_i^{\mathrm{en}}, \mathrm{W}_j^{\mathrm{ru}})$ between selected topics in English and Russian versions of *Origin of Species*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c″) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

8 Approximate word clustering in selected Uralic languages

In this section, we present the approximate clustering algorithms for two representative Uralic languages: Finnish and Hungarian.

The major challenges in information processing for Finnish and Hungarian are the following:

- Both languages have extremely complex case systems (15 for Finnish and 18 for Hungarian, not counting rarely used cases), and the situation is compounded by the fact that several case endings may attach sequentially to the same word.
- Finnish has consonant gradation (mutation of stem-final consonants) during declension. (Estonian, another Uralic language not treated here, behaves similarly.) Hungarian case endings sometimes also cause changes in the stem-final consonants.
- Their verb morphologies have a similar complexity as the Romance languages.
- Similar to the Germanic counterpart, both Finnish and Hungarian orthographies require that certain compounds be written as single words, with no spaces or hyphens between their constituents.

Finnish and Hungarian are two major agglutinative languages spoken in Europe. Basque, a language isolate spoken in France and Spain, also has agglutinative morphology. The agglutinative nature of Turkish (as well as other Turkic languages spoken in Eurasia) is similar to Finnish and Hungarian. We will cover the word clustering algorithms for Basque and Turkish in §9.

8.1 Modified Porter stemming algorithm for Finnish

Definition 8.1 (Finnish stop words)**.** If a word belongs to the following list[96]:

aika, *aina*, *ainainen*, *ainoa*, *ainoaa*, *ainoain*, *ainoan*, *ainoana*, *ainoat*, *ainoata*, *ainoiden*, *ainoihin*, *ainoiksi*, *ainoilla*, *ainoille*, *ainoilta*, *ainoin*, *ainoina*, *ainoine*, *ainoisiin*, *ainoissa*, *ainoista*, *ainoita*, *ainoitta*, *ainoitten*, *ainuiden*, *ainuihin*, *ainuiksi*, *ainuilla*, *ainuille*, *ainuilta*, *ainuin*, *ainuina*, *ainuineen*, *ainuisiin*, *ainuissa*, *ainuista*, *ainuita*, *ainuitta*, *ainuitten*, *ainut*, *ainutta*, *ainuuksi*, *ainuulla*, *ainuulle*, *ainuulta*, *ainuun*, *ainuuna*, *ainuuseen*, *ainuussa*, *ainuusta*, *ainuut*, *ainuutta*, *aivan*, *ala*, *alakkain*, *alapuolella*, *alas*, *alatusten*, *alhaalla*, *alhaalle*, *alhaalta*, *ali*, *alitse*, *alla*, *alle*, *allekkain*, *alta*, *ansiosta*, *asti*, *edelle*, *edellä*, *edeltä*, *edes*, *edessä*, *edestä*, *editse*, *eduksi*, *ehkä*, *ei*, *elative*, *eli*, *ellei*, *emme*, *en*, *enemmän*, *eniten*, *ennen*, *ennenkuin*, *ennestä*, *ennestään*, *ensi*, *ensin*, *entä*, *enää*, *erittäin*, *eräiden*, *eräihin*, *eräiksi*, *eräille*, *eräillä*, *eräiltä*, *eräin*, *eräine*, *eräinä*, *eräisiin*, *eräissä*, *eräistä*, *eräitten*, *eräittä*, *eräitä*, *eräs*, *erästä*, *erääksi*, *eräälle*, *eräällä*, *eräältä*, *erään*, *eräänä*, *erääseen*, *eräässä*, *eräästä*, *erät*, *eräättä*, *esi*, *esiin*, *et*, *eteemme*, *eteen*, *eteensä*, *etenkin*, *etenkään*, *ette*, *ettei*, *etten*, *ettet*, *että*, *halki*, *he*, *heidän*, *heidät*, *heihin*, *heille*, *heillä*, *heiltä*, *heissä*, *heistä*, *heitä*, *heti*, *hiukan*, *hyvin*, *hän*, *häneen*, *hänelle*, *hänellä*, *häneltä*, *hänessä*, *hänestä*, *hänet*, *häntä*, *ihan*, *ikään*, *ilman*, *itse*, *itseemme*, *itseeni*, *itseenne*, *itseensä*, *itseesi*, *itseksesi*, *itsellesi*, *itselläsi*, *itsellään*, *itseltäsi*, *itseltään*, *itsen*, *itsenä*, *itsenäsi*, *itsenään*, *itsesi*, *itsessäsi*, *itsessään*, *itsestäsi*, *itsestään*, *itseä*, *itseäsi*, *itseään*, *ja*, *jk*, *jkhun*, *jkn*, *jkta*, *jo*, *johdosta*, *johon*, *johonkuhun*, *joiden*, *joidenkuiden*, *joihin*, *joihinkuihin*, *joiksi*, *joiksikuiksi*, *joilla*, *joillain*, *joillakuilla*, *joille*, *joillekuille*, *joilta*, *joiltain*, *joiltakuilta*, *join*, *joina*, *joinain*, *joinakuina*, *joine*, *joinekuine*, *joinkuin*, *joissa*, *joissain*, *joissakuissa*, *joista*, *joistain*, *joistakuista*, *joita*, *joitain*, *joitakuita*, *joitse*, *joitta*, *joittakuitta*, *joitten*, *joittenkuitten*, *jokainen*, *jokaiseen*, *jokaiseksi*, *jokaisella*, *jokaiselle*, *jokaiselta*, *jokaisen*, *jokaisena*, *jokaisessa*, *jokaisesta*, *jokaiset*, *jokaisetta*, *jokaisia*, *jokaisien*, *jokaisiin*, *jokaisiksi*, *jokaisilla*, *jokaisille*, *jokaisilta*, *jokaisin*, *jokaisina*, *jokaisine*, *jokaisissa*, *jokaisista*, *jokaisitta*, *jokaista*, *jokaisten*, *joksikuksi*, *joku*, *jollain*, *jollakulla*, *jollei*, *jollekulle*, *jollen*, *jollet*, *jolloin*, *joltain*, *joltakulta*, *jommaksikummaksi*, *jommallakummalla*, *jommallekummalle*, *jommaltakummalta*, *jommankumman*, *jommassakummassa*, *jommastakummasta*, *jommatkummat*, *jommattakummatta*, *jommiksikummiksi*, *jommillakummilla*, *jommillekummille*, *jommiltakummilta*, *jomminkummin*, *jommissakummissa*, *jommistakummista*, *jommittakummitta*, *jompaakumpaa*, *jompaankumpaan*, *jompanakumpana*, *jompiakumpia*, *jompiinkumpiin*, *jompikumpi*, *jompinakumpina*, *jompinekumpine*, *jona*, *jonain*, *jonakuna*, *jonka*, *jonkin*, *jonkun*, *jos*, *joskus*, *jossain*, *jossakussa*, *jostain*, *jostakusta*, *jota*, *jotain*, *jotakuta*, *joten*, *jotenkuten*, *jotka*, *jotkin*, *jotkut*, *jottakutta*, *joukoon*, *joukossa*, *joukosta*, *juuri*, *jäljessä*, *jälkeen*, *jälleen*, *kai*, *kaikeksi*, *kaikella*, *kaikelle*, *kaikelta*, *kaiken*, *kaikessa*, *kaikesta*, *kaiket*, *kaiketta*, *kaikiksi*, *kaikilla*, *kaikille*, *kaikilta*, *kaikissa*, *kaikista*, *kaikitta*, *kaikkea*, *kaikkeen*, *kaikkein*, *kaikkena*, *kaikki*, *kaikkia*, *kaikkiin*, *kaikkina*, *kaikkine*, *kanssa*, *kautta*, *keiden*, *keihin*, *keiksi*, *keille*, *keillä*, *keiltä*, *keinä*, *keissä*, *keistä*, *keitä*, *keneen*, *keneksi*, *kenelle*, *kenellä*, *keneltä*, *kenen*, *kenenä*, *kenessä*, *kenestä*, *kenet*, *kenties*, *kerran*, *keskelle*,

[96] Our list of Finnish stop words is based on `http://snowball.tartarus.org/algorithms/finnish/stop.txt`, with extensive additions to roughly match their counterparts in English.

keskellä, *keskeltä*, *kesken*, *keskenään*, *keskeä*, *keski*, *keskitse*, *keskuudeksi*, *keskuudella*, *keskuudelle*, *keskuudelta*, *keskuuden*, *keskuudessa*, *keskuudesta*, *keskuudet*, *keskuudetta*, *keskuuksia*, *keskuuksien*, *keskuuksiin*, *keskuuksiksi*, *keskuuksilla*, *keskuuksille*, *keskuuksilta*, *keskuuksin*, *keskuuksina*, *keskuuksineen*, *keskuuksissa*, *keskuuksista*, *keskuuksitta*, *keskuus*, *keskuuteen*, *keskuutena*, *keskuutta*, *ketkä*, *ketä*, *kohdaksi*, *kohdalla*, *kohdalle*, *kohdalta*, *kohdan*, *kohdassa*, *kohdasta*, *kohdat*, *kohdatta*, *kohdiksi*, *kohdilla*, *kohdille*, *kohdilta*, *kohdin*, *kohdissa*, *kohdista*, *kohditta*, *kohta*, *kohtaa*, *kohtain*, *kohtana*, *kohti*, *kohtia*, *kohtiin*, *kohtina*, *kohtineen*, *koko*, *koska*, *kovin*, *kuhunkin*, *kuidenkin*, *kuihinkin*, *kuiksikin*, *kuillakin*, *kuillekin*, *kuiltakin*, *kuin*, *kuinakin*, *kuissakin*, *kuistakin*, *kuitakin*, *kuitenkin*, *kuittakin*, *kuittenkin*, *kuka*, *kukin*, *kuksikin*, *kullakin*, *kullekin*, *kulloinkin*, *kultakin*, *kummaksi*, *kummalla*, *kummalle*, *kummalta*, *kumman*, *kummassa*, *kummasta*, *kummat*, *kummatta*, *kummiksi*, *kummilla*, *kummille*, *kummilta*, *kummin*, *kummissa*, *kummista*, *kummitta*, *kumpaa*, *kumpaan*, *kumpain*, *kumpana*, *kumpi*, *kumpia*, *kumpiin*, *kumpina*, *kumpine*, *kun*, *kunakin*, *kunnekin*, *kunnes*, *kussakin*, *kustakin*, *kutakin*, *kuten*, *kutkin*, *kuttakin*, *kykene*, *kykenee*, *kykeneminen*, *kykenemistä*, *kykenemä*, *kykenemäisillään*, *kykenemän*, *kykenemättömien*, *kykenemättömiin*, *kykenemättömiksi*, *kykenemättömille*, *kykenemättömillä*, *kykenemättömiltä*, *kykenemättömin*, *kykenemättömine*, *kykenemättöminä*, *kykenemättömissä*, *kykenemättömistä*, *kykenemättömittä*, *kykenemättömiä*, *kykenemättömäksi*, *kykenemättömälle*, *kykenemättömällä*, *kykenemättömältä*, *kykenemättömän*, *kykenemättömänä*, *kykenemättömässä*, *kykenemättömästä*, *kykenemättömät*, *kykenemättömättä*, *kykenemättömään*, *kykenemätön*, *kykenemätönten*, *kykenemätöntä*, *kykenemään*, *kykenen*, *kykenet*, *kykenevien*, *kykeneviin*, *kykeneviksi*, *kykeneville*, *kykenevillä*, *kykeneviltä*, *kykenevin*, *kykenevine*, *kykenevinä*, *kykenevissä*, *kykenevistä*, *kykenevittä*, *kykeneviä*, *kykenevä*, *kykeneväin*, *kykenevän*, *kykenevänä*, *kykenevää*, *kykenevään*, *kykeni*, *kykenin*, *kykenisi*, *kykenisin*, *kykenisit*, *kykenit*, *kyllin*, *kyllä*, *lain*, *liene*, *lienee*, *lienen*, *lienet*, *liian*, *luo*, *luokse*, *luona*, *luota*, *lähekkäin*, *lähelle*, *lähellä*, *läheltä*, *lähes*, *lähetysten*, *lähi*, *lähitse*, *läpi*, *läsnä*, *lävitse*, *me*, *meidän*, *meidät*, *meihin*, *meille*, *meillä*, *meiltä*, *meissä*, *meistä*, *meitä*, *melko*, *mieluummin*, *mihin*, *miksi*, *mikä*, *mikäli*, *mikään*, *millainen*, *millaiseen*, *millaiseksi*, *millaisella*, *millaiselle*, *millaiselta*, *millaisen*, *millaisena*, *millaisessa*, *millaisesta*, *millaiset*, *millaisetta*, *millaisia*, *millaisien*, *millaisiin*, *millaisiksi*, *millaisilla*, *millaisille*, *millaisilta*, *millaisin*, *millaisina*, *millaisineenn*, *millaisissa*, *millaisista*, *millaisitta*, *millaista*, *millaisten*, *mille*, *milloin*, *millä*, *millään*, *miltä*, *miltään*, *minkä*, *minkälainen*, *minkälaiseen*, *minkälaiseksi*, *minkälaisella*, *minkälaiselle*, *minkälaiselta*, *minkälaisen*, *minkälaisena*, *minkälaisessa*, *minkälaisesta*, *minkälaiset*, *minkälaisetta*, *minkälaisia*, *minkälaisien*, *minkälaisiin*, *minkälaisiksi*, *minkälaisilla*, *minkälaisille*, *minkälaisilta*, *minkälaisin*, *minkälaisina*, *minkälaisineen*, *minkälaisissa*, *minkälaisista*, *minkälaisitta*, *minkälaista*, *minkälaisten*, *minkään*, *minne*, *minua*, *minulla*, *minulle*, *minulta*, *minun*, *minussa*, *minusta*, *minut*, *minuun*, *minä*, *minään*, *missä*, *missään*, *mistä*, *mistään*, *miten*, *mitkä*, *mitkään*, *mitä*, *mitään*, *moinen*, *moiseen*, *moiseksi*, *moisella*, *moiselle*, *moiselta*, *moisen*, *moisena*, *moisessa*, *moisesta*, *moiset*, *moisetta*, *moisia*, *moisien*, *moisiin*, *moisiksi*, *moisilla*, *moisille*, *moisilta*, *moisin*, *moisina*, *moisine*, *moisissa*, *moisista*, *moisitta*, *moista*, *moisten*, *molemmaksi*, *molemmalla*, *molemmalle*, *molemmalta*, *molemman*, *molemmassa*, *molemmasta*, *molemmat*, *molemmatta*, *molemmiksi*, *molemmilla*, *molemmille*, *molemmilta*, *molemmin*, *molemmissa*, *molemmista*, *molemmitta*, *molempaa*, *molempaan*, *molempana*, *molempi*, *molempia*, *molempiin*, *molempina*, *molempine*, *moneen*, *moneksi*, *monella*, *monelle*, *monelta*, *monen*, *monena*, *monessa*, *monesta*, *monesti*, *monet*, *monetta*, *moni*, *monia*, *moniaalla*, *moniaalle*, *moniaalta*, *moniin*, *monin*, *monina*, *monine*, *monna*, *monta*, *muiden*, *muihin*, *muiksi*, *muilla*, *muille*, *muilta*, *muin*, *muina*, *muine*, *muissa*, *muista*, *muita*, *muitta*, *muitten*, *mukaan*, *mukana*, *mutta*, *mutte*, *muttei*, *mutten*, *muttet*, *muu*, *muualla*, *muualle*, *muualta*, *muuhun*, *muulloin*, *muun*, *muuna*, *muut*, *muuta*, *muutama*, *muutamaa*, *muutamain*, *muutaman*, *muutamana*, *muutamat*, *muutamia*, *muutamien*, *muutamiin*, *muutamiksi*, *muutamilla*, *muutamille*, *muutamilta*, *muutamin*, *muutamina*, *muutamineen*, *muutamissa*, *muutamista*, *muutamitta*, *myös*, *myötä*, *ne*, *niiden*, *niihin*, *niiksi*, *niille*, *niillä*, *niiltä*, *niin*, *niine*, *niinkuin*, *niinä*, *niissä*, *niistä*, *niitten*, *niittä*, *niitä*, *no*, *noiden*, *noihin*, *noiksi*, *noilla*, *noille*, *noilta*, *noin*, *noina*, *noissa*, *noista*, *noita*, *nuo*, *nyt*, *näiden*, *näihin*, *näiksi*, *näille*, *näillä*, *näiltä*, *näin*, *näine*, *näinä*, *näissä*, *näistä*, *näittä*, *näitä*, *nämä*, *ohi*, *ohitse*, *oikein*, *oitis*, *ole*, *olema*, *olemaa*, *olemain*, *olemaisillaan*, *oleman*, *olemana*, *olemat*, *olematon*, *olematonta*, *olematonten*, *olemattomaan*, *olemattomaksi*, *olemattomalla*, *olemattomalle*, *olemattomalta*, *olemattoman*, *olemattomana*, *olemattomassa*, *olemattomasta*, *olemattomat*, *olemattomatta*, *olemattomia*, *olemattomien*, *olemattomiin*, *olemattomiksi*, *olemattomilla*, *olemattomille*, *olemattomilta*, *olemattomin*, *olemattomina*, *olemattomine*, *olemattomissa*, *olemattomista*, *olemattomitta*, *olemia*, *olemien*, *olemiin*, *olemiksi*, *olemilla*, *olemille*, *olemilta*, *olemin*, *olemina*, *olemineen*, *olemissa*, *olemista*, *olemitta*, *olen*, *olet*, *oleva*, *olevaa*, *olevain*, *olevan*, *olevana*, *olevasi*, *olevia*, *olevien*, *oleviin*, *oleviksi*, *olevilla*, *oleville*, *olevilta*, *olevin*, *olevina*, *olevine*, *olevissa*, *olevista*, *olevitta*, *oli*, *olin*, *olisi*, *olisin*, *olisit*, *olit*, *olkaa*, *olko*, *olkoon*, *olkoot*, *olla*, *olleeksi*, *olleella*, *olleelle*, *olleelta*, *olleen*, *olleena*, *olleeseen*, *olleessa*, *olleesta*, *olleet*, *olleetta*, *olleiden*, *olleihin*, *olleiksi*, *olleilla*, *olleille*, *olleilta*, *ollein*, *olleina*, *olleine*, *olleisiin*, *olleissa*, *olleista*, *olleita*, *olleitta*, *olleitten*, *ollen*, *ollessa*, *olluiksi*, *olluilla*, *olluille*, *olluilta*, *olluin*, *olluissa*, *olluista*, *olluitta*, *olluksi*, *ollulla*, *ollulle*, *ollulta*, *ollun*, *ollussa*, *ollusta*, *ollut*, *ollutta*, *oltaessa*, *oltaisi*, *oltaisiin*, *oltako*, *oltakoon*, *oltaman*, *oltane*, *oltava*, *oltiin*, *oltu*, *oltua*, *oltuihin*, *oltuina*, *oltuine*, *oltuja*, *oltujen*, *oltuna*, *oltuun*, *oma*, *omaa*, *omain*, *oman*, *omana*, *omat*, *omia*, *omien*, *omiin*, *omiksi*, *omilla*, *omille*, *omilta*, *omin*, *omina*, *omine*, *omissa*, *omista*, *omitta*, *on*, *ovat*, *paitsi*, *paljo*, *paljoa*, *paljoihin*, *paljoiksi*, *paljoilla*, *paljoille*, *paljoilta*, *paljoin*,

paljoina, *paljoineen*, *paljoissa*, *paljoista*, *paljoitta*, *paljoja*, *paljojen*, *paljon*, *paljona*, *paljoon*, *paljot*, *pelkiksi*, *pelkille*, *pelkillä*, *pelkiltä*, *pelkin*, *pelkissä*, *pelkistä*, *pelkittä*, *pelkkien*, *pelkkiin*, *pelkkine*, *pelkkinä*, *pelkkiä*, *pelkkä*, *pelkkäin*, *pelkkänä*, *pelkkää*, *pelkkään*, *pelkäksi*, *pelkälle*, *pelkällä*, *pelkältä*, *pelkän*, *pelkässä*, *pelkästä*, *pelkät*, *pelkättä*, *perässä*, *pian*, *pitkin*, *poikki*, *pois*, *poissa*, *pysty*, *pystyessä*, *pystyi*, *pystyin*, *pystyisi*, *pystyisin*, *pystyisit*, *pystyit*, *pystykää*, *pystyköön*, *pystyööt*, *pystyminen*, *pystymistä*, *pystymä*, *pystymäisillään*, *pystymän*, *pystymätön*, *pystymään*, *pystyn*, *pystyne*, *pystynee*, *pystyneenä*, *pystyneeseen*, *pystyneet*, *pystyneiden*, *pystyneihin*, *pystyneiksi*, *pystyneille*, *pystyneillä*, *pystyneiltä*, *pystynein*, *pystyneine*, *pystyneinä*, *pystyneisiin*, *pystyneissä*, *pystyneistä*, *pystyneitten*, *pystyneittä*, *pystyneitä*, *pystynet*, *pystynyt*, *pystynyttä*, *pystyt*, *pystyttiin*, *pystytty*, *pystyttäessä*, *pystyttäisi*, *pystyttäisiin*, *pystyttäkö*, *pystyttäköön*, *pystyttämän*, *pystyttäne*, *pystyttävä*, *pystytä*, *pystytään*, *pystyvä*, *pystyy*, *pystyä*, *päin*, *rinnalla*, *saa*, *saada*, *saaden*, *saadessa*, *saakaa*, *saakka*, *saakoon*, *saakoot*, *saama*, *saamaisillaan*, *saaman*, *saamaton*, *saaminen*, *saamista*, *saan*, *saane*, *saanee*, *saaneet*, *saanet*, *saanut*, *saat*, *saataessa*, *saataisi*, *saataisiin*, *saatako*, *saatakoon*, *saataman*, *saatane*, *saatava*, *saatiin*, *saatu*, *saava*, *sai*, *sain*, *saisi*, *saisin*, *saisit*, *sait*, *sama*, *samaa*, *saman*, *samana*, *samat*, *samoihin*, *samoiksi*, *samoilla*, *samoille*, *samoilta*, *samoin*, *samoina*, *samoine*, *samoissa*, *samoista*, *samoitta*, *samoja*, *samojen*, *sangen*, *se*, *sellainen*, *sellaiseen*, *sellaiseksi*, *sellaisella*, *sellaiselle*, *sellaiselta*, *sellaisen*, *sellaisena*, *sellaisessa*, *sellaisesta*, *sellaiset*, *sellaisetta*, *sellaisia*, *sellaisien*, *sellaisiin*, *sellaisiksi*, *sellaisilla*, *sellaisille*, *sellaisilta*, *sellaisin*, *sellaisina*, *sellaisineen*, *sellaisissa*, *sellaisista*, *sellaisitta*, *sellaista*, *sellaisten*, *semmoinen*, *semmoiseen*, *semmoiseksi*, *semmoisella*, *semmoiselle*, *semmoiselta*, *semmoisen*, *semmoisena*, *semmoisessa*, *semmoisesta*, *semmoiset*, *semmoisetta*, *semmoisia*, *semmoisien*, *semmoisiin*, *semmoisiksi*, *semmoisilla*, *semmoisille*, *semmoisilta*, *semmoisin*, *semmoisina*, *semmoisine*, *semmoisissa*, *semmoisista*, *semmoisitta*, *semmoista*, *semmoisten*, *sen*, *sentään*, *senvuoksi*, *siellä*, *sieltä*, *siihen*, *siinä*, *siis*, *siitä*, *sijaan*, *siksi*, *sille*, *silloin*, *sillä*, *siltä*, *sinne*, *sinua*, *sinulla*, *sinulle*, *sinulta*, *sinun*, *sinussa*, *sinusta*, *sinut*, *sinuun*, *sinä*, *sisä*, *sisäkkäin*, *sisätysten*, *sisään*, *siten*, *sitten*, *sitä*, *taa*, *taakse*, *taannoin*, *taas*, *tahansa*, *tahi*, *tai*, *taitse*, *taka*, *takaa*, *takaisin*, *takana*, *takia*, *te*, *tee*, *teet*, *tehden*, *tehdessä*, *tehdä*, *tehdään*, *tehkää*, *tehkö*, *tehköön*, *tehööt*, *tehne*, *tehnee*, *tehneet*, *tehnen*, *tehnet*, *tehnyt*, *tehtiin*, *tehty*, *tehtäessä*, *tehtäisi*, *tehtäisiin*, *tehtäkö*, *tehtäköön*, *tehtämän*, *tehtäne*, *tehtävä*, *teidän*, *teidät*, *teihin*, *teille*, *teillä*, *teiltä*, *teimme*, *tein*, *teissä*, *teistä*, *teit*, *teitte*, *teitä*, *tekee*, *tekeminen*, *tekemistä*, *tekemä*, *tekemäisillään*, *tekemän*, *tekemätön*, *tekemään*, *tekevä*, *tekevät*, *teki*, *tekisi*, *tekisin*, *tekisit*, *todella*, *toinen*, *toisaalla*, *toisaalle*, *toisaalta*, *toiseen*, *toiseksi*, *toisella*, *toiselle*, *toiselta*, *toisen*, *toisena*, *toisensa*, *toisessa*, *toisesta*, *toiset*, *toisetta*, *toisia*, *toisien*, *toisiensa*, *toisiin*, *toisiinsa*, *toisikseen*, *toisiksensa*, *toisiksi*, *toisilla*, *toisille*, *toisilta*, *toisin*, *toisina*, *toisine*, *toisinsa*, *toisissa*, *toisista*, *toisitta*, *toista*, *toisten*, *toistensa*, *tuo*, *tuohon*, *tuollainen*, *tuollaiseen*, *tuollaiseksi*, *tuollaisella*, *tuollaiselle*, *tuollaiselta*, *tuollaisen*, *tuollaisena*, *tuollaisessa*, *tuollaisesta*, *tuollaiset*, *tuollaisetta*, *tuollaisia*, *tuollaisien*, *tuollaisiin*, *tuollaisiksi*, *tuollaisilla*, *tuollaisille*, *tuollaisilta*, *tuollaisin*, *tuollaisina*, *tuollaisineen*, *tuollaisissa*, *tuollaisista*, *tuollaisitta*, *tuollaista*, *tuollaisten*, *tuon*, *tuona*, *tuota*, *tuskin*, *tähden*, *tähän*, *täksi*, *tällainen*, *tällaiseen*, *tällaiseksi*, *tällaisella*, *tällaiselle*, *tällaiselta*, *tällaisen*, *tällaisena*, *tällaisessa*, *tällaisesta*, *tällaiset*, *tällaisetta*, *tällaisia*, *tällaisien*, *tällaisiin*, *tällaisiksi*, *tällaisilla*, *tällaisille*, *tällaisilta*, *tällaisin*, *tällaisina*, *tällaisineen*, *tällaisissa*, *tällaisista*, *tällaisitta*, *tällaista*, *tällaisten*, *tälle*, *tällä*, *tällöin*, *tältä*, *tämä*, *tämän*, *tänne*, *tänä*, *tässä*, *tästä*, *täten*, *tätä*, *täydy*, *täytyi*, *täytyisi*, *täytykö*, *täytyköön*, *täytyminen*, *täytymistä*, *täytyne*, *täytynee*, *täytynyt*, *täytyy*, *täytyä*, *täällä*, *täältä*, *ulkoa*, *ulkona*, *ulos*, *usea*, *useaa*, *useain*, *useammaksi*, *useammalla*, *useammalle*, *useammalta*, *useamman*, *useammassa*, *useammasta*, *useammat*, *useammatta*, *useammiksi*, *useammilla*, *useammille*, *useammilta*, *useammin*, *useammissa*, *useammista*, *useammitta*, *useampaa*, *useampaan*, *useampain*, *useampana*, *useampi*, *useampia*, *useampiin*, *useampina*, *useampine*, *usean*, *useana*, *useat*, *useata*, *useiden*, *useihin*, *useiksi*, *useilla*, *useille*, *useilta*, *useimmaksi*, *useimmalla*, *useimmalle*, *useimmalta*, *useimman*, *useimmassa*, *useimmasta*, *useimmat*, *useimmatta*, *useimmiksi*, *useimmilla*, *useimmille*, *useimmilta*, *useimmin*, *useimmissa*, *useimmista*, *useimmitta*, *useimpaan*, *useimpain*, *useimpana*, *useimpia*, *useimpien*, *useimpiin*, *useimpina*, *useimpine*, *usein*, *useina*, *useine*, *useinta*, *useinten*, *useisiin*, *useissa*, *useista*, *useita*, *useitta*, *useitten*, *uudelleen*, *uudestaan*, *vaan*, *vai*, *vaikka*, *vaikkei*, *vaikkemme*, *vaikken*, *vaikket*, *vaikkette*, *vain*, *vallan*, *varsin*, *vasta*, *vastedes*, *vasten*, *vastikään*, *vielä*, *viereen*, *vierekkäin*, *vierelle*, *vierellä*, *viereltä*, *vieren*, *vieressä*, *vierestä*, *vieret*, *vieretysten*, *vieri*, *vieriin*, *vieritse*, *vieriä*, *viertä*, *voi*, *voida*, *voiden*, *voidessa*, *voikaa*, *voikoon*, *voikoot*, *voima*, *voimaisillaan*, *voiman*, *voimaton*, *voiminen*, *voimista*, *voin*, *voine*, *voinee*, *voineet*, *voinet*, *voinut*, *voisi*, *voisin*, *voisit*, *voit*, *voitaessa*, *voitaisi*, *voitaisiin*, *voitako*, *voitakoon*, *voitaman*, *voitane*, *voitava*, *voitiin*, *voitu*, *voiva*, *vuoksi*, *väliin*, *välillä*, *välissä*, *välistä*, *yhdeksi*, *yhdelle*, *yhdellä*, *yhdeltä*, *yhden*, *yhdessä*, *yhdesti*, *yhdestä*, *yhdet*, *yhdettä*, *yhteen*, *yhtenä*, *yhtä*, *yhtäälle*, *yhtäällä*, *yhtäältä*, *yhtään*, *yhä*, *yksi*, *yksiin*, *yksin*, *yksine*, *yksinä*, *yksittäin*, *yksiä*, *ylen*, *yli*, *ylitse*, *yläpuolella*, *ylös*, *ympäri*, *ympäriinsä*, *ympäritse*, *älkää*, *älköön*, *älkööt*, *ällös*, *älä*,

then we consider it a Finnish stop root (notation: **FinnishStopRoot**). A Finnish stop word matches the following string pattern:

$$\textbf{FinnishStopRoot}(\varnothing|ksi|lla|lle|llä|lta|ltä|ssa|ssä|sta|stä|tta|ttä)(\varnothing|kse)$$
$$(\varnothing|an|en|mme|ni|nne|nsa|nsä|si|tte|vat|vät)(\varnothing|han|hän|ka|kaan|kin|ko|kä|kään|kö|pa|pä)_m,$$

or (*jok*|*vast*)*'***X**.[97] All the Finnish stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

Up to some modifications, our method is modeled after the Porter stemming algorithm for Finnish (`http://snowball.tartarus.org/algorithms/finnish/stemmer.html`), devised by Martin F. Porter. Our modifications of the Porter stemming algorithm aim at two goals: to fully take care of verb conjugations and to fully accommodate to consonant/vowel alternations in declensions.

8.1.1 Effective spelling and essential root

It is assumed that all Finnish words are converted to lowercase before going through any of the procedures below.

Unlike the situation in German, diacritical marks in Finnish can affect the meaning of a word immensely. For example, we have *rohkaista* "to encourage" vs. *röhkäistä* "to grunt". Therefore, we are going to retain diacritical marks in the construction of Finnish effective spellings and essential roots, contrary to the practice for German.

Before we start, we recapitulate Martin F. Porter's definition of Finnish vowels, restricted vowels and long vowels.

Definition 8.2 (Finnish Vowels, Restricted Vowels and Long Vowels)**.** Hereafter in §8.1, the symbol **V** stands for any member from the list of Finnish vowels {*a*, *e*, *i*, *o*, *u*, *y*, *ä*, *ö*}. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Finnish vowels.

Dual to the notations above, the symbol **C** stands for any character that does not belong to the list {*a*, *e*, *i*, *o*, *u*, *y*, *ä*, *ö*}, and $\mathbf{C}_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list.

The symbol $\mathbf{V}^*$ stands for any member from the list of Finnish restricted vowels {*a*, *e*, *i*, *o*, *u*, *ä*, *ö*}, while $\mathbf{V}^{**}$ represents any one in the list of Finnish long vowels {*aa*, *ee*, *ii*, *oo*, *uu*, *ää*, *öö*}. □

In the following, we define the first and second protected ranges of a Finnish word, following Martin F. Porter's definition of regions *R*1 and *R*2.

Definition 8.3 (Finnish First protected range)**.** Let $\hat{\sigma}$ be a Finnish text string, then its first protected range $\mathrm{ProtRg}_1(\hat{\sigma})$ is specified through the following procedures:

- Look for the string pattern $\mathbf{C}_{m_0}\mathbf{V}_m\mathbf{C}\sim$ in the string $\hat{\sigma}$;
- If the string pattern above is found, the last position occupied by such a string defines $\mathrm{ProtRg}_1(\hat{\sigma})$; otherwise, set $\mathrm{ProtRg}_1(\hat{\sigma}) = \ell(\hat{\sigma})$.

Once $\mathrm{ProtRg}_1(\hat{\sigma})$ is determined, the text string $\hat{\sigma}$ needs to be explicitly marked with a separator signifying the position of the first protected range. Concretely speaking, we define $\mathrm{MarkRg}_1(\hat{\sigma})$ by inserting a dagger symbol ("†") at position $\mathrm{ProtRg}_1(\hat{\sigma}) + 1$ to the string $\hat{\sigma}$. □

Definition 8.4 (Finnish Second protected range)**.** Let $\hat{\sigma}'$ be a Finnish text string marked explicitly with its first protected range, then its second protected range $\mathrm{ProtRg}_2(\hat{\sigma}')$ is determined by the following procedures:

- Look for the string pattern $\dagger\mathbf{C}_{m_0}\mathbf{V}_m\mathbf{C}$ in the string $\hat{\sigma}'$;
- If the string pattern above is found, the last position occupied by such a string defines $\mathrm{ProtRg}_2(\hat{\sigma}')$; otherwise, set $\mathrm{ProtRg}_2(\hat{\sigma}') = \ell(\hat{\sigma}')$.

Once $\mathrm{ProtRg}_2(\hat{\sigma}')$ is determined, the text string $\hat{\sigma}'$ needs to be further marked with a separator signifying the position of the second protected range. Concretely speaking, we define $\mathrm{MarkRg}_2(\hat{\sigma}')$ by inserting a double dagger symbol ("‡") at position $\mathrm{ProtRg}_2(\hat{\sigma}') + 1$ to the string $\hat{\sigma}'$. □

Example 8.4.1. We use some English words to illustrate the workings of ProtRg_1, MarkRg_1, ProtRg_2 and MarkRg_2:

$\hat{\sigma}$	*beau*	*beauty*	*beautiful*	*beautifully*
$\mathrm{ProtRg}_1(\hat{\sigma})$	4	5	5	5
$\mathrm{MarkRg}_1(\hat{\sigma})$	*beau*†	*beaut*†*y*	*beaut*†*iful*	*beaut*†*ifully*
$\mathrm{ProtRg}_2(\mathrm{MarkRg}_1(\hat{\sigma}))$	5	7	8	8
$\mathrm{MarkRg}_2(\mathrm{MarkRg}_1(\hat{\sigma}))$	*beau*†‡	*beaut*†*y*‡	*beaut*†*if*‡*ul*	*beaut*†*if*‡*ully*

[97] Here, the apostrophe refers to either Apostrophe U+0027 or Right Single Quotation Mark U+2019, as is appropriate for the particular text mining task.

Algorithm 8.5 (Finnish effective spelling)**.** *For a Finnish word* $\hat{\sigma}$ *(converted to lower case form), its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in sequential steps:*

(1) *Convert to lowercase and do* $\sim ksest(a|ä) \rightarrow s$, $\sharp\mathbf{X} \rightarrow \varnothing$.

(2) *Do uudelleen*$\sim\!\!\rightarrow \rho\varepsilon$.

(3) *Do* $nt(o|ö)(i|j)(\varnothing|hi)(\varnothing|n)(\varnothing|a|en|ä) \rightarrow nta$, $\sim\hat{\chi}_1 eeseen\mathbf{X} \rightarrow \hat{\chi}_1 e$, $\sim\hat{\chi}_2 m(a|ä)sill\mathbf{X} \rightarrow \hat{\chi}_2$, $\sim\hat{\chi}_3 sest(a|ä)\mathbf{X} \rightarrow \hat{\chi}_3 nen$, $\underline{\mathbf{X}\hat{\chi}m(a|ä)(t|tt)(o|ö)(m|n)(\varnothing|a|aa|i|t|ä|ää)(\varnothing|en|iin|ksi|ll|lt|n|ss|st|t|tt)(\varnothing|a|e|ä)} \rightarrow \tau\nu\tau\text{-}\mathbf{X}\hat{\chi}$.

(4) *Replace*

(sisar|velj)enpo(i|j)**X**~	(sisar|velj)entyt**X**~	amerika	avioliit(∅|t)~	etei(n|s)~
võefω	*võiece*	*αμερικα*	*μαρρ*	*αιλs*

help~	henkilö	ihme**X**~	ihmi	isänne	john~	ka(d|t)o	kai(∅|k)u**X**~	kaika(a|isi)**X**~	kalle~
ηελρ	*περσö*	*ωοõδρε*	*inhimi*	*isä*	*γιωñn*	*λοσso*	*εχο*	*εχο*	*expnse*

kalli	kasvo	kats**X**	kirjasto	koir	kuol~	kylj~	kä(d|t)e~	käyt(∅|t)(e|i|ö)(∅|j)	luopu
expnsi	*φaso*	*katσ*	*λiβρo*	*dog*	*δiεl*	*kγlj*	*käsi*	*uσä*	*luovu*

läpikotaisin	maine	musta(ank|sukk)~	naimis~	ou(d|t)o~	pilot~	pool~	puh(e|u)**X**	puista	puolu
θurγ	*fame*	*jaλωσ*	*μarρ*	*ωeirδo*	*πiλοτ*	*πωολl*	*λογγ*	*ξaκa*	*defu*

puolust	päiv	selv~	seuralai(n|s)	silm	sisaru	sito~	sul(a|e|i)~	sulo~	summa	syntyperä
defens	*päiv*	*celar*	*σoclais*	*εγε*	*σiβλu*	*bĩvdo*	*μελτa*	*ξαρμο*	*σuμma*	*syntyenperä*

taho~	tall	tarkoi~	tarkoi~	tullut**X**~	tult	tun(n|t)o	tunnust	tuntei(s|t)**X**	tuntisi~	töi
auθo	*taλll*	*μeαñi*	*μeαñi*	*tuli*	*tul*	*tuntea*	*confeß*	*tuntea*	*tunteasi*	*työ*

usko~	uu(de|si)m(m|p)**X**~	vaikut(∅|t)**X**	val(l|t)$^{\mathbf{X}\in}$(a|o)	valli~	viha	vilp(∅|p)~
φaθo	*uusi*	*infλ*	*γawlt***X**	*δομĩi*	*aĩγρa*	*fardp*

vuo(si|teen|tena|tt)	yh(d|t)~	miss	nai(∅|si|va)(∅|n|t)(∅|e|ut)(∅|e|et|mme|tte|vat)
jahρi	*yks*	*neiti*	*μarρ*

uusitte	vuonna	~aikoi	~käyt(iin|te)	~käyttäy(d|t)**X**	~vuode(ksi|lla|lle|lta|n|ssa|sta|t)
uusi	*jahρ*	*aio*	*käy*	*baχs*	*jahρ*

(5) *Replace*

(mamm|muts|äip|äipp|äisk)~	**X**$^{\mathbf{X}\in}$(**V**|s)tapa	**X**$^{\mathbf{X}\in}$(∅|lähi)omai(n|s)~
äit	**X***tava*	**X***ρelτβa*

X$^{\mathbf{X}\in}$(∅|oiko)lu(e|i)~	ahdist	aiko~	alu	asem~	asu	aut(∅|t)~	avio~	avit(∅|t)~	avu~	eilis(∅|t)~
X*luki*	*aγiσt*	*aio*	*alku*	*σοτωεm*	*δερσu*	*aput*	*μarρo*	*aput*	*apu*	*eilin*

har(r|t)~	harm~	harv~	heng~	huone	ihm~	iho~	illalli~	isän(n|t)	jou(d|t)u~
δowt	*hram*	*ρar*	*henk*	*ζiμe*	*ημμm*	*σγκõo*	*σιπερlli*	*χοστon*	*γετu*

juo(ks|sk|ss)~	kahv	kakist~	katker~	kaun(a|oi|oj)~	kes(i|ä)~	kirjall	kirjav	kirjoit
juost	*kofφ*	*κof*	*βitτer*	*χατa*	*εετεa*	*ωiρtall*	*κολραν*	*ωiρt*

kitty~	ko(d|t)(e|i)	kohtalo	kuul~	kuun(n|t)el~	käyttä	käytöks	laakso	lan(g|k)o~
κiττy	*κoti*	*φατo*	*κuul*	*κuul*	*uσä*	*baχs*	*ωαληο*	*λαĩγo*

loist~	lom~	luist~	luk(∅|k)o	luki(nn|t|ts|tt)	luku	luv(a|i)	luvu	läh(d|t)	maal**V**~
σiηiĩ	*ηολδ*	*σiλδ*	*λοκο*	*λοκ*	*ταζπu*	*lupa*	*ταζπu*	*λεαb*	*αaμλl***V**

mail(e|ei|eja|i)~	maini(n|nn|t|ts|tt)~	mary~	matk	mielipah~	muist~	naam~
μαιλi	*μeĩτζion*	*μaργ*	*τeριπ*	*ργγuτ*	*μeĩμt*	*faσam*

nai(d|k|m|t)~	nei(d|t)o	oiv(a|i)~	onnist	ost~	parv	perust~	pi(d|t)em(m|p)~	pien~
μarρ	*μαιδο*	*φiĩva*	*σukc*	*βuη*	*guρp*	*βaζ*	*λοĩγ*	*σaμlλ*

pisi(mm|mp|n|nt)~	pitk**X**$^{\mathbf{X}\in}$(i|ä)~	pitu~	puh(d|t)	puisto	päivälli~	sa(d|t)(a|e|oi)~
λοĩγ	*λοĩγ***X**	*λοĩγu*	*κελaĩd*	*παρκο*	*δiĩvερlli*	*ραιĩa*

sal(e|ei|eja|i)	*seitsem~*	*selk~*	*seurustel*	*sie(d|t)~*	*siev(i|ä)~*	*suvai(n|nn|t|ts|tt)*
ζaαλi	*7em*	*sel*	*σocl*	*σotd*	*περττi*	*τολρat*

säily~	*sääst~*	*tahd*	*talou*	*tapa~*	*tapp~*	*tee~*	*tilan(n|t)~*	*to(de|si)*
räμṽy	*σavat*	*taht*	*εεκωu*	*tava*	*κiλ*	*ξai*	*σiτyn*	*totta*

tuha(n|nn|nt|t|tt)	*tulk$_{\hat{\chi}\notin}$(a|o)~*	*tun(n|t)(ei|en|ev|u)(∅|u)*	*tunti(n|s)*	*turh~*	*tyt(∅|t)är~*
θat	*τuλkkχ̂*	*tuntea*	*τuntis*	*ṽiλh*	*δoχτr*

tyt(∅|t)ö~	*tädi~*	*tänään~*	*vapi~*	*varh*	*varm*	*vel(∅|k)(a|o)*	*veli~*	*vih(∅|k)i*	*ylell*	*ylist~*
γiρλö	*täti*	*τoδηan*	*vavi*	*εarλh*	*κuσρ*	*δeβτa*	*veljen*	*μarρi*	*luxll*	*πuρazs*

aie(∅|tta)	*eilen*	*enoX*	*huomen(issa|na)*	*leskirouvX*	*nainen*
aike	*eilinen*	*ṽoκλo*	*τομροω*	*leski*	*naisen*

~aik(a|aa|oi|ojen)(∅|in|n|na|neen)	*~kaunis*	*~käytä(∅|mme|tte)*	*~käytös(∅|ten)*	*~varu(ks|s|st)X*	*~veli*
aja	*kaunii*	*uσä*	*baχs*	*κiβmt*	*veljen*

(6) *Replace*

(hyv|parah|paras|parh)~	**V** *'*	**X**$^{\in}$*(yö|öi***C**$_{\hat{\chi}\notin}$*)~*	*é*	*hdeks*	*herist~*	*hert~*	*herä~*	*huom~*	*ilo~*	*inh~*
parempi	**V***Q*	*q***X**	*e*	*hdex*	*ξak*	*σewt*	*ωakä*	*σim*	*ilzo*	*inzh*

is(i|ä)~	*jano~*	*juo(k|s)***C**$_{\hat{\chi}\notin}$*~*	*kirje~*	*kuum~*	*lm*	*miel*	*sall~*	*sellai***X**$^{\in}$*(n|s)~*	*siro~*	*sisko***X***~*	*toi(m|nt)*
faija	*γiano*	*qojo*	*girje*	*guum*	*λμ*	*μiel*	*sallq*	*zenlai***X**	*σiro*	*sisar*	*ηoβ*

toiv~	*tun(n|t)(ej|i)$_{\hat{\chi}\notin}$(e|s)*	*tunne(i|r)*	*tunnis(s|t)a*	*tuntei(hin|na|neen)*	*tuom(∅|ar)i~*
χoπ	*τuntiχ̂*	*τunti*	*τunti*	*τunti*	*δoμi*

vaj~	*ystäv(∅|yks|yst)(∅|i)~*	*hra*	*nti*	*rva*	*~(∅|si)tte*	*~***X**$^{\in}$*($\hat{\chi}_m$)z$\hat{\chi}_{m_0}$*	**~C ʼX**
waj	*ystew*	*herra*	*neiti*	*rouva*	*∅*	*χ̂z*	**C**

*~***X**$^{\in}$*(l|s)k(o|ö)*	*~***X**$^{\in}$*(u|y)(ksin|s|tta|ttä)*	*~ej(a|ä)*	*~et*	*~isk(aa|ää)*	*~lleen*	*~mies*	*~tunti*
X	**X***d*	*e*	*etkin*	*isk*	*ll*	*mieh*	*τunti*

(hyv|parah|paras|parh)~	**V** *'*	**X**$^{\in}$*(yö|öi***C**$_{\hat{\chi}\notin}$*)~*	*é*	*hdeks*	*herist~*	*hert~*	*herä~*	*huom~*	*ilo~*	*inh~*
parempi	**V***Q*	*q***X**	*e*	*hdex*	*ξak*	*σewt*	*ωakä*	*σim*	*ilzo*	*inzh*

is(i|ä)~	*jano~*	*juo(k|s)***C**$_{\hat{\chi}\notin}$*~*	*kirje~*	*kuum~*	*lm*	*miel*	*sall~*	*sellai***X**$^{\in}$*(n|s)~*	*siro~*	*sisko***X***~*	*toi(m|nt)*
faija	*γiano*	*qojo*	*girje*	*guum*	*λμ*	*μiel*	*sallq*	*zenlai***X**	*σiro*	*sisar*	*ηoβ*

toiv~	*tun(n|t)(ej|i)$_{\hat{\chi}\notin}$(e|s)*	*tunne(i|r)*	*tunnis(s|t)a*	*tuntei(hin|na|neen)*	*tuom(∅|ar)i~*
χoπ	*τuntiχ̂*	*τunti*	*τunti*	*τunti*	*δoμi*

vaj~	*ystäv(∅|yks|yst)(∅|i)~*	*hra*	*nti*	*rva*	*~(∅|si)tte*	*~***X**$^{\in}$*($\hat{\chi}_m$)z$\hat{\chi}_{m_0}$*	**~C ʼX**
waj	*ystew*	*herra*	*neiti*	*rouva*	*∅*	**X***z*	**C**

*~***X**$^{\in}$*(l|s)k(o|ö)*	*~***X**$^{\in}$*(u|y)(ksin|s|tta|ttä)*	*~ej(a|ä)*	*~et*	*~isk(aa|ää)*	*~lleen*	*~mies*	*~tunti*
X	**X***d*	*e*	*etkin*	*isk*	*ll*	*mieh*	*τunti*

(7) *If the result so far start with* **C**$_m$**VV~** *(where the two occurrences of* **V** *may or may not represent the same vowel), perform the following replacements on the word initial pattern* **C**$_m$**VV~***:*[98]

$\hat{\xi}^{\in}$**(V)***i*	**V****	*ie*	*uo*	*yö*
$\hat{\xi}ij\hat{\xi}i$	(**V****)$^{\{1\}}$*iji*	*eije*	*oijo*	*öijö*

If the result after step (6) does not start with **C**$_m$**VV~***, leave it as is.*

(8) *Mark the first protected range with* †.

(9) *Replace*

~X$^{\in}$*((a|e|i|n|o|t|u|y|ä|ö)(∅|†))(han|hän|kaan|kään|kin|ko|kö|pa|pä)*$_m$
X

[98] Here, the two instances of $\hat{\xi}$ in $\hat{\xi}ij\hat{\xi}i$ represent the same letter; (**V****)$^{\{1\}}$*iji* stands the first letter in the long vowel **V****, followed immediately by *iji*.

$\sim(\varnothing\|s)(iess(a\|ä)\|k(aa\|ää)(\varnothing\|mme)\|k(oo\|öö)(n\|t)\|minen\|nee(\varnothing\|t)\|n(e\|u\|y)t\|si(mme\|n\|t)\|tt(aessa\|äessä\|äisi\|aman\|ämän\|ane\|äne\|aneen\|äneen\|iin\|u\|y)\|v(a\|ä\|at\|ät))$
$\varnothing$

$\sim\dagger\mathbf{X}(l\|v)(a\|ä)i(n\|s)\hat{\chi}_{m_0}$	$\sim\dagger\mathbf{X}hk\hat{\chi}_{m_0}$	$\sim\dagger\mathbf{X}lli(n\|s)\hat{\chi}_{m_0}$
$\dagger\mathbf{X}$	$\dagger\mathbf{X}$	$\dagger\mathbf{X}$
$\sim\dagger\mathbf{X}luont(a\|e)\hat{\chi}_{m_0}$	$\sim\dagger\mathbf{X}suh(d\|t)\hat{\chi}_{m_0}$	$\sim\dagger\mathbf{X}t(a\|ä)r\hat{\chi}_{m_0}$
$\dagger\mathbf{X}$	$\dagger\mathbf{X}$	$\dagger\mathbf{X}$

(10) *Mark the second protected range with* ‡.

(11) *Do* $\sim\ddagger\mathbf{X}sti\to\ddagger\mathbf{X}$, $\sim\ddagger\mathbf{X}'(ks|ll|lt|ss|st)\mathbf{X}''\to\ddagger\mathbf{X}'$, $(k\ddagger s|l\ddagger l|l\ddagger t|s\ddagger s|s\ddagger t)\mathbf{X}'''\to\varnothing$.

(12) *Do* $\ddagger\to\varnothing$.

(13) *Replace*

$\sim\dagger it$	$\sim\dagger\mathbf{X}\mathbf{C}it$	$\sim\dagger\mathbf{X}(mme\|nne\|nsa\|nsä)$	$\sim\dagger\mathbf{X}_{\hat{\chi}\notin}(k)si$	$\sim\dagger\mathbf{X}^{\mathbf{X}_1\in}(\varnothing\|kse)ni$
i	$\dagger\mathbf{X}\mathbf{C}i$	$\dagger\mathbf{X}$	$\dagger\mathbf{X}\hat{\chi}$	$\sim\dagger\mathbf{X}\mathbf{X}'_1$

$\sim^{\mathbf{X}_2\in}((\dagger\mathbf{X}(in\|ll))\|i\dagger n\|l\dagger l)een$	$\sim^{\mathbf{X}_3\in}((\dagger\mathbf{X}(ll\|lt\|n\|ss\|st\|t))\|l\dagger l\|l\dagger t\|n\dagger\|s\dagger s\|s\dagger t\|t\dagger)^{\mathbf{X}_4\in}(aa\|ää)n$
$\mathbf{X}_2e$	$\mathbf{X}_3\mathbf{X}_4^{\{1\}}$

where one obtains $\mathbf{X}'_1$ *after doing* $kse\to ksi$ *on* $\mathbf{X}_1$.

(14) *Do* $\dagger\to\varnothing$.

(15) *Do* $maij\sim\to\varepsilon a\rho\theta$, $mm^{\hat{\chi}\in}(a|i|ä)(\varnothing|n)\to M\hat{\chi}$, $sisar\sim\to\sigma i\sigma a\rho$, $\tau\nu\tau$-$j\mathbf{V}_m\to\tau\nu\tau$-$Z\mathbf{V}_m$, $\tau\nu\tau$-$käy\to\tau\nu\tau$-$kkäy$, $\sim nt(a|ä)\to nt$, $\sim x(a|ä)s\to x$, $\underline{kahdexi}\to kaksi$.

Algorithm 8.6 (Finnish essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Finnish word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Mark the first protected range with* †.

(2) *Do* $\sim\dagger\mathbf{X}\hat{\xi}h\hat{\xi}n\to\dagger\mathbf{X}\hat{\xi}*$ *if* $\hat{\xi}\in\{a, e, i, o, u, ä, ö\}$. *Make the following replacements in the same sweep:*

$\sim\dagger\mathbf{X}\mathbf{V}^*i(den\|siin\|tten)$	$\sim^{\mathbf{X}_1\in}(\dagger\mathbf{X}(\mathbf{V}^{**}\|(a\|ä)k))seen$	$\sim^{\mathbf{X}_2\in}(\dagger\mathbf{X}\mathbf{C}\mathbf{V}\|\mathbf{C}\dagger\mathbf{V})(\varnothing\|ej)(a\|ä)$	$\sim\dagger\mathbf{X}\mathbf{V}tt(a\|ä)$
$\dagger\mathbf{X}\mathbf{V}^*i*$	$\mathbf{X}'_1*$	$\mathbf{X}_2*$	$\dagger\mathbf{X}\mathbf{V}*$

$\sim\dagger\mathbf{X}(ine\|ksi\|ll(a\|ä\|e)\|lt(a\|ä)\|n(a\|ä)\|ss(a\|ä)\|st(a\|ä))$	$\sim^{\mathbf{X}_3\in}(\dagger\mathbf{X})n$
$\dagger\mathbf{X}*$	$\dagger\mathbf{X}'_3*$

where one constructs $\mathbf{X}'_1$ *by doing* $\sim(a|ä)k\to\varnothing$ *on* $\mathbf{X}_1$, *and* $\mathbf{X}'_3$ *by doing* $\sim^{\mathbf{X}\in}(aa|ee|ie|ii|oo|uu|ää|öö)\to\mathbf{X}^{\{1\}}$ *on* $\mathbf{X}_3$.

(3) *Do* $\sim\dagger\mathbf{X}(m(a|ä)(\varnothing|isi)|m(ato|ätö)|mi|ne|t(a|ä)|tt(a|ä)i)\to\dagger\mathbf{X}*$.

(4) *Mark the second protected range with* ‡.

(5) *If the result so far does not end with* $*$, *append a symbol* # *at the end of the string.*

(6) *Replace*

$\sim\ddagger\mathbf{X}i(M\|mp)(a\|i\|ä)$	$\sim\ddagger\mathbf{X}(M\|mp)(a\|i\|ä)$
$\ddagger\mathbf{X}$	$\ddagger\mathbf{X}$ *if* $\mathbf{X}$ *does not end with* $\sim po$

(7) *Replace*

$\sim\dagger(\hat{\chi}\|\ddagger)_{m_0}(i\|j)*$	$\sim^{\mathbf{X}\in}(\dagger(\hat{\chi}\|\ddagger)_{m_0}\mathbf{V}\|\mathbf{V}(\dagger\|\ddagger)_m)t(\varnothing\|\ddagger)\#$
$\dagger(\hat{\chi}\|\ddagger)_{m_0}*$	$\mathbf{X}\tau$

(8) *If the pattern* $\sim\ddagger\mathbf{X}Ma\tau$ *is found, replace it with* $\ddagger\mathbf{X}'$. *Here,* $\mathbf{X}'$ *is obtained by removing* $\sim(\varnothing|i)Ma$ *from* $\mathbf{X}Ma$ *if* $\mathbf{X}$ *does not end with* $\sim po$; *otherwise, set* $\mathbf{X}'$ *as* $\mathbf{X}Ma$.

(9) *Do* $(\ddagger|*|\#|\tau)\to\varnothing$.

(10) *Do* $\sim\dagger\mathbf{XV}^{**} \to \dagger\mathbf{X}(\mathbf{V}^{**})^{\{1\}}$.

(11) *Do* $\sim\dagger\mathbf{XC}(a|e|i|\ddot{a}) \to \dagger\mathbf{XC}$.

(12) *Do* $\sim\dagger\mathbf{X}^{\hat{\chi}\in}(o|u)j \to \dagger\mathbf{X}\hat{\chi}$.

(13) *Do* $\sim\dagger\mathbf{X}_1 i(l)_m\mathbf{X}_2 \to \dagger\mathbf{X}_1$, $\sim\dagger\mathbf{XC}it \to \dagger\mathbf{XC}i$, $\sim\dagger\mathbf{X}jo \to \dagger\mathbf{X}j$, $\sim\dagger\mathbf{X}ks \to \dagger\mathbf{X}d$, $\sim\dagger\mathbf{X}(st|ts) \to \dagger\mathbf{X}$.

(14) *Do* $\dagger \to \varnothing$.

(15) *Replace*

ng	*nk*	*kah*(∅|*d*|*dest*|*t*)	*kä*(*v*|*y*)**X**	*nä*(*e*|*emm*|*en*|*et*|*h*|*hkö*|*ht*|*hty*|*ijä*|*ijäi*|*ke*|*ki*	*tun*(∅|*n*|*s*)	*vase*(∅|*nt*)	*~oi*	*~ti*
γ	*κ*	*kaks*	*käy*	*nähd*	*tunt*	*vaseM*	∅	*t*

(16) *Do* $\hat{\chi}_{\times 2} \to \hat{\chi}^{+}$, *that is, replace all double letters by the capital form of the corresponding single letter.*

(17) *Do* $U(d|t)\sim \to Usi$, $\sim\hat{\chi}(U|Y)(\varnothing|d|s|t) \to \hat{\chi}$, *Us* $\to$ *Usi.*

For clustering purposes, we will also define an operation to undouble final consonants in the effective spelling of a Finnish word.

Algorithm 8.7 (Reduction of Final Double Consonants)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Finnish word, then* UndCons($\hat{\sigma}$) *is defined by the following procedure:*

- *Look for the pattern* $\sim\hat{\chi}_{\times 2}\mathbf{V}_{m_0}$ *in* $\hat{\sigma}$.
- *If the pattern is not found or if* $\hat{\chi}$ *is one of the Finnish vowels* {*a, e, i, o, u, y, ä, ö*}*, do nothing; otherwise, replace the pattern by* $\hat{\chi}\mathbf{V}_{m_0}$.

8.1.2 Admissible mutation and approximate clustering

Algorithm 8.8 (Simple heredity test)**.** *Construct* $\hat{\alpha}'$ *by doing* $\sim(d|e|i|n|t|v) \to \varnothing$ *on* $\hat{\alpha}$. *Define* $\hat{\beta}'$ *similarly. The Boolean-valued function* SimpHrdTest($\hat{\alpha},\hat{\beta}$) *returns* **TRUE** *if the lowercase form of* $\hat{\alpha}$ *contains at least one instance of* {*a, e, i, o, u, y, ä, ö*} **AND** *at least one of the following three conditions holds:*[99]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\alpha}' = \hat{\beta}'$;

(iii) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\beta} = \hat{\alpha}\mathbf{VX}$.

In what follows, we define SuffixNW($\hat{\alpha},\hat{\beta}$), RootNW($\hat{\alpha},\hat{\beta}$), NW*($\hat{\alpha},\hat{\beta}$) and SuffixSW($\hat{\alpha},\hat{\beta}$), SuffixSW($\hat{\alpha},\hat{\beta}$), SW*($\hat{\alpha},\hat{\beta}$) as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5.

Consonant gradations refer to consonant mutation patterns found in Finnish and its close relative, Estonian. (While Hungarian is also a Uralic language, it does not exhibit systematic consonant mutations as in Finnish.) In order to accommodate to this particular feature of Finnish, we need to define admissible consonant alternation patterns in Finnish effective spellings and essential roots.

Definition 8.9 (Finnish consonant gradations)**.** Define $\mathbb{G}$ = [∅, ∅]||[*k*, ∅]||[*k*, *j*]||[*k*, *Q*]||[*k*, *v*]||[*K*, *k*]||[*L*, *l*]||[*L*, *lt*]||[*M*, *mp*]||[*mp*, *nt*]||[*N*, *n*]||[*N*, *nt*]||[*p*, *v*]||[*P*, *p*]||[*R*, *r*]||[*R*, *rt*]||[*s*, *d*]||[*s*, *t*]||[*S*, *s*]||[*S*, *st*]||[*t*, *d*]||[*t*, *s*]||[*T*, *s*]||[*T*, *t*]||[*κ*, *γ*] as admissible Finnish consonant gradations.

Algorithm 8.10 (Admissible suffix mismatch and vowel alternation)**.** *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$*, the Boolean-valued function*

$$\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}), \text{SuffixNW}(\hat{\alpha},\hat{\beta}), \text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if at least one of the following three conditions holds:*

(i) SuffixNW($\hat{\alpha},\hat{\beta}$) = [∅|*h*|$\mathbf{V}_m$, ∅|*h*|$\mathbf{V}_m$] **AND** (NW*($\hat{\alpha},\hat{\beta}$) = $\mathbb{G}$ **OR** NW*($\hat{\beta},\hat{\alpha}$) = $\mathbb{G}$ **OR** (NW*($\hat{\alpha},\hat{\beta}$) = ∅ **AND** *the lowercase form of* RootNW($\hat{\alpha},\hat{\beta}$) *contains at least one instance of* **V**))*;*

(ii) NW*($\hat{\alpha},\hat{\beta}$) = ∅ **AND** SuffixNW$'$($\hat{\alpha},\hat{\beta}$) = $\mathbb{G}$|[*a, o*]||[*m, nt*]||[*n, t*]*, where one obtains* SuffixNW$'$($\hat{\alpha},\hat{\beta}$) *by doing* $\sim(h|s|\mathbf{V}_m) \to \varnothing$ *on both components of* SuffixNW($\hat{\alpha},\hat{\beta}$)*;*

[99] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

(iii) $\mathrm{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$ **AND** $\mathrm{SuffixNW}''(\hat{\beta},\hat{\alpha}) = \mathbb{G}$, *where* $\mathrm{SuffixNW}''(\hat{\beta},\hat{\alpha})$ *results from doing* $\sim(h|\mathbf{V}_m) \to \varnothing$ *on both components of* $\mathrm{SuffixNW}(\hat{\beta},\hat{\alpha})$.

Similarly, one can evaluate another Boolean-valued function

$$\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW.

The following heredity test function is very similar to the German version (Algorithm), except that we additionally require that the two strings in question must start with the same letter — constant gradations do not apply to the initial consonants.

Algorithm 8.11 (Heredity test function)**.** *In what follows,* $\ell([\hat{\sigma},\hat{\tau}]) = \min\{\ell(\hat{\sigma}),\ell(\hat{\tau})\}$. *For two strings* $\hat{\alpha}$ *and* $\hat{\beta}$, *the Boolean-valued function* $\mathrm{HrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if at least one of the following three conditions holds:*

(i) $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta}) = \mathbf{TRUE}$;

(ii) $\ell(\mathrm{RootNW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SuffixNW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{NW}^*(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootNW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixNW}(\hat{\alpha},\hat{\beta}), \mathrm{NW}^*(\hat{\alpha},\hat{\beta})) = \mathbf{TRUE}$ **AND** *both* $\hat{\alpha}$ *and* $\hat{\beta}$ *start with the same letter;*

(iii) $\ell(\mathrm{RootSW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SuffixSW}(\hat{\alpha},\hat{\beta})) + \ell(\mathrm{SW}^*(\hat{\alpha},\hat{\beta})) \geq \frac{\max\{\ell(\hat{\alpha}),\ell(\hat{\beta})\}}{2}$ **AND** $\mathrm{AdmMut}(\mathrm{RootSW}(\hat{\alpha},\hat{\beta}), \mathrm{SuffixSW}(\hat{\alpha},\hat{\beta}), \mathrm{SW}^*(\hat{\alpha},\hat{\beta})) = \mathbf{TRUE}$ **AND** *both* $\hat{\alpha}$ *and* $\hat{\beta}$ *start with the same letter.*

Algorithm 8.12 (Approximate clustering of Finnish words)**.** *The approximate clustering of a list of Finnish words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1, \mathrm{EffSpell}(\hat{\alpha}_1), \mathrm{UndCons}(\mathrm{EffSpell}(\hat{\alpha}_1))), \ldots, (\hat{\alpha}_N, \mathrm{EffSpell}(\hat{\alpha}_N), \mathrm{UndCons}(\mathrm{EffSpell}(\hat{\alpha}_N)))\}$ *alphabetically according to the third component (with higher priority) and the second component (with lower priority). If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1)}), \mathrm{UndCons}(\mathrm{EffSpell}(\hat{\alpha}_{(1)}))), \ldots, (\hat{\alpha}_{(N)}, \mathrm{EffSpell}(\hat{\alpha}_{(N)}), \mathrm{UndCons}(\mathrm{EffSpell}(\hat{\alpha}_{(N)})))\}$ *satisfies both* $\mathrm{HrdTest}(\mathrm{EffSpell}(\hat{\alpha}_{(n)}), \mathrm{EffSpell}(\hat{\alpha}_{(n+1)})) = \mathbf{FALSE}$ *and* $\mathrm{HrdTest}(\mathrm{EffSpell}(\hat{\alpha}_{(n+1)}), \mathrm{EffSpell}(\hat{\alpha}_{(n)})) = \mathbf{FALSE}$, *where* $n \in \mathbb{Z} \cap [1, N)$, *then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words with tags:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,1)}), \ldots), \ldots, (\hat{\alpha}_{(1,n_1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,n_1)}), \ldots)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(M,1)}), \ldots), \ldots, (\hat{\alpha}_{(M,n_M)}, \mathrm{EffSpell}(\hat{\alpha}_{(M,n_M)}), \ldots)\}\}$, *where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)})), \ldots, (\hat{\alpha}_{(m,n_m)}, \mathrm{EffSpell}(\hat{\alpha}_{(m,n_m)}))\}$ *where* $m \in \mathbb{Z} \cap [1, M]$, *we augment it into a tagged entry* $G_m = (g_m, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)}), \mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})), T_m)$, *where one constructs the tag* T_m *from* $\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))$ *in two steps:*

- *Replace all capital letters (including those in the standard Latin alphabet, as well as Ä and Ö) by double letters in lowercase form, and do* $\sim o \to a$.
- *Apply* UndCons.

The list $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* T_m *(with highest priority),* $\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with medium priority), and* $\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}, \hat{\gamma}'''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)}, \hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\mathrm{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \mathbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \mathbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, *then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$. *Finally, generate a list of word clusters* $\{\check{\Gamma}_1 = (\check{G}_{(1,1)}, \ldots), \ldots, \check{\Gamma}_K = (\check{G}_{(K,1)}, \ldots)\}$ *by discarding all the tags.*

Example 8.12.1. There are 78 inflection classes in the Finnish language, according to the classification scheme of KOTUS (`https://en.wiktionary.org/wiki/Appendix:Finnish_nominal_inflection` and `https://en.wiktionary.org/wiki/Appendix:Finnish_conjugation`). Among these classes, the first 51 classes describe nominal declensions (nouns, adjectives and numerals) while the rest are classes for verb conjugations. We pick one representative word from each inflection class, with or without gradation (consonant alternation in the stem).

The KOTUS class number will be shown parenthetically after the English translation for each word family of Finnish nominals below.

valo, *valoa*, *valoihin*, *valoiksi*, *valoilla*, *valoille*, *valoilta*, *valoin*, *valoina*, *valoineen*, *valoissa*, *valoista*, *valoitta*, *valoja*, *valojen*, *valoksi*, *valolla*, *valolle*, *valolta*, *valon*, *valona*, *valoon*, *valossa*, *valosta*, *valot*, *valotta* — “light” (1);

palvelu, *palvelua*, *palveluiden*, *palveluihin*, *palveluiksi*, *palveluilla*, *palveluille*, *palveluilta*, *palveluin*, *palveluina*, *palveluineen*, *palveluissa*, *palveluista*, *palveluita*, *palveluitta*, *palveluitten*, *palveluja*, *palvelujen*, *palveluksi*, *palvelulla*, *palvelulle*, *palvelulta*, *palvelun*, *palveluna*, *palvelussa*, *palvelusta*, *palvelut*, *palvelutta*, *palveluun* — “service” (2);

valtio, *valtioiden*, *valtioihin*, *valtioiksi*, *valtioilla*, *valtioille*, *valtioilta*, *valtioin*, *valtioina*, *valtioineen*, *valtioissa*, *valtioista*, *valtioita*, *valtioitta*, *valtioitten*, *valtioksi*, *valtiolla*, *valtiolle*, *valtiolta*, *valtion*, *valtiona*, *valtioon*, *valtiossa*, *valtiosta*, *valtiot*, *valtiota*, *valtiotta* — “government” (3);

laatikko, *laatikkoa*, *laatikkoihin*, *laatikkoina*, *laatikkoineen*, *laatikkoja*, *laatikkojen*, *laatikkona*, *laatikkoon*, *laatikoiden*, *laatikoihin*, *laatikoiksi*, *laatikoilla*, *laatikoille*, *laatikoilta*, *laatikoin*, *laatikoissa*, *laatikoista*, *laatikoita*, *laatikoitta*, *laatikoitten*, *laatikoksi*, *laatikolla*, *laatikolle*, *laatikolta*, *laatikon*, *laatikossa*, *laatikosta*, *laatikot*, *laatikotta* — “box” (4);

risteihin, *risteiksi*, *risteille*, *risteillä*, *risteiltä*, *ristein*, *risteineen*, *risteinä*, *risteissä*, *risteistä*, *risteittä*, *ristejä*, *risti*, *ristien*, *ristiin*, *ristiksi*, *ristille*, *ristillä*, *ristiltä*, *ristin*, *ristinä*, *ristissä*, *rististä*, *ristit*, *ristittä*, *ristiä* — “cross” (5);

papereiden, *papereihin*, *papereiksi*, *papereilla*, *papereille*, *papereilta*, *paperein*, *papereina*, *papereineen*, *papereissa*, *papereista*, *papereita*, *papereitta*, *papereitten*, *papereja*, *paperi*, *paperia*, *paperien*, *paperiin*, *paperiksi*, *paperilla*, *paperille*, *paperilta*, *paperin*, *paperina*, *paperissa*, *paperista*, *paperit*, *paperitta* — “paper” (6);

ovea, *oveen*, *oveksi*, *ovella*, *ovelle*, *ovelta*, *oven*, *ovena*, *ovessa*, *ovesta*, *ovet*, *ovetta*, *ovi*, *ovia*, *ovien*, *oviin*, *oviksi*, *ovilla*, *oville*, *ovilta*, *ovin*, *ovina*, *ovineen*, *ovissa*, *ovista*, *ovitta* — “door” (7);

nalle, *nallea*, *nalleen*, *nalleihin*, *nalleiksi*, *nalleilla*, *nalleille*, *nalleilta*, *nallein*, *nallein*, *nalleina*, *nalleineen*, *nalleissa*, *nalleista*, *nalleitta*, *nalleja*, *nallejen*, *nalleksi*, *nallella*, *nallelle*, *nallelta*, *nallen*, *nallena*, *nallessa*, *nallesta*, *nallet*, *nalletta* — “teddy bear” (8);

kala, *kalaa*, *kalaan*, *kalain*, *kalaksi*, *kalalla*, *kalalle*, *kalalta*, *kalan*, *kalana*, *kalassa*, *kalasta*, *kalat*, *kalatta*, *kaloihin*, *kaloiksi*, *kaloilla*, *kaloille*, *kaloilta*, *kaloin*, *kaloina*, *kaloineen*, *kaloissa*, *kaloista*, *kaloitta*, *kaloja*, *kalojen* — “fish” (9);

koira, *koiraa*, *koiraan*, *koirain*, *koiraksi*, *koiralla*, *koiralle*, *koiralta*, *koiran*, *koirana*, *koirassa*, *koirasta*, *koirat*, *koiratta*, *koiria*, *koirien*, *koiriin*, *koiriksi*, *koirilla*, *koirille*, *koirilta*, *koirin*, *koirina*, *koirineen*, *koirissa*, *koirista*, *koiritta* — “dog” (10);

omena, *omenaa*, *omenaan*, *omenain*, *omenaksi*, *omenalla*, *omenalle*, *omenalta*, *omenan*, *omenana*, *omenassa*, *omenasta*, *omenat*, *omenatta*, *omenia*, *omenien*, *omeniin*, *omeniksi*, *omenilla*, *omenille*, *omenilta*, *omenin*, *omenina*, *omenineen*, *omenissa*, *omenista*, *omenitta*, *omenoiden*, *omenoihin*, *omenoiksi*, *omenoilla*, *omenoille*, *omenoilta*, *omenoin*, *omenoina*, *omenoineen*, *omenoissa*, *omenoista*, *omenoita*, *omenoitta*, *omenoitten*, *omenoja*, *omenojen* — “apple” (11);

kulkija, *kulkijaa*, *kulkijaan*, *kulkijain*, *kulkijaksi*, *kulkijalla*, *kulkijalle*, *kulkijalta*, *kulkijan*, *kulkijana*, *kulkijassa*, *kulkijasta*, *kulkijat*, *kulkijatta*, *kulkijoiden*, *kulkijoihin*, *kulkijoiksi*, *kulkijoilla*, *kulkijoille*, *kulkijoilta*, *kulkijoin*, *kulkijoina*, *kulkijoineen*, *kulkijoissa*, *kulkijoista*, *kulkijoita*, *kulkijoitta*, *kulkijoitten* — “traveller” (12);

katiska, *katiskaa*, *katiskaan*, *katiskain*, *katiskaksi*, *katiskalla*, *katiskalle*, *katiskalta*, *katiskan*, *katiskana*, *katiskassa*, *katiskasta*, *katiskat*, *katiskatta*, *katiskoiden*, *katiskoihin*, *katiskoiksi*, *katiskoilla*, *katiskoille*, *katiskoilta*, *katiskoin*, *katiskoina*, *katiskoineen*, *katiskoissa*, *katiskoista*, *katiskoita*, *katiskoitta*, *katiskoitten*, *katiskoja*, *katiskojen* — “fishtrap” (13);

solakaksi, *solakalla*, *solakalle*, *solakalta*, *solakan*, *solakassa*, *solakasta*, *solakat*, *solakatta*, *solakka*, *solakkaa*, *solakkaan*, *solakkain*, *solakkana*, *solakkoihin*, *solakkoina*, *solakkoineen*, *solakkoja*, *solakkojen*, *solakoiden*, *solakoihin*, *solakoiksi*, *solakoilla*, *solakoille*, *solakoilta*, *solakoin*, *solakoissa*, *solakoista*, *solakoita*, *solakoitta*, *solakoitten* — “slender” (14);

korkea, *korkeaa*, *korkeaan*, *korkeain*, *korkeaksi*, *korkealla*, *korkealle*, *korkealta*, *korkean*, *korkeana*, *korkeassa*, *korkeasta*, *korkeat*, *korkeata*, *korkeatta*, *korkeiden*, *korkeihin*, *korkeiksi*, *korkeilla*, *korkeille*, *korkeilta*, *korkein*, *korkeina*, *korkeine*, *korkeisiin*, *korkeissa*, *korkeista*, *korkeita*, *korkeitta*, *korkeitten* — “tall” (15);

vanhemmaksi, *vanhemmalla*, *vanhemmalle*, *vanhemmalta*, *vanhemman*, *vanhemmassa*, *vanhemmasta*, *vanhemmat*, *vanhemmatta*, *vanhemmiksi*, *vanhemmilla*, *vanhemmille*, *vanhemmilta*, *vanhemmin*, *vanhemmissa*, *vanhemmista*, *vanhemmitta*, *vanhempaa*, *vanhempaan*, *vanhempain*, *vanhempana*, *vanhempi*, *vanhempia*, *vanhempien*, *vanhempiin*, *vanhempina*, *vanhempine* — "senior" (16);

vapaa, *vapaaksi*, *vapaalla*, *vapaalle*, *vapaalta*, *vapaan*, *vapaana*, *vapaaseen*, *vapaassa*, *vapaasta*, *vapaat*, *vapaata*, *vapaatta*, *vapaiden*, *vapaihin*, *vapaiksi*, *vapailla*, *vapaille*, *vapailta*, *vapain*, *vapaina*, *vapaine*, *vapaisiin*, *vapaissa*, *vapaista*, *vapaita*, *vapaitta*, *vapaitten* — "free" (17);

maa, *maahan*, *maaksi*, *maalla*, *maalle*, *maalta*, *maan*, *maana*, *maassa*, *maasta*, *maat*, *maata*, *maatta*, *maiden*, *maihin*, *maiksi*, *mailla*, *maille*, *mailta*, *main*, *maina*, *maineen*, *maissa*, *maista*, *maita*, *maitta*, *maitten* — "earth" (18);

soiden, *soihin*, *soiksi*, *soilla*, *soille*, *soilta*, *soin*, *soina*, *soineen*, *soissa*, *soista*, *soita*, *soitta*, *soitten*, *suo*, *suohon*, *suoksi*, *suolla*, *suolle*, *suolta*, *suon*, *suona*, *suossa*, *suosta*, *suot*, *suota*, *suotta* — "swamp" (19);

filee, *fileehen*, *fileeksi*, *fileelle*, *fileellä*, *fileeltä*, *fileen*, *fileenä*, *fileeseen*, *fileessä*, *fileestä*, *fileet*, *fileettä*, *fileetä*, *fileiden*, *fileihin*, *fileiksi*, *fileille*, *fileillä*, *fileiltä*, *filein*, *fileineen*, *fileinä*, *fileisiin*, *fileissä*, *fileistä*, *fileitten*, *fileittä*, *fileitä* — "fillet" (20);

rosé, *roséhen*, *roséiden*, *roséihin*, *roséiksi*, *roséilla*, *roséille*, *roséilta*, *roséin*, *roséina*, *roséineen*, *roséissa*, *roséista*, *roséita*, *roséitta*, *roséitten*, *roséksi*, *rosélla*, *rosélle*, *roselta*, *rosén*, *roséna*, *roséssa*, *rosésta*, *rosét*, *roséta*, *rosétta* — "pinkwine" (21);

parfait, *parfait'hen*, *parfait'iden*, *parfait'ihin*, *parfait'iksi*, *parfait'illa*, *parfait'ille*, *parfait'ilta*, *parfait'in*, *parfait'ina*, *parfait'ineen*, *parfait'issa*, *parfait'ista*, *parfait'ita*, *parfait'itta*, *parfait'itten*, *parfait'ksi*, *parfait'lla*, *parfait'lle*, *parfait'lta*, *parfait'n*, *parfait'na*, *parfait'ssa*, *parfait'sta*, *parfait't*, *parfait'ta*, *parfait'tta* — "parfait" (22);

tiileen, *tiileksi*, *tiilelle*, *tiilellä*, *tiileltä*, *tiilen*, *tiilenä*, *tiilessä*, *tiilestä*, *tiilet*, *tiilettä*, *tiili*, *tiilien*, *tiiliin*, *tiiliksi*, *tiilille*, *tiilillä*, *tiililtä*, *tiilin*, *tiilineen*, *tiilinä*, *tiilissä*, *tiilistä*, *tiilittä*, *tiiliä*, *tiiltä* — "brick" (23);

uneen, *uneksi*, *unella*, *unelle*, *unelta*, *unen*, *unena*, *unessa*, *unesta*, *unet*, *unetta*, *uni*, *unia*, *unien*, *uniin*, *uniksi*, *unilla*, *unille*, *unilta*, *unin*, *unina*, *unineen*, *unissa*, *unista*, *unitta*, *unta*, *unten* — "dream" (24);

toimea, *toimeen*, *toimeksi*, *toimella*, *toimelle*, *toimelta*, *toimen*, *toimena*, *toimessa*, *toimesta*, *toimet*, *toimetta*, *toimi*, *toimia*, *toimien*, *toimiin*, *toimiksi*, *toimilla*, *toimille*, *toimilta*, *toimin*, *toimina*, *toimineen*, *toimissa*, *toimista*, *toimitta*, *tointa*, *tointen* — "chore" (25);

pieneen, *pieneksi*, *pienelle*, *pienellä*, *pieneltä*, *pienen*, *pienenä*, *pienessä*, *pienestä*, *pienet*, *pienettä*, *pieni*, *pienien*, *pieniin*, *pieniksi*, *pienille*, *pienillä*, *pieniltä*, *pienin*, *pienine*, *pieninä*, *pienissä*, *pienistä*, *pienittä*, *pieniä*, *pienten*, *pientä* — "small" (26);

kädeksi, *kädelle*, *kädellä*, *kädeltä*, *käden*, *kädessä*, *kädestä*, *kädet*, *kädettä*, *käsi*, *käsien*, *käsiin*, *käsiksi*, *käsille*, *käsillä*, *käsiltä*, *käsin*, *käsineen*, *käsinä*, *käsissä*, *käsistä*, *käsittä*, *käsiä*, *käteen*, *kätenä*, *kätten*, *kättä* — "arm" (27);

kynneksi, *kynnelle*, *kynnellä*, *kynneltä*, *kynnen*, *kynnessä*, *kynnestä*, *kynnet*, *kynnettä*, *kynsi*, *kynsien*, *kynsiin*, *kynsiksi*, *kynsille*, *kynsillä*, *kynsiltä*, *kynsin*, *kynsineen*, *kynsinä*, *kynsissä*, *kynsistä*, *kynsittä*, *kynsiä*, *kynteen*, *kyntenä*, *kyntten*, *kynttä* — "fingernail" (28);

lapseen, *lapseksi*, *lapsella*, *lapselle*, *lapselta*, *lapsen*, *lapsena*, *lapsessa*, *lapsesta*, *lapset*, *lapsetta*, *lapsi*, *lapsia*, *lapsien*, *lapsiin*, *lapsiksi*, *lapsilla*, *lapsille*, *lapsilta*, *lapsin*, *lapsina*, *lapsineen*, *lapsissa*, *lapsista*, *lapsitta* — "child" (29);

veisten, *veistä*, *veitseen*, *veitseksi*, *veitselle*, *veitsellä*, *veitseltä*, *veitsen*, *veitsenä*, *veitsessä*, *veitsestä*, *veitset*, *veitsettä*, *veitsi*, *veitsien*, *veitsiin*, *veitsiksi*, *veitsille*, *veitsillä*, *veitsiltä*, *veitsin*, *veitsineen*, *veitsinä*, *veitsissä*, *veitsistä*, *veitsittä*, *veitsiä* — "knife" (30);

kahdeksi, *kahdella*, *kahdelle*, *kahdelta*, *kahden*, *kahdessa*, *kahdesta*, *kahdesti*, *kahdet*, *kahdetta*, *kahta*, *kahtaalla*, *kahtaalle*, *kahtaalta*, *kahteen*, *kahtena*, *kahtia*, *kaksi*, *kaksia*, *kaksien*, *kaksiin*, *kaksiksi*, *kaksilla*, *kaksille*, *kaksilta*, *kaksin*, *kaksina*, *kaksine*, *kaksissa*, *kaksista*, *kaksitta*, *kaksittain* — "two" (31);

sisar, *sisareen*, *sisareksi*, *sisarella*, *sisarelle*, *sisarelta*, *sisaren*, *sisarena*, *sisaressa*, *sisaresta*, *sisaret*, *sisaretta*, *sisaria*, *sisarien*, *sisariin*, *sisariksi*, *sisarilla*, *sisarille*, *sisarilta*, *sisarin*, *sisarina*, *sisarineen*, *sisarissa*, *sisarista*, *sisaritta*, *sisarta*, *sisarten* — “sister” (32);

kytkimeen, *kytkimeksi*, *kytkimelle*, *kytkimellä*, *kytkimeltä*, *kytkimen*, *kytkimenä*, *kytkimessä*, *kytkimestä*, *kytkimet*, *kytkimettä*, *kytkimien*, *kytkimiin*, *kytkimiksi*, *kytkimille*, *kytkimillä*, *kytkimiltä*, *kytkimin*, *kytkimineen*, *kytkiminä*, *kytkimissä*, *kytkimistä*, *kytkimittä*, *kytkimiä*, *kytkin*, *kytkinten*, *kytkintä* — “switch” (33);

onneton, *onnetonta*, *onnetonten*, *onnettomaan*, *onnettomaksi*, *onnettomalla*, *onnettomalle*, *onnettomalta*, *onnettoman*, *onnettomana*, *onnettomassa*, *onnettomasta*, *onnettomat*, *onnettomatta*, *onnettomia*, *onnettomien*, *onnettomiin*, *onnettomiksi*, *onnettomilla*, *onnettomille*, *onnettomilta*, *onnettomin*, *onnettomina*, *onnettomine*, *onnettomissa*, *onnettomista*, *onnettomitta* — “unhappy” (34);

lämmin, *lämmintä*, *lämpimien*, *lämpimiin*, *lämpimiksi*, *lämpimille*, *lämpimillä*, *lämpimiltä*, *lämpimin*, *lämpimine*, *lämpiminä*, *lämpimissä*, *lämpimistä*, *lämpimittä*, *lämpimiä*, *lämpimäin*, *lämpimäksi*, *lämpimälle*, *lämpimällä*, *lämpimältä*, *lämpimän*, *lämpimänä*, *lämpimässä*, *lämpimästä*, *lämpimät*, *lämpimättä*, *lämpimään* — “warm” (35);

sisimmiksi, *sisimmille*, *sisimmillä*, *sisimmiltä*, *sisimmin*, *sisimmissä*, *sisimmistä*, *sisimmittä*, *sisimmäksi*, *sisimmälle*, *sisimmällä*, *sisimmältä*, *sisimmän*, *sisimmässä*, *sisimmästä*, *sisimmät*, *sisimmättä*, *sisimpien*, *sisimpiin*, *sisimpine*, *sisimpinä*, *sisimpiä*, *sisimpäin*, *sisimpänä*, *sisimpään*, *sisin*, *sisinten*, *sisintä* — “innermost” (36);

vasemmaksi, *vasemmalla*, *vasemmalle*, *vasemmalta*, *vasemman*, *vasemmassa*, *vasemmasta*, *vasemmat*, *vasemmatta*, *vasemmiksi*, *vasemmilla*, *vasemmille*, *vasemmilta*, *vasemmin*, *vasemmissa*, *vasemmista*, *vasemmitta*, *vasempaa*, *vasempaan*, *vasempain*, *vasempana*, *vasempia*, *vasempien*, *vasempiin*, *vasempina*, *vasempine*, *vasen*, *vasenta*, *vasenten* — “left” (37);

nainen, *naiseen*, *naiseksi*, *naisella*, *naiselle*, *naiselta*, *naisen*, *naisena*, *naisessa*, *naisesta*, *naiset*, *naisetta*, *naisia*, *naisien*, *naisiin*, *naisiksi*, *naisilla*, *naisille*, *naisilta*, *naisin*, *naisina*, *naisineen*, *naisissa*, *naisista*, *naisitta*, *naista*, *naisten* — “woman” (38);

vastaukseen, *vastaukseksi*, *vastauksella*, *vastaukselle*, *vastaukselta*, *vastauksen*, *vastauksena*, *vastauksessa*, *vastauksesta*, *vastaukset*, *vastauksetta*, *vastauksia*, *vastauksien*, *vastauksiin*, *vastauksiksi*, *vastauksilla*, *vastauksille*, *vastauksilta*, *vastauksin*, *vastauksina*, *vastauksineen*, *vastauksissa*, *vastauksista*, *vastauksitta*, *vastaus*, *vastausta*, *vastausten* — “answer” (39);

kalleudeksi, *kalleudella*, *kalleudelle*, *kalleudelta*, *kalleuden*, *kalleudessa*, *kalleudesta*, *kalleudet*, *kalleudetta*, *kalleuksia*, *kalleuksien*, *kalleuksiin*, *kalleuksiksi*, *kalleuksilla*, *kalleuksille*, *kalleuksilta*, *kalleuksin*, *kalleuksina*, *kalleuksineen*, *kalleuksissa*, *kalleuksista*, *kalleuksitta*, *kalleus*, *kalleuteen*, *kalleutena*, *kalleutta* — “expensiveness” (40);

vieraaksi, *vieraalla*, *vieraalle*, *vieraalta*, *vieraan*, *vieraana*, *vieraaseen*, *vieraassa*, *vieraasta*, *vieraat*, *vieraatta*, *vieraiden*, *vieraihin*, *vieraiksi*, *vierailla*, *vieraille*, *vierailta*, *vierain*, *vieraina*, *vieraine*, *vieraisiin*, *vieraissa*, *vieraista*, *vieraita*, *vieraitta*, *vieraitten*, *vieras*, *vierasta* — “unfamiliar” (41);

mieheen, *mieheksi*, *miehelle*, *miehellä*, *mieheltä*, *miehen*, *miehenä*, *miehessä*, *miehestä*, *miehet*, *miehettä*, *miehien*, *miehiin*, *miehiksi*, *miehille*, *miehillä*, *miehiltä*, *miehin*, *miehineen*, *miehinä*, *miehissä*, *miehistä*, *miehittä*, *miehiä*, *mies*, *miesten*, *miestä* — “man” (42);

ohueen, *ohueksi*, *ohuella*, *ohuelle*, *ohuelta*, *ohuen*, *ohuena*, *ohuessa*, *ohuesta*, *ohuet*, *ohuetta*, *ohuiden*, *ohuihin*, *ohuiksi*, *ohuilla*, *ohuille*, *ohuilta*, *ohuin*, *ohuina*, *ohuine*, *ohuisiin*, *ohuissa*, *ohuista*, *ohuita*, *ohuitta*, *ohuitten*, *ohut*, *ohutta* — “thin” (43);

keväiden, *keväihin*, *keväiksi*, *keväille*, *keväillä*, *keväiltä*, *keväin*, *keväineen*, *keväinä*, *keväisiin*, *keväissä*, *keväistä*, *keväitten*, *keväittä*, *keväitä*, *kevät*, *kevättä*, *kevääksi*, *kevää̈lle*, *kevää̈llä*, *kevää̈ltä*, *kevään*, *kevää̈nä*, *kevääseen*, *kevää̈ssä*, *kevää̈stä*, *kevää̈t*, *kevää̈ttä* — “spring” (44);

kahdeksanneksi, *kahdeksannella*, *kahdeksannelle*, *kahdeksannelta*, *kahdeksannen*, *kahdeksannessa*, *kahdeksannesta*, *kahdeksannet*, *kahdeksannetta*, *kahdeksansia*, *kahdeksansien*, *kahdeksansiin*, *kahdeksansiksi*, *kahdeksansilla*, *kahdeksansille*, *kahdeksansilta*, *kahdeksansin*, *kahdeksansina*, *kahdeksansineen*, *kahdeksansissa*, *kahdeksansista*, *kahdeksansitta*, *kahdeksanteen*, *kahdeksantena*, *kahdeksas*, *kahdeksatta* — “eighth” (45);

tuhanneksi, *tuhannella*, *tuhannelle*, *tuhannelta*, *tuhannen*, *tuhannessa*, *tuhannesta*, *tuhannet*, *tuhannetta*, *tuhansia*, *tuhansien*, *tuhansiin*, *tuhansiksi*, *tuhansilla*, *tuhansille*, *tuhansilta*, *tuhansin*, *tuhansina*, *tuhansine*, *tuhansissa*, *tuhansista*, *tuhansitta*, *tuhanteen*, *tuhanten*, *tuhantena*, *tuhat*, *tuhatta* — "thousand" (46);

kuolleeksi, *kuolleella*, *kuolleelle*, *kuolleelta*, *kuolleen*, *kuolleena*, *kuolleeseen*, *kuolleessa*, *kuolleesta*, *kuolleet*, *kuolleetta*, *kuolleiden*, *kuolleihin*, *kuolleiksi*, *kuolleilla*, *kuolleille*, *kuolleilta*, *kuollein*, *kuolleina*, *kuolleine*, *kuolleisiin*, *kuolleissa*, *kuolleista*, *kuolleita*, *kuolleitta*, *kuolleitten*, *kuollut*, *kuollutta* — "dead" (47);

hame, *hameeksi*, *hameella*, *hameelle*, *hameelta*, *hameen*, *hameena*, *hameeseen*, *hameessa*, *hameesta*, *hameet*, *hameetta*, *hameiden*, *hameihin*, *hameiksi*, *hameilla*, *hameille*, *hameilta*, *hamein*, *hameina*, *hameineen*, *hameisiin*, *hameissa*, *hameista*, *hameita*, *hameitta*, *hameitten*, *hametta* — "skirt" (48);

askel, *askele*, *askeleeksi*, *askeleella*, *askeleelle*, *askeleelta*, *askeleen*, *askeleena*, *askeleeseen*, *askeleessa*, *askeleesta*, *askeleet*, *askeleetta*, *askeleiden*, *askeleiksi*, *askeleilla*, *askeleille*, *askeleilta*, *askelein*, *askeleina*, *askeleineen*, *askeleisiin*, *askeleissa*, *askeleista*, *askeleita*, *askeleitta*, *askeleitten*, *askeleksi*, *askelella*, *askelelle*, *askelelta*, *askelen*, *askelena*, *askelessa*, *askelesta*, *askelet*, *askeletta*, *askelia*, *askelien*, *askeliin*, *askeliksi*, *askelilla*, *askelille*, *askelilta*, *askelin*, *askelina*, *askelineen*, *askelissa*, *askelista*, *askelitta*, *askelta*, *askelten* — "step" (49);

isoäideiksi, *isoäideille*, *isoäideillä*, *isoäideiltä*, *isoäidein*, *isoäideissä*, *isoäideistä*, *isoäideittä*, *isoäidiksi*, *isoäidille*, *isoäidillä*, *isoäidiltä*, *isoäidin*, *isoäidissä*, *isoäidistä*, *isoäidit*, *isoäidittä*, *isoäiteihin*, *isoäiteineen*, *isoäiteinä*, *isoäitejä*, *isoäiti*, *isoäitien*, *isoäitiin*, *isoäitinä*, *isoäitiä* — "grandmother" (50);

nuoreenpariin, *nuoreksipariksi*, *nuorellaparilla*, *nuorelleparille*, *nuoreltaparilta*, *nuorenaparina*, *nuorenparin*, *nuoressaparissa*, *nuorestaparista*, *nuoretparit*, *nuorettaparitta*, *nuoriapareja*, *nuorienparien*, *nuoriinpareihin*, *nuoriksipareiksi*, *nuorillapareilla*, *nuorillepareille*, *nuoriltapareilta*, *nuorinapareina*, *nuorinepareineen*, *nuorinparein*, *nuoripari*, *nuorissapareissa*, *nuoristapareista*, *nuorittapareitta*, *nuortaparia*, *nuortenparien* — "newly-wed couple" (51).

Throwing all the words above to our algorithm, we receive the following output:

{*valo*, *valoa*, *valoihin*, *valoiksi*, *valoilla*, *valoille*, *valoilta*, *valoin*, *valoina*, *valoineen*, *valoissa*, *valoista*, *valoitta*, *valoja*, *valojen*, *valoksi*, *valolla*, *valolle*, *valolta*, *valon*, *valona*, *valoon*, *valossa*, *valosta*, *valot*, *valotta*} — (1) 26 words;

{*palvelu*, *palvelua*, *palveluiden*, *palveluihin*, *palveluiksi*, *palveluilla*, *palveluille*, *palveluilta*, *palveluin*, *palveluina*, *palveluineen*, *palveluissa*, *palveluista*, *palveluita*, *palveluitta*, *palveluitten*, *palveluja*, *palvelujen*, *palveluksi*, *palvelulla*, *palvelulle*, *palvelulta*, *palvelun*, *palveluna*, *palvelussa*, *palvelusta*, *palvelut*, *palvelutta*, *palveluun*} — (2) 29 words;

{*valtio*, *valtioiden*, *valtioihin*, *valtioiksi*, *valtioilla*, *valtioille*, *valtioilta*, *valtioin*, *valtioina*, *valtioineen*, *valtioissa*, *valtioista*, *valtioita*, *valtioitta*, *valtioitten*, *valtioksi*, *valtiolla*, *valtiolle*, *valtiolta*, *valtion*, *valtiona*, *valtioon*, *valtiossa*, *valtiosta*, *valtiot*, *valtiota*, *valtiotta*} — (3) 27 words;

{*laatikko*, *laatikkoa*, *laatikkoihin*, *laatikkoina*, *laatikkoineen*, *laatikkoja*, *laatikkojen*, *laatikkona*, *laatikkoon*, *laatikoiden*, *laatikoihin*, *laatikoiksi*, *laatikoilla*, *laatikoille*, *laatikoilta*, *laatikoin*, *laatikoissa*, *laatikoista*, *laatikoita*, *laatikoitta*, *laatikoitten*, *laatikoksi*, *laatikolla*, *laatikolle*, *laatikolta*, *laatikon*, *laatikossa*, *laatikosta*, *laatikot*, *laatikotta*} — (4) 30 words;

{*risteihin*, *risteiksi*, *risteille*, *risteillä*, *risteiltä*, *ristein*, *risteineen*, *risteinä*, *risteissä*, *risteistä*, *risteittä*, *ristejä*, *risti*, *ristien*, *ristiin*, *ristiksi*, *ristille*, *ristillä*, *ristiltä*, *ristin*, *ristinä*, *ristissä*, *rististä*, *ristit*, *ristittä*, *ristiä*} — (5) 26 words;

{*papereiden*, *papereihin*, *papereiksi*, *papereilla*, *papereille*, *papereilta*, *paperein*, *papereina*, *papereineen*, *papereissa*, *papereista*, *papereita*, *papereitta*, *papereitten*, *papereja*, *paperi*, *paperia*, *paperien*, *paperiin*, *paperiksi*, *paperilla*, *paperille*, *paperilta*, *paperin*, *paperina*, *paperissa*, *paperista*, *paperit*, *paperitta*} — (6) 29 words;

{*ovea*, *oveen*, *oveksi*, *ovella*, *ovelle*, *ovelta*, *oven*, *ovena*, *ovessa*, *ovesta*, *ovet*, *ovetta*, *ovi*, *ovia*, *ovien*, *oviin*, *oviksi*, *ovilla*, *oville*, *ovilta*, *ovin*, *ovina*, *ovineen*, *ovissa*, *ovista*, *ovitta*} — (7) 26 words;

{*nalle*, *nallea*, *nalleen*, *nalleihin*, *nalleiksi*, *nalleilla*, *nalleille*, *nalleilta*, *nallein*, *nalleina*, *nalleineen*, *nalleissa*, *nalleista*, *nalleitta*, *nalleja*, *nallejen*, *nalleksi*, *nallella*, *nallelle*, *nallelta*, *nallen*, *nallena*, *nallessa*, *nallesta*, *nallet*, *nalletta*} — (8) 26 words;

{*kala*, *kalaa*, *kalaan*, *kalain*, *kalaksi*, *kalalla*, *kalalle*, *kalalta*, *kalan*, *kalana*, *kalassa*, *kalasta*, *kalat*, *kalatta*, *kaloihin*, *kaloiksi*, *kaloilla*, *kaloille*, *kaloilta*, *kaloin*, *kaloina*, *kaloineen*, *kaloissa*, *kaloista*, *kaloitta*, *kaloja*, *kalojen*} — (9) 27 words;

{*koira*, *koiraa*, *koiraan*, *koirain*, *koiraksi*, *koiralla*, *koiralle*, *koiralta*, *koiran*, *koirana*, *koirassa*, *koirasta*, *koirat*, *koiratta*, *koiria*, *koirien*, *koiriin*, *koiriksi*, *koirilla*, *koirille*, *koirilta*, *koirin*, *koirina*, *koirineen*, *koirissa*, *koirista*, *koiritta*} — (10) 27 words;

{*omena*, *omenaa*, *omenaan*, *omenain*, *omenaksi*, *omenalla*, *omenalle*, *omenalta*, *omenan*, *omenana*, *omenassa*, *omenasta*, *omenat*, *omenatta*, *omenia*, *omenien*, *omeniin*, *omeniksi*, *omenilla*, *omenille*, *omenilta*, *omenin*, *omenina*, *omenineen*, *omenissa*, *omenista*, *omenitta*, *omenoiden*, *omenoihin*, *omenoiksi*, *omenoilla*, *omenoille*, *omenoilta*, *omenoin*, *omenoina*, *omenoineen*, *omenoissa*, *omenoista*, *omenoita*, *omenoitta*, *omenoitten*, *omenoja*, *omenojen*} — (11) 43 words;

{*kulkija*, *kulkijaa*, *kulkijaan*, *kulkijain*, *kulkijaksi*, *kulkijalla*, *kulkijalle*, *kulkijalta*, *kulkijan*, *kulkijana*, *kulkijassa*, *kulkijasta*, *kulkijat*, *kulkijatta*, *kulkijoiden*, *kulkijoihin*, *kulkijoiksi*, *kulkijoilla*, *kulkijoille*, *kulkijoilta*, *kulkijoin*, *kulkijoina*, *kulkijoineen*, *kulkijoissa*, *kulkijoista*, *kulkijoita*, *kulkijoitta*, *kulkijoitten*} — (12) 28 words;

{*katiska*, *katiskaa*, *katiskaan*, *katiskain*, *katiskaksi*, *katiskalla*, *katiskalle*, *katiskalta*, *katiskan*, *katiskana*, *katiskassa*, *katiskasta*, *katiskat*, *katiskatta*, *katiskoiden*, *katiskoihin*, *katiskoiksi*, *katiskoilla*, *katiskoille*, *katiskoilta*, *katiskoin*, *katiskoina*, *katiskoineen*, *katiskoissa*, *katiskoista*, *katiskoita*, *katiskoitta*, *katiskoitten*, *katiskoja*, *katiskojen*} — (13) 30 words;

{*solakaksi*, *solakalla*, *solakalle*, *solakalta*, *solakan*, *solakassa*, *solakasta*, *solakat*, *solakatta*, *solakka*, *solakkaa*, *solakkaan*, *solakkain*, *solakkana*, *solakkoihin*, *solakkoina*, *solakkoineen*, *solakkoja*, *solakkojen*, *solakoiden*, *solakoihin*, *solakoiksi*, *solakoilla*, *solakoille*, *solakoilta*, *solakoin*, *solakoissa*, *solakoista*, *solakoita*, *solakoitta*, *solakoitten*} — (14) 31 words;

{*korkea*, *korkeaa*, *korkeaan*, *korkeain*, *korkeaksi*, *korkealla*, *korkealle*, *korkealta*, *korkean*, *korkeana*, *korkeassa*, *korkeasta*, *korkeat*, *korkeata*, *korkeatta*, *korkeiden*, *korkeihin*, *korkeiksi*, *korkeilla*, *korkeille*, *korkeilta*, *korkein*, *korkeina*, *korkeine*, *korkeisiin*, *korkeissa*, *korkeista*, *korkeita*, *korkeitta*, *korkeitten*} — (15) 30 words;

{*vanhemmaksi*, *vanhemmalla*, *vanhemmalle*, *vanhemmalta*, *vanhemman*, *vanhemmassa*, *vanhemmasta*, *vanhemmat*, *vanhemmatta*, *vanhemmiksi*, *vanhemmilla*, *vanhemmille*, *vanhemmilta*, *vanhemmin*, *vanhemmissa*, *vanhemmista*, *vanhemmitta*, *vanhempaa*, *vanhempaan*, *vanhempain*, *vanhempana*, *vanhempi*, *vanhempia*, *vanhempien*, *vanhempiin*, *vanhempina*, *vanhempine*} — (16) 27 words;

{*vapaa*, *vapaaksi*, *vapaalla*, *vapaalle*, *vapaalta*, *vapaan*, *vapaana*, *vapaaseen*, *vapaassa*, *vapaasta*, *vapaat*, *vapaata*, *vapaatta*, *vapaiden*, *vapaihin*, *vapaiksi*, *vapailla*, *vapaille*, *vapailta*, *vapain*, *vapaina*, *vapaine*, *vapaisiin*, *vapaissa*, *vapaista*, *vapaita*, *vapaitta*, *vapaitten*} — (17) 28 words;

{*maineen*} — (18) 1 word;

{*maa*, *maahan*, *maaksi*, *maalla*, *maalle*, *maalta*, *maan*, *maana*, *maassa*, *maasta*, *maat*, *maata*, *maatta*, *maiden*, *maihin*, *maiksi*, *mailla*, *maille*, *mailta*, *main*, *maina*, *maissa*, *maista*, *maita*, *maitta*, *maitten*} — (18) 26 words;

{*soiden*, *soihin*, *soiksi*, *soilla*, *soille*, *soilta*, *soin*, *soina*, *soineen*, *soissa*, *soista*, *soita*, *soitta*, *soitten*, *suo*, *suohon*, *suoksi*, *suolla*, *suolle*, *suolta*, *suon*, *suona*, *suossa*, *suosta*, *suot*, *suota*, *suotta*} — (19) 27 words;

{*filee*, *fileehen*, *fileeksi*, *fileelle*, *fileellä*, *fileeltä*, *fileen*, *fileenä*, *fileeseen*, *fileessä*, *fileestä*, *fileet*, *fileettä*, *fileetä*, *fileiden*, *fileihin*, *fileiksi*, *fileille*, *fileillä*, *fileiltä*, *filein*, *fileineen*, *fileinä*, *fileisiin*, *fileissä*, *fileistä*, *fileitten*, *fileittä*, *fileitä*} — (20) 29 words;

{*rosé*, *roséhen*, *roséiden*, *roséihin*, *roséiksi*, *roséilla*, *roséille*, *roséilta*, *roséin*, *roséina*, *roséineen*, *roséissa*, *roséista*, *roséita*, *roséitta*, *roséitten*, *roséksi*, *rosélla*, *rosélle*, *rosélta*, *rosén*, *roséna*, *roséssa*, *rosésta*, *rosét*, *roséta*, *rosétta*} — (21) 27 words;

{*parfait*, *parfait'hen*, *parfait'iden*, *parfait'ihin*, *parfait'iksi*, *parfait'illa*, *parfait'ille*, *parfait'ilta*, *parfait'in*, *parfait'ina*, *parfait'ineen*, *parfait'issa*, *parfait'ista*, *parfait'ita*, *parfait'itta*, *parfait'itten*, *parfait'ksi*, *parfait'lla*, *parfait'lle*, *parfait'lta*, *parfait'n*, *parfait'na*, *parfait'ssa*, *parfait'sta*, *parfait't*, *parfait'ta*, *parfait'tta*} — (22) 27 words;

{*tiileen*, *tiileksi*, *tiilelle*, *tiilellä*, *tiileltä*, *tiilen*, *tiilenä*, *tiilessä*, *tiilestä*, *tiilet*, *tiilettä*, *tiili*, *tiilien*, *tiiliin*, *tiiliksi*, *tiilille*, *tiilillä*, *tiililtä*, *tiilin*, *tiilineen*, *tiilinä*, *tiilissä*, *tiilistä*, *tiilittä*, *tiiliä*, *tiiltä*} — (23) 26 words;

{*uneen*, *uneksi*, *unella*, *unelle*, *unelta*, *unen*, *unena*, *unessa*, *unesta*, *unet*, *unetta*, *uni*, *unia*, *unien*, *uniin*, *uniksi*, *unilla*, *unille*, *unilta*, *unin*, *unina*, *unineen*, *unissa*, *unista*, *unitta*, *unta*, *unten*} — (24) 27 words;

{*toimea*, *toimeen*, *toimeksi*, *toimella*, *toimelle*, *toimelta*, *toimen*, *toimena*, *toimessa*, *toimesta*, *toimet*, *toimetta*, *toimi*, *toimia*, *toimien*, *toimiin*, *toimiksi*, *toimilla*, *toimille*, *toimilta*, *toimin*, *toimina*, *toimineen*, *toimissa*, *toimista*, *toimitta*, *tointa*, *tointen*} — (25) 28 words;

{*pieneen*, *pieneksi*, *pienelle*, *pienellä*, *pieneltä*, *pienen*, *pienenä*, *pienessä*, *pienestä*, *pienet*, *pienettä*, *pieni*, *pienien*, *pieniin*, *pieniksi*, *pienille*, *pienillä*, *pieniltä*, *pienin*, *pienine*, *pieninä*, *pienissä*, *pienistä*, *pienittä*, *pieniä*, *pienten*, *pientä*} — (26) 27 words;

{*kädeksi*, *kädelle*, *kädellä*, *kädeltä*, *käden*, *kädessä*, *kädestä*, *kädet*, *kädettä*, *käsi*, *käsien*, *käsiin*, *käsiksi*, *käsille*, *käsillä*, *käsiltä*, *käsin*, *käsineen*, *käsinä*, *käsissä*, *käsistä*, *käsittä*, *käsiä*, *käteen*, *kätenä*, *kätten*, *kättä*} — (27) 27 words;

{*kynneksi*$_{28}$, *kynnelle*$_{28}$, *kynnellä*$_{28}$, *kynneltä*$_{28}$, *kynnen*$_{28}$, *kynnessä*$_{28}$, *kynnestä*$_{28}$, *kynnet*$_{28}$, *kynnettä*$_{28}$, *kynsi*$_{28}$, *kynsien*$_{28}$, *kynsiin*$_{28}$, *kynsiksi*$_{28}$, *kynsille*$_{28}$, *kynsillä*$_{28}$, *kynsiltä*$_{28}$, *kynsin*$_{28}$, *kynsineen*$_{28}$, *kynsinä*$_{28}$, *kynsissä*$_{28}$, *kynsistä*$_{28}$, *kynsittä*$_{28}$, *kynsiä*$_{28}$, *kynteen*$_{28}$, *kyntenä*$_{28}$, *kyntten*$_{28}$, *kynttä*$_{28}$, *kytkin*$_{33}$} — (28) 27 words; (33) 1 word;

{*lapseen*, *lapseksi*, *lapsella*, *lapselle*, *lapselta*, *lapsen*, *lapsena*, *lapsessa*, *lapsesta*, *lapset*, *lapsetta*, *lapsi*, *lapsia*, *lapsien*, *lapsiin*, *lapsiksi*, *lapsilla*, *lapsille*, *lapsilta*, *lapsin*, *lapsina*, *lapsineen*, *lapsissa*, *lapsista*, *lapsitta*} — (29) 25 words;

{*veisten*, *veistä*, *veitseen*, *veitseksi*, *veitselle*, *veitsellä*, *veitseltä*, *veitsen*, *veitsenä*, *veitsessä*, *veitsestä*, *veitset*, *veitsettä*, *veitsi*, *veitsien*, *veitsiin*, *veitsiksi*, *veitsille*, *veitsillä*, *veitsiltä*, *veitsin*, *veitsineen*, *veitsinä*, *veitsissä*, *veitsistä*, *veitsittä*, *veitsiä*} — (30) 27 words;

{*kahdeksi*, *kahdella*, *kahdelle*, *kahdelta*, *kahden*, *kahdessa*, *kahdesta*, *kahdesti*, *kahdet*, *kahdetta*, *kahta*, *kahtaalla*, *kahtaalle*, *kahtaalta*, *kahteen*, *kahtena*, *kahtia*, *kaksi*, *kaksia*, *kaksien*, *kaksiin*, *kaksiksi*, *kaksilla*, *kaksille*, *kaksilta*, *kaksin*, *kaksina*, *kaksine*, *kaksissa*, *kaksista*, *kaksitta*, *kaksittain*} — (31) 32 words;

{*sisar*, *sisareen*, *sisareksi*, *sisarella*, *sisarelle*, *sisarelta*, *sisaren*, *sisarena*, *sisaressa*, *sisaresta*, *sisaret*, *sisaretta*, *sisaria*, *sisarien*, *sisariin*, *sisariksi*, *sisarilla*, *sisarille*, *sisarilta*, *sisarin*, *sisarina*, *sisarineen*, *sisarissa*, *sisarista*, *sisaritta*, *sisarta*, *sisarten*} — (32) 27 words;

{*kytkimeen*, *kytkimeksi*, *kytkimelle*, *kytkimellä*, *kytkimeltä*, *kytkimen*, *kytkimenä*, *kytkimessä*, *kytkimestä*, *kytkimet*, *kytkimettä*, *kytkimien*, *kytkimiin*, *kytkimiksi*, *kytkimille*, *kytkimillä*, *kytkimiltä*, *kytkimin*, *kytkimineen*, *kytkiminä*, *kytkimissä*, *kytkimistä*, *kytkimittä*, *kytkimiä*, *kytkinten*, *kytkintä*} — (33) 26 words;

{*onneton*, *onnetonta*, *onnetonten*, *onnettomaan*, *onnettomaksi*, *onnettomalla*, *onnettomalle*, *onnettomalta*, *onnettoman*, *onnettomana*, *onnettomassa*, *onnettomasta*, *onnettomat*, *onnettomatta*, *onnettomia*, *onnettomien*, *onnettomiin*, *onnettomiksi*, *onnettomilla*, *onnettomille*, *onnettomilta*, *onnettomin*, *onnettomina*, *onnettomine*, *onnettomissa*, *onnettomista*, *onnettomitta*} — (34) 27 words;

{*lämmin*, *lämmintä*, *lämpimien*, *lämpimiin*, *lämpimiksi*, *lämpimille*, *lämpimillä*, *lämpimiltä*, *lämpimin*, *lämpimine*, *lämpiminä*, *lämpimissä*, *lämpimistä*, *lämpimittä*, *lämpimiä*, *lämpimäin*, *lämpimäksi*, *lämpimälle*, *lämpimällä*, *lämpimältä*, *lämpimän*, *lämpimänä*, *lämpimässä*, *lämpimästä*, *lämpimät*, *lämpimättä*, *lämpimään*} — (35) 27 words;

{*sisimmiksi*, *sisimmille*, *sisimmillä*, *sisimmiltä*, *sisimmin*, *sisimmissä*, *sisimmistä*, *sisimmittä*, *sisimmäksi*, *sisimmälle*, *sisimmällä*, *sisimmältä*, *sisimmän*, *sisimmässä*, *sisimmästä*, *sisimmät*, *sisimmättä*, *sisimpien*, *sisimpiin*, *sisimpine*, *sisimpinä*, *sisimpiä*, *sisimpäin*, *sisimpänä*, *sisimpään*, *sisin*, *sisinten*, *sisintä*} — (36) 28 words;

{*vasemmaksi*, *vasemmalla*, *vasemmalle*, *vasemmalta*, *vasemman*, *vasemmassa*, *vasemmasta*, *vasemmat*, *vasemmatta*, *vasemmiksi*, *vasemmilla*, *vasemmille*, *vasemmilta*, *vasemmin*, *vasemmissa*, *vasemmista*, *vasemmitta*, *vasempaa*, *vasempaan*, *vasempain*, *vasempana*, *vasempia*, *vasempien*, *vasempiin*, *vasempina*, *vasempine*, *vasen*, *vasenta*, *vasenten*} — (37) 29 words;

{*naisin*} — (38) 1 word;

{*nainen*, *naiseen*, *naiseksi*, *naisella*, *naiselle*, *naiselta*, *naisen*, *naisena*, *naisessa*, *naisesta*, *naiset*, *naisetta*, *naisia*, *naisien*, *naisiin*, *naisiksi*, *naisilla*, *naisille*, *naisilta*, *naisina*, *naisineen*, *naisissa*, *naisista*, *naisitta*, *naista*, *naisten*} — (38) 26 words;

{*vastaukseen*, *vastaukseksi*, *vastauksella*, *vastaukselle*, *vastaukselta*, *vastauksen*, *vastauksena*, *vastauksessa*, *vastauksesta*, *vastaukset*, *vastauksetta*, *vastauksia*, *vastauksien*, *vastauksiin*, *vastauksiksi*, *vastauksilla*, *vastauksille*, *vastauksilta*, *vastauksin*, *vastauksina*, *vastauksineen*, *vastauksissa*, *vastauksista*, *vastauksitta*, *vastaus*, *vastausta*, *vastausten*} — (39) 27 words;

{*kalleudeksi*, *kalleudella*, *kalleudelle*, *kalleudelta*, *kalleuden*, *kalleudessa*, *kalleudesta*, *kalleudet*, *kalleudetta*, *kalleuksia*, *kalleuksien*, *kalleuksiin*, *kalleuksiksi*, *kalleuksilla*, *kalleuksille*, *kalleuksilta*, *kalleuksin*, *kalleuksina*, *kalleuksineen*, *kalleuksissa*, *kalleuksista*, *kalleuksitta*, *kalleus*, *kalleuteen*, *kalleutena*, *kalleutta*} — (40) 26 words;

{*vieraaksi*, *vieraalla*, *vieraalle*, *vieraalta*, *vieraan*, *vieraana*, *vieraaseen*, *vieraassa*, *vieraasta*, *vieraat*, *vieraatta*, *vieraiden*, *vieraihin*, *vieraiksi*, *vierailla*, *vieraille*, *vierailta*, *vierain*, *vieraina*, *vieraine*, *vieraisiin*, *vieraissa*, *vieraista*, *vieraita*, *vieraitta*, *vieraitten*, *vieras*, *vierasta*} — (41) 28 words;

{*mieheen*, *mieheksi*, *miehelle*, *miehellä*, *mieheltä*, *miehen*, *miehenä*, *miehessä*, *miehestä*, *miehet*, *miehettä*, *miehien*, *miehiin*, *miehiksi*, *miehille*, *miehillä*, *miehiltä*, *miehin*, *miehineen*, *miehinä*, *miehissä*, *miehistä*, *miehittä*, *miehiä*, *mies*, *miesten*, *miestä*} — (42) 27 words;

{*ohueen*, *ohueksi*, *ohuella*, *ohuelle*, *ohuelta*, *ohuen*, *ohuena*, *ohuessa*, *ohuesta*, *ohuet*, *ohuetta*, *ohuiden*, *ohuihin*, *ohuiksi*, *ohuilla*, *ohuille*, *ohuilta*, *ohuin*, *ohuina*, *ohuine*, *ohuisiin*, *ohuissa*, *ohuista*, *ohuita*, *ohuitta*, *ohuitten*, *ohut*, *ohutta*} — (43) 28 words;

{*keväiden*, *keväihin*, *keväiksi*, *keväille*, *keväillä*, *keväiltä*, *keväin*, *keväineen*, *keväinä*, *keväisiin*, *keväissä*, *keväistä*, *keväitten*, *keväittä*, *keväitä*, *kevät*, *kevättä*, *kevääksi*, *keväälle*, *keväällä*, *keväältä*, *kevään*, *keväänä*, *kevääseen*, *keväässä*, *keväästä*, *kevään*, *kevättä*} — (44) 28 words;

{*kahdeksanneksi*, *kahdeksannella*, *kahdeksannelle*, *kahdeksannelta*, *kahdeksannen*, *kahdeksannessa*, *kahdeksannesta*, *kahdeksannet*, *kahdeksannetta*, *kahdeksansia*, *kahdeksansien*, *kahdeksansiin*, *kahdeksansiksi*, *kahdeksansilla*, *kahdeksansille*, *kahdeksansilta*, *kahdeksansin*, *kahdeksansina*, *kahdeksansineen*, *kahdeksansissa*, *kahdeksansista*, *kahdeksansitta*, *kahdeksanteen*, *kahdeksantena*, *kahdeksas*, *kahdeksatta*} — (45) 26 words;

{*tuhanneksi*, *tuhannella*, *tuhannelle*, *tuhannelta*, *tuhannen*, *tuhannessa*, *tuhannesta*, *tuhannet*, *tuhannetta*, *tuhansia*, *tuhansien*, *tuhansiin*, *tuhansiksi*, *tuhansilla*, *tuhansille*, *tuhansilta*, *tuhansin*, *tuhansina*, *tuhansine*, *tuhansissa*, *tuhansista*, *tuhansitta*, *tuhanteen*, *tuhanten*, *tuhantena*, *tuhat*, *tuhatta*} — (46) 27 words;

{*kuolleeksi*, *kuolleella*, *kuolleelle*, *kuolleelta*, *kuolleen*, *kuolleena*, *kuolleeseen*, *kuolleessa*, *kuolleesta*, *kuolleet*, *kuolleetta*, *kuolleiden*, *kuolleihin*, *kuolleiksi*, *kuolleilla*, *kuolleille*, *kuolleilta*, *kuollein*, *kuolleina*, *kuolleine*, *kuolleisiin*, *kuolleissa*, *kuolleista*, *kuolleita*, *kuolleitta*, *kuolleitten*, *kuollut*, *kuollutta*} — (47) 28 words;

{*hame*, *hameeksi*, *hameella*, *hameelle*, *hameelta*, *hameen*, *hameena*, *hameeseen*, *hameessa*, *hameesta*, *hameet*, *hameetta*, *hameiden*, *hameihin*, *hameiksi*, *hameilla*, *hameille*, *hameilta*, *hamein*, *hameina*, *hameineen*, *hameisiin*, *hameissa*, *hameista*, *hameita*, *hameitta*, *hameitten*, *hametta*} — (48) 28 words;

{*askel*, *askele*, *askeleeksi*, *askeleella*, *askeleelle*, *askeleelta*, *askeleen*, *askeleena*, *askeleeseen*, *askeleessa*, *askeleesta*, *askeleet*, *askeleetta*, *askeleiden*, *askeleiksi*, *askeleilla*, *askeleille*, *askeleilta*, *askelein*, *askeleina*, *askeleineen*, *askeleisiin*, *askeleissa*, *askeleista*, *askeleita*, *askeleitta*, *askeleitten*, *askeleksi*, *askelella*, *askelelle*, *askelelta*, *askelen*, *askelena*, *askelessa*, *askelesta*, *askelet*, *askeletta*, *askelia*, *askelien*, *askeliin*, *askeliksi*, *askelilla*, *askelille*, *askelilta*, *askelin*, *askelina*, *askelineen*, *askelissa*, *askelista*, *askelitta*, *askelta*, *askelten*} — (49) 52 words;

{*isoäideiksi*, *isoäideille*, *isoäideillä*, *isoäideiltä*, *isoäidein*, *isoäideissä*, *isoäideistä*, *isoäideittä*, *isoäidiksi*, *isoäidille*, *isoäidillä*, *isoäidiltä*, *isoäidin*, *isoäidissä*, *isoäidistä*, *isoäidit*, *isoäiditä*, *isoäiteihin*, *isoäiteineen*, *isoäiteinä*, *isoäitejä*, *isoäiti*, *isoäitien*, *isoäitiin*, *isoäitinä*, *isoäitiä*} — (50) 26 words;

{*nuoreenpariin*} — (51) 1 word;

{*nuoreksipariksi*} — (51) 1 word;

{*nuorellaparilla*} — (51) 1 word;

{*nuorelleparille*} — (51) 1 word;

{*nuoreltaparilta*} — (51) 1 word;

{*nuorenaparina*} — (51) 1 word;

{*nuorenparin*} — (51) 1 word;

{*nuoretparit*} — (51) 1 word;

{*nuorettaparitta*} — (51) 1 word;

{*nuoriapareja*} — (51) 1 word;

{*nuorienparien*} — (51) 1 word;

{*nuoriinpareihin*} — (51) 1 word;

{*nuoriksipareiksi*} — (51) 1 word;

{*nuorinapareina*} — (51) 1 word;

{*nuorinepareineen*} — (51) 1 word;

{*nuorinparein*} — (51) 1 word;

{*nuoripari*} — (51) 1 word;

{*nuorittapareitta*} — (51) 1 word;

{*nuortaparia*} — (51) 1 word;

{*nuortenparien*} — (51) 1 word;

{*nuoressaparissa*, *nuorestaparista*} — (51) 2 words;

{*nuorissapareissa*, *nuoristapareista*} — (51) 2 words;

{*nuorillapareilla*, *nuorillepareille*, *nuoriltapareilta*} — (51) 3 words;

Here, as we present the results above, each cluster is followed by a brief description about the KOTUS affiliations of the words therein. When a cluster contains words from more than one KOTUS class, the words within the cluster are labeled with class numbers in subscripts.

The word *maineen* is both the comitative plural form of *maa* "earth" and the genitive singular form of *maine* "fame". Since the comitative case is rarely used, it is understandable that our algorithm separated *maineen* from the other inflected forms of *maa*.

For a similar reason, the word *naisin* was not treated as the instructive plural (a rarely used inflection) form of *nainen* "woman", but was considered as the first personal singular conditional present of *naida* "marry".

For more information regarding the compound word in KOTUS-51, see Example 8.14.1.

As we test our algorithm against the following sample verbs:

sano, *sanoa*, *sanoakseen*, *sanoen*, *sanoessa*, *sanoi*, *sanoimme*, *sanoin*, *sanoisi*, *sanoisimme*, *sanoisin*, *sanoisit*, *sanoisitte*, *sanoisivat*, *sanoit*, *sanoitte*, *sanoivat*, *sanokaa*, *sanokaamme*, *sanoko*, *sanokoon*, *sanokoot*, *sanoma*, *sanomaan*, *sanomaisillaan*, *sanomalla*, *sanoman*, *sanomassa*, *sanomasta*, *sanomaton*, *sanomatta*, *sanominen*, *sanomista*, *sanomme*, *sanon*, *sanone*, *sanonee*, *sanoneet*, *sanonemme*, *sanonen*, *sanonet*, *sanonette*, *sanonevat*, *sanonut*, *sanoo*, *sanot*, *sanota*, *sanotaan*, *sanottaessa*, *sanottaisi*, *sanottaisiin*, *sanottako*, *sanottakoon*, *sanottaman*, *sanottane*, *sanottaneen*, *sanottava*, *sanotte*, *sanottiin*, *sanottu*, *sanova*, *sanovat* — "tell" (52);

muista, *muistaa*, *muistaakseen*, *muistaen*, *muistaessa*, *muistaisi*, *muistaisimme*, *muistaisin*, *muistaisit*, *muistaisitte*, *muistaisivat*, *muistakaa*, *muistakaamme*, *muistako*, *muistakoon*, *muistakoot*, *muistama*, *muistamaan*, *muistamaisillaan*, *muistamalla*, *muistaman*, *muistamassa*, *muistamasta*, *muistamaton*, *muistamatta*, *muistaminen*, *muistamista*, *muistamme*, *muistan*, *muistane*, *muistanee*, *muistaneet*, *muistanemme*, *muistanen*, *muistanet*, *muistanette*, *muistanevat*, *muistanut*, *muistat*, *muistatte*, *muistava*, *muistavat*, *muisteta*, *muistetaan*, *muistettaessa*, *muistettaisi*, *muistettaisiin*, *muistettako*, *muistettakoon*, *muistettaman*, *muistettane*, *muistettaneen*, *muistettava*, *muistettiin*, *muistettu*, *muisti*, *muistimme*, *muistin*, *muistit*, *muistitte*, *muistivat* — "remember" (53);

huuda, *huudamme*, *huudan*, *huudat*, *huudatte*, *huudeta*, *huudetaan*, *huudettaessa*, *huudettaisi*, *huudettaisiin*, *huudettako*, *huudettakoon*, *huudettaman*, *huudettane*, *huudettaneen*, *huudettava*, *huudettiin*, *huudettu*, *huusi*, *huusimme*, *huusin*, *huusit*, *huusitte*, *huusivat*, *huutaa*, *huutaakseen*, *huutaen*, *huutaessa*, *huutaisi*, *huutaisimme*, *huutaisin*, *huutaisit*, *huutaisitte*, *huutaisivat*, *huutakaa*, *huutakaamme*, *huutako*, *huutakoon*, *huutakoot*, *huutama*, *huutamaan*, *huutamaisillaan*, *huutamalla*, *huutaman*, *huutamassa*, *huutamasta*, *huutamaton*, *huutamatta*, *huutaminen*, *huutamista*, *huutane*, *huutanee*, *huutaneet*, *huutanemme*, *huutanen*, *huutanet*, *huutanette*, *huutanevat*, *huutanut*, *huutava*, *huutavat* — "shout" (54);

souda, *soudamme*, *soudan*, *soudat*, *soudatte*, *soudeta*, *soudetaan*, *soudettaessa*, *soudettaisi*, *soudettaisiin*, *soudettako*, *soudettakoon*, *soudettaman*, *soudettane*, *soudettaneen*, *soudettava*, *soudettiin*, *soudettu*, *soudimme*, *soudin*, *soudit*, *souditte*, *sousi*, *sousimme*, *sousin*, *sousit*, *sousitte*, *sousivat*, *soutaa*, *soutaakseen*, *soutaen*, *soutaessa*, *soutaisi*, *soutaisimme*, *soutaisin*, *soutaisit*, *soutaisitte*, *soutaisivat*, *soutakaa*, *soutakaamme*, *soutako*, *soutakoon*, *soutakoot*, *soutama*, *soutamaan*, *soutamaisillaan*, *soutamalla*, *soutaman*, *soutamassa*, *soutamasta*, *soutamaton*, *soutamatta*, *soutaminen*, *soutamista*, *soutane*, *soutanee*, *soutaneet*, *soutanemme*, *soutanen*, *soutanet*, *soutanette*, *soutanevat*, *soutanut*, *soutava*, *soutavat*, *souti*, *soutivat* — "row" (55);

kaiva, *kaivaa*, *kaivaisi*, *kaivaisimme*, *kaivaisin*, *kaivaisit*, *kaivaisitte*, *kaivaisivat*, *kaivakaa*, *kaivakaamme*, *kaivako*, *kaivakoon*, *kaivakoot*, *kaivamme*, *kaivan*, *kaivaneet*, *kaivanut*, *kaivat*, *kaivatte*, *kaivavat*, *kaiveta*, *kaivetaan*, *kaivettaisi*, *kaivettaisiin*, *kaivettako*, *kaivettakoon*, *kaivettiin*, *kaivettu*, *kaivoi*, *kaivoimme*, *kaivoin*, *kaivoit*, *kaivoitte*, *kaivoivat* — "dig" (56);

saarra, *saarramme*, *saarran*, *saarrat*, *saarratte*, *saarreta*, *saarretaan*, *saarrettaessa*, *saarrettaisi*, *saarrettaisiin*, *saarrettako*, *saarrettakoon*, *saarrettaman*, *saarrettane*, *saarrettaneen*, *saarrettava*, *saarrettiin*, *saarrettu*, *saarroimme*, *saarroin*, *saarroit*, *saarroitte*, *saarsi*, *saarsimme*, *saarsin*, *saarsit*, *saarsitte*, *saarsivat*, *saartaa*, *saartaakseen*, *saartaen*, *saartaessa*, *saartaisi*, *saartaisimme*, *saartaisin*, *saartaisit*, *saartaisitte*, *saartaisivat*, *saartakaa*, *saartakaamme*, *saartako*, *saartakoon*, *saartakoot*, *saartama*, *saartamaan*, *saartamaisillaan*, *saartamalla*, *saartaman*, *saartamassa*, *saartamasta*, *saartamaton*, *saartamatta*, *saartaminen*, *saartamista*, *saartane*, *saartanee*, *saartaneet*, *saartanemme*, *saartanen*, *saartanet*, *saartanette*, *saartanevat*, *saartanut*, *saartava*, *saartavat*, *saartoi*, *saartoivat* — "surround" (57);

laske, *laskea*, *laskeakseen*, *laskee*, *laskekaa*, *laskekaamme*, *laskeko*, *laskekoon*, *laskekoot*, *laskema*, *laskemaan*, *laskemaisillaan*, *laskemalla*, *laskeman*, *laskemassa*, *laskemasta*, *laskematon*, *laskematta*, *laskeminen*, *laskemista*, *laskemme*, *lasken*, *laskene*, *laskenee*, *laskeneet*, *laskenemme*, *laskenen*, *laskenet*, *laskenette*, *laskenevat*, *laskenut*, *lasket*, *lasketa*, *lasketaan*, *laskettaessa*, *laskettaisi*, *laskettaisiin*, *laskettako*, *laskettakoon*, *laskettaman*, *laskettane*, *laskettaneen*, *laskettava*, *laskette*, *laskettiin*, *laskettu*, *laskeva*, *laskevat*, *laski*, *laskien*, *laskiessa*, *laskimme*, *laskin*, *laskisi*, *laskisimme*, *laskisin*, *laskisit*, *laskisitte*, *laskisivat*, *laskit*, *laskitte*, *laskivat* — "drop" (58);

tunne, *tunnemme*, *tunnen*, *tunnet*, *tunneta*, *tunnetaan*, *tunnettaessa*, *tunnettaisi*, *tunnettaisiin*, *tunnettako*, *tunnettakoon*, *tunnettaman*, *tunnettane*, *tunnettaneen*, *tunnettava*, *tunnette*, *tunnettiin*, *tunnettu*, *tunsi*, *tunsimme*, *tunsin*, *tunsit*, *tunsitte*, *tunsivat*, *tuntea*, *tunteakseen*, *tuntee*, *tuntekaa*, *tuntekaamme*, *tunteko*, *tuntekoon*, *tuntekoot*, *tuntema*, *tuntemaan*, *tuntemaisillaan*, *tuntemalla*, *tunteman*, *tuntemassa*, *tuntemasta*, *tuntematon*, *tuntematta*, *tuntеminen*, *tuntemista*, *tuntene*, *tuntenee*, *tunteneet*, *tuntenemme*, *tuntenen*, *tuntenet*, *tuntenette*, *tuntenevat*, *tuntenut*, *tunteva*, *tuntevat*, *tuntien*, *tuntiessa*, *tuntisi*, *tuntisimme*, *tuntisin*, *tuntisit*, *tuntisitte*, *tuntisivat* — "feel" (59);

lähde, *lähdemme*, *lähden*, *lähdet*, *lähdette*, *lähdettiin*, *lähdetty*, *lähdettäessä*, *lähdettäisi*, *lähdettäisiin*, *lähdettäkö*, *lähdettäköön*, *lähdettämän*, *lähdettäne*, *lähdettäneen*, *lähdettävä*, *lähdetä*, *lähdetään*, *lähdimme*, *lähdin*, *lähdit*, *lähditte*, *lähtee*, *lähtekää*, *lähtekäämme*, *lähtekö*, *lähteköön*, *lähtekööt*, *lähteminen*, *lähtemistä*, *lähtemä*, *lähtemäisillään*, *lähtemällä*, *lähtemän*, *lähtemässä*, *lähtemästä*, *lähtemättä*, *lähtemätön*, *lähtemään*, *lähtene*, *lähtenee*, *lähteneet*, *lähtenemme*, *lähtenen*, *lähtenet*, *lähtenette*, *lähtenevät*, *lähtenyt*, *lähtevä*, *lähtevät*, *lähteä*, *lähteäkseen*, *lähti*, *lähtien*, *lähtiessä*, *lähtisi*, *lähtisimme*, *lähtisin*, *lähtisit*, *lähtisitte*, *lähtisivät*, *lähtivät* — "leave" (60);

salli, *sallia*, *salliakseen*, *sallien*, *salliessa*, *sallii*, *sallikaa*, *sallikaamme*, *salliko*, *sallikoon*, *sallikoot*, *sallima*, *sallimaan*, *sallimaisillaan*, *sallimalla*, *salliman*, *sallimassa*, *sallimasta*, *sallimaton*, *sallimatta*, *salliminen*, *sallimista*, *sallimme*, *sallin*, *salline*, *sallinee*, *sallineet*, *sallinemme*, *sallinen*, *sallinet*, *sallinette*, *sallinevat*, *sallinut*, *sallisi*, *sallisimme*, *sallisin*, *sallisit*, *sallisitte*, *sallisivat*, *sallit*, *sallita*, *sallitaan*, *sallittaessa*, *sallittaisi*, *sallittaisiin*, *sallittako*, *sallittakoon*, *sallittaman*, *sallittane*, *sallittaneen*, *sallittava*, *sallitte*, *sallittiin*, *sallittu*, *salliva*, *sallivat* — "allow" (61);

voi, *voida*, *voidaan*, *voidakseen*, *voiden*, *voidessa*, *voikaa*, *voikaamme*, *voiko*, *voikoon*, *voikoot*, *voima*, *voimaan*, *voimaisillaan*, *voimalla*, *voiman*, *voimassa*, *voimasta*, *voimaton*, *voimatta*, *voiminen*, *voimista*, *voimme*, *voin*, *voine*, *voinee*, *voineet*, *voinemme*, *voinen*, *voinet*, *voinette*, *voinevat*, *voinut*, *voisi*, *voisimme*, *voisin*, *voisit*, *voisitte*,

voisivat, *voit*, *voitaessa*, *voitaisi*, *voitaisiin*, *voitako*, *voitakoon*, *voitaman*, *voitane*, *voitaneen*, *voitava*, *voitiin*, *voitte*, *voitu*, *voiva*, *voivat* — "can" (62);

saa, *saada*, *saadaan*, *saadakseen*, *saaden*, *saadessa*, *saakaa*, *saakaamme*, *saako*, *saakoon*, *saakoot*, *saama*, *saamaan*, *saamaisillaan*, *saamalla*, *saaman*, *saamassa*, *saamasta*, *saamaton*, *saamatta*, *saaminen*, *saamista*, *saamme*, *saan*, *saane*, *saanee*, *saaneet*, *saanemme*, *saanen*, *saanet*, *saanette*, *saanevat*, *saanut*, *saat*, *saataessa*, *saataisi*, *saataisiin*, *saatako*, *saatakoon*, *saataman*, *saatane*, *saataneen*, *saatava*, *saatiin*, *saatte*, *saatu*, *saava*, *saavat* — "receive" (63);

juo, *juoda*, *juodaan*, *juodakseen*, *juoden*, *juodessa*, *juokaa*, *juokaamme*, *juoko*, *juokoon*, *juokoot*, *juoma*, *juomaan*, *juomaisillaan*, *juomalla*, *juoman*, *juomassa*, *juomasta*, *juomaton*, *juomatta*, *juominen*, *juomista*, *juomme*, *juon*, *juone*, *juonee*, *juoneet*, *juonemme*, *juonen*, *juonet*, *juonette*, *juonevat*, *juonut*, *juot*, *juotaessa*, *juotaisi*, *juotaisiin*, *juotako*, *juotakoon*, *juotaman*, *juotane*, *juotaneen*, *juotava*, *juotiin*, *juotte*, *juotu*, *juova*, *juovat* — "drink" (64);

kävi, *kävimme*, *kävin*, *kävisi*, *kävisimme*, *kävisin*, *kävisit*, *kävisitte*, *kävisivät*, *kävit*, *kävitte*, *kävivät*, *käy*, *käyden*, *käydessä*, *käydä*, *käydäkseen*, *käydään*, *käykää*, *käykäämme*, *käykö*, *käyköön*, *käyköot*, *käyminen*, *käymistä*, *käymme*, *käymä*, *käymäisillään*, *käymällä*, *käymän*, *käymässä*, *käymästä*, *käymättä*, *käymätön*, *käymään*, *käyn*, *käyne*, *käynee*, *käyneet*, *käynemme*, *käynen*, *käynet*, *käynette*, *käynevät*, *käynyt*, *käyt*, *käytiin*, *käytte*, *käyty*, *käytäessä*, *käytäisi*, *käytäisiin*, *käytäkö*, *käytäköön*, *käytämän*, *käytäne*, *käytäneen*, *käytävä*, *käyvä*, *käyvät* — "go" (65);

rohkaise, *rohkaisee*, *rohkaisema*, *rohkaisemaan*, *rohkaisemaisillaan*, *rohkaisemalla*, *rohkaiseman*, *rohkaisemassa*, *rohkaisemasta*, *rohkaisematon*, *rohkaisematta*, *rohkaiseminen*, *rohkaisemista*, *rohkaisemme*, *rohkaisen*, *rohkaiset*, *rohkaisette*, *rohkaiseva*, *rohkaisevat*, *rohkaisi*, *rohkaisimme*, *rohkaisin*, *rohkaisisi*, *rohkaisisimme*, *rohkaisisin*, *rohkaisisit*, *rohkaisisitte*, *rohkaisisivat*, *rohkaisit*, *rohkaisitte*, *rohkaisivat*, *rohkaiskaa*, *rohkaiskaamme*, *rohkaisko*, *rohkaiskoon*, *rohkaiskoot*, *rohkaisse*, *rohkaissee*, *rohkaisseet*, *rohkaissemme*, *rohkaissen*, *rohkaisset*, *rohkaissette*, *rohkaissevat*, *rohkaissut*, *rohkaista*, *rohkaistaan*, *rohkaistaessa*, *rohkaistaisi*, *rohkaistaisiin*, *rohkaistako*, *rohkaistakoon*, *rohkaistakseen*, *rohkaistaman*, *rohkaistane*, *rohkaistaneen*, *rohkaistava*, *rohkaisten*, *rohkaistessa*, *rohkaistiin*, *rohkaistu* — "encourage" (66);

tule, *tulee*, *tulema*, *tulemaan*, *tulemaisillaan*, *tulemalla*, *tuleman*, *tulemassa*, *tulemasta*, *tulematon*, *tulematta*, *tuleminen*, *tulemista*, *tulemme*, *tulen*, *tulet*, *tulette*, *tuleva*, *tulevat*, *tuli*, *tulimme*, *tulin*, *tulisi*, *tulisimme*, *tulisin*, *tulisit*, *tulisitte*, *tulisivat*, *tulit*, *tulitte*, *tulivat*, *tulkaa*, *tulkaamme*, *tulko*, *tulkoon*, *tulkoot*, *tulla*, *tullaan*, *tullakseen*, *tulle*, *tullee*, *tulleet*, *tullemme*, *tullen*, *tullessa*, *tullet*, *tullette*, *tullevat*, *tullut*, *tultaessa*, *tultaisi*, *tultaisiin*, *tultako*, *tultakoon*, *tultaman*, *tultane*, *tultaneen*, *tultava*, *tultiin*, *tultu* — "come" (67);

tupakoi, *tupakoida*, *tupakoidaan*, *tupakoidakseen*, *tupakoiden*, *tupakoidessa*, *tupakoikaa*, *tupakoikaamme*, *tupakoiko*, *tupakoikoon*, *tupakoikoot*, *tupakoima*, *tupakoimaan*, *tupakoimaisillaan*, *tupakoimalla*, *tupakoiman*, *tupakoimassa*, *tupakoimasta*, *tupakoimaton*, *tupakoimatta*, *tupakoiminen*, *tupakoimista*, *tupakoimme*, *tupakoin*, *tupakoine*, *tupakoinee*, *tupakoineet*, *tupakoinemme*, *tupakoinen*, *tupakoinet*, *tupakoinette*, *tupakoinevat*, *tupakoinut*, *tupakoisi*, *tupakoisimme*, *tupakoisin*, *tupakoisit*, *tupakoisitte*, *tupakoisivat*, *tupakoit*, *tupakoitaessa*, *tupakoitaisi*, *tupakoitaisiin*, *tupakoitako*, *tupakoitakoon*, *tupakoitaman*, *tupakoitane*, *tupakoitaneen*, *tupakoitava*, *tupakoitiin*, *tupakoitse*, *tupakoitsee*, *tupakoitsema*, *tupakoitsemaan*, *tupakoitsemaisillaan*, *tupakoitsemalla*, *tupakoitseman*, *tupakoitsemassa*, *tupakoitsemasta*, *tupakoitsematon*, *tupakoitsematta*, *tupakoitseminen*, *tupakoitsemista*, *tupakoitsemme*, *tupakoitsen*, *tupakoitset*, *tupakoitsette*, *tupakoitseva*, *tupakoitsevat*, *tupakoitsi*, *tupakoitsimme*, *tupakoitsin*, *tupakoitsisi*, *tupakoitsisimme*, *tupakoitsisin*, *tupakoitsisit*, *tupakoitsisitte*, *tupakoitsisivat*, *tupakoitsit*, *tupakoitsitte*, *tupakoitsivat*, *tupakoitte*, *tupakoitu*, *tupakoiva*, *tupakoivat* — "smoke" (68);

valinne, *valinnee*, *valinneet*, *valinnemme*, *valinnen*, *valinnet*, *valinnette*, *valinnevat*, *valinnut*, *valita*, *valitaan*, *valitakseen*, *valiten*, *valitessa*, *valitkaa*, *valitkaamme*, *valitko*, *valitkoon*, *valitkoot*, *valitse*, *valitsee*, *valitsema*, *valitsemaan*, *valitsemaisillaan*, *valitsemalla*, *valitseman*, *valitsemassa*, *valitsemasta*, *valitsematon*, *valitsematta*, *valitseminen*, *valitsemista*, *valitsemme*, *valitsen*, *valitset*, *valitsette*, *valitseva*, *valitsevat*, *valitsi*, *valitsimme*, *valitsin*, *valitsisi*, *valitsisimme*, *valitsisin*, *valitsisit*, *valitsisitte*, *valitsisivat*, *valitsit*, *valitsitte*, *valitsivat*, *valittaessa*, *valittaisi*, *valittaisiin*, *valittako*, *valittakoon*, *valittaman*, *valittane*, *valittaneen*, *valittava*, *valittiin*, *valittu* — "choose" (69);

juokse, *juoksee*, *juoksema*, *juoksemaan*, *juoksemaisillaan*, *juoksemalla*, *juokseman*, *juoksemassa*, *juoksemasta*, *juoksematon*, *juoksematta*, *juokseminen*, *juoksemista*, *juoksemme*, *juoksen*, *juokset*, *juoksette*, *juokseva*, *juoksevat*, *juoksi*, *juoksimme*, *juoksin*, *juoksisi*, *juoksisimme*, *juoksisin*, *juoksisit*, *juoksisitte*, *juoksisivat*, *juoksit*, *juoksitte*, *juoksivat*, *juoskaa*, *juoskaamme*, *juosko*, *juoskoon*, *juoskoot*, *juosse*, *juossee*, *juosseet*, *juossemme*, *juossen*, *juosset*, *juossette*, *juossevat*, *juossut*, *juosta*, *juostaan*, *juostaessa*, *juostaisi*, *juostaisiin*, *juostako*, *juostakoon*, *juostakseen*, *juostaman*, *juostane*, *juostaneen*, *juostava*, *juosten*, *juostessa*, *juostiin*, *juostu* — "run" (70);

näe, *näemme*, *näen*, *näet*, *näette*, *nähden*, *nähdessä*, *nähdä*, *nähdäkseen*, *nähdään*, *nähkää*, *nähkäämme*, *nähkö*, *nähköön*, *nähkööt*, *nähne*, *nähnee*, *nähneet*, *nähnemme*, *nähnen*, *nähnet*, *nähnette*, *nähnevät*, *nähnyt*, *nähtiin*, *nähty*, *nähtäessä*, *nähtäisi*, *nähtäisiin*, *nähtäkö*, *nähtäköön*, *nähtämän*, *nähtäne*, *nähtäneen*, *nähtävä*, *näimme*, *näin*, *näit*, *näitte*, *näkee*, *näkeminen*, *näkemistä*, *näkemä*, *näkemäisillään*, *näkemällä*, *näkemän*, *näkemässä*, *näkemästä*, *näkemättä*, *näkemätön*, *näkemään*, *näkevä*, *näkevät*, *näki*, *näkisi*, *näkisimme*, *näkisin*, *näkisit*, *näkisitte*, *näkisivät*, *näkivät* — “see” (71);

vanhene, *vanhenee*, *vanhenema*, *vanhenemaan*, *vanhenemaisillaan*, *vanhenemalla*, *vanheneman*, *vanhenemassa*, *vanhenemasta*, *vanhenematon*, *vanhenematta*, *vanheneminen*, *vanhenemista*, *vanhenemme*, *vanhenen*, *vanhenet*, *vanhenette*, *vanheneva*, *vanhenevat*, *vanheni*, *vanhenimme*, *vanhenin*, *vanhenisi*, *vanhenisimme*, *vanhenisin*, *vanhenisit*, *vanhenisitte*, *vanhenisivat*, *vanhenit*, *vanhenitte*, *vanhenivat*, *vanhenne*, *vanhennee*, *vanhenneet*, *vanhennemme*, *vanhennen*, *vanhennet*, *vanhennette*, *vanhennevat*, *vanhennut*, *vanheta*, *vanhetaan*, *vanhetakseen*, *vanheten*, *vanhetessa*, *vanhetkaa*, *vanhetkaamme*, *vanhetko*, *vanhetkoon*, *vanhetkoot*, *vanhettaessa*, *vanhettaisi*, *vanhettaisiin*, *vanhettako*, *vanhettakoon*, *vanhettaman*, *vanhettane*, *vanhettaneen*, *vanhettava*, *vanhettiin*, *vanhettu* — “age” (72);

salaa, *salaama*, *salaamaan*, *salaamaisillaan*, *salaamalla*, *salaaman*, *salaamassa*, *salaamasta*, *salaamaton*, *salaamatta*, *salaaminen*, *salaamista*, *salaamme*, *salaan*, *salaat*, *salaatte*, *salaava*, *salaavat*, *salaisi*, *salaisimme*, *salaisin*, *salaisit*, *salaisitte*, *salaisivat*, *salanne*, *salannee*, *salanneet*, *salannemme*, *salannen*, *salannet*, *salannette*, *salannevat*, *salannut*, *salasi*, *salasimme*, *salasin*, *salasit*, *salasitte*, *salasivat*, *salata*, *salataan*, *salatakseen*, *salaten*, *salatessa*, *salatkaa*, *salatkaamme*, *salatko*, *salatkoon*, *salatkoot*, *salattaessa*, *salattaisi*, *salattaisiin*, *salattako*, *salattakoon*, *salattaman*, *salattane*, *salattaneen*, *salattava*, *salattiin*, *salattu* — “conceal” (73);

katkea, *katkeaa*, *katkeaisi*, *katkeaisimme*, *katkeaisin*, *katkeaisit*, *katkeaisitte*, *katkeaisivat*, *katkeama*, *katkeamaan*, *katkeamaisillaan*, *katkeamalla*, *katkeaman*, *katkeamassa*, *katkeamasta*, *katkeamaton*, *katkeamatta*, *katkeaminen*, *katkeamista*, *katkeamme*, *katkean*, *katkeat*, *katkeatte*, *katkeava*, *katkeavat*, *katkeisi*, *katkeisimme*, *katkeisin*, *katkeisit*, *katkeisitte*, *katkeisivat*, *katkenne*, *katkennee*, *katkenneet*, *katkennemme*, *katkennen*, *katkennet*, *katkennette*, *katkennevat*, *katkennut*, *katkesi*, *katkesimme*, *katkesin*, *katkesit*, *katkesitte*, *katkesivat*, *katketa*, *katketaan*, *katketakseen*, *katketen*, *katketessa*, *katketkaa*, *katketkaamme*, *katketko*, *katketkoon*, *katketkoot*, *katkettaessa*, *katkettaisi*, *katkettaisiin*, *katkettako*, *katkettakoon*, *katkettaman*, *katkettane*, *katkettaneen*, *katkettava*, *katkettiin*, *katkettu* — “cut” (74);

selvinne, *selvinnee*, *selvinneet*, *selvinnemme*, *selvinnen*, *selvinnet*, *selvinnette*, *selvinnevät*, *selvinnyt*, *selvisi*, *selvisimme*, *selvisin*, *selvisit*, *selvisitte*, *selvisivät*, *selviten*, *selvitessä*, *selvitkää*, *selvitkäämme*, *selvitkö*, *selvitköön*, *selvitkööt*, *selvittiin*, *selvitty*, *selvittäessä*, *selvittäisi*, *selvittäisiin*, *selvittäkö*, *selvittäköön*, *selvittämän*, *selvittäne*, *selvittäneen*, *selvittävä*, *selvitä*, *selvitäkseen*, *selvitään*, *selviä*, *selviäisi*, *selviäisimme*, *selviäisin*, *selviäisit*, *selviäisitte*, *selviäisivät*, *selviäminen*, *selviämistä*, *selviämme*, *selviämä*, *selviämäisillään*, *selviämällä*, *selviämän*, *selviämässä*, *selviämästä*, *selviämättä*, *selviämätön*, *selviämään*, *selviän*, *selviät*, *selviätte*, *selviävä*, *selviävät*, *selviää* — “escape” (75);

taida, *taidamme*, *taidan*, *taidat*, *taidatte*, *taideta*, *taidetaan*, *taidettaessa*, *taidettaisi*, *taidettaisiin*, *taidettako*, *taidettakoon*, *taidettaman*, *taidettane*, *taidettaneen*, *taidettava*, *taidettiin*, *taidettu*, *tainne*, *tainnee*, *tainneet*, *tainnemme*, *tainnen*, *tainnet*, *tainnette*, *tainnevat*, *tainnut*, *taisi*, *taisimme*, *taisin*, *taisit*, *taisitte*, *taisivat*, *taitaa*, *taitaakseen*, *taitaen*, *taitaessa*, *taitaisi*, *taitaisimme*, *taitaisin*, *taitaisit*, *taitaisitte*, *taitaisivat*, *taitakaa*, *taitakaamme*, *taitako*, *taitakoon*, *taitakoot*, *taitama*, *taitamaan*, *taitamaisillaan*, *taitamalla*, *taitaman*, *taitamassa*, *taitamasta*, *taitamaton*, *taitamatta*, *taitaminen*, *taitamista*, *taitane*, *taitanee*, *taitaneet*, *taitanemme*, *taitanen*, *taitanet*, *taitanette*, *taitanevat*, *taitanut*, *taitava*, *taitavat* — “master” (76);

kumajaa, *kumajaisi*, *kumaji* — “boom” (77);

kaikaa, *kaikaavat*, *kaikaisi*, *kaikaisivat* — “echo” (78),

we obtain

{*sanomaton*} — (52) 1 word;

{*sano*, *sanoa*, *sanoakseen*, *sanoen*, *sanoessa*, *sanoi*, *sanoimme*, *sanoin*, *sanoisi*, *sanoisimme*, *sanoisin*, *sanoisit*, *sanoisitte*, *sanoisivat*, *sanoit*, *sanoitte*, *sanoivat*, *sanokaa*, *sanokaamme*, *sanoko*, *sanokoon*, *sanokoot*, *sanoma*, *sanomaan*, *sanomaisillaan*, *sanomalla*, *sanoman*, *sanomassa*, *sanomasta*, *sanomatta*, *sanominen*, *sanomista*, *sanomme*, *sanon*, *sanone*, *sanonee*, *sanoneet*, *sanonemme*, *sanonen*, *sanonet*, *sanonette*, *sanonevat*, *sanonut*, *sanoo*, *sanot*, *sanota*, *sanotaan*, *sanottaessa*, *sanottaisi*, *sanottaisiin*, *sanottako*, *sanottakoon*, *sanottaman*, *sanottane*, *sanottaneen*, *sanottava*, *sanotte*, *sanottiin*, *sanottu*, *sanova*, *sanovat*} — (52) 61 words;

{*muistamaton*} — (53) 1 word;

{*muista*, *muistaa*, *muistaakseen*, *muistaen*, *muistaessa*, *muistaisi*, *muistaisimme*, *muistaisin*, *muistaisit*, *muistaisitte*, *muistaisivat*, *muistakaa*, *muistakaamme*, *muistako*, *muistakoon*, *muistakoot*, *muistama*, *muistamaan*, *muistamaisillaan*, *muistamalla*, *muistaman*, *muistamassa*, *muistamasta*, *muistamatta*, *muistaminen*, *muistamista*, *muistamme*, *muistan*, *muistane*, *muistanee*, *muistaneet*, *muistanemme*, *muistanen*, *muistanet*, *muistanette*, *muistanevat*, *muistanut*, *muistat*, *muistatte*, *muistava*, *muistavat*, *muisteta*, *muistetaan*, *muistettaessa*, *muistettaisi*, *muistettaisiin*, *muistettako*, *muistettakoon*, *muistettaman*, *muistettane*, *muistettaneen*, *muistettava*, *muistettiin*, *muistettu*, *muisti*, *muistimme*, *muistin*, *muistit*, *muistitte*, *muistivat*} — (53) 60 words;

{*huutamaton*} — (54) 1 word;

{*huuda*, *huudamme*, *huudan*, *huudat*, *huudatte*, *huudeta*, *huudetaan*, *huudettaessa*, *huudettaisi*, *huudettaisiin*, *huudettako*, *huudettakoon*, *huudettaman*, *huudettane*, *huudettaneen*, *huudettava*, *huudettiin*, *huudettu*, *huusi*, *huusimme*, *huusin*, *huusit*, *huusitte*, *huusivat*, *huutaa*, *huutaakseen*, *huutaen*, *huutaessa*, *huutaisi*, *huutaisimme*, *huutaisin*, *huutaisit*, *huutaisitte*, *huutaisivat*, *huutakaa*, *huutakaamme*, *huutako*, *huutakoon*, *huutakoot*, *huutama*, *huutamaan*, *huutamaisillaan*, *huutamalla*, *huutaman*, *huutamassa*, *huutamasta*, *huutamatta*, *huutaminen*, *huutamista*, *huutane*, *huutanee*, *huutaneet*, *huutanemme*, *huutanen*, *huutanet*, *huutanette*, *huutanevat*, *huutanut*, *huutava*, *huutavat*} — (54) 60 words;

{*sousitte*} — (55) 1 word;

{*soutamaton*} — (55) 1 word;

{*souda*, *soudamme*, *soudan*, *soudat*, *soudatte*, *soudeta*, *soudetaan*, *soudettaessa*, *soudettaisi*, *soudettaisiin*, *soudettako*, *soudettakoon*, *soudettaman*, *soudettane*, *soudettaneen*, *soudettava*, *soudettiin*, *soudettu*, *soudimme*, *soudin*, *soudit*, *souditte*, *sousi*, *sousimme*, *sousin*, *sousit*, *sousivat*, *soutaa*, *soutaakseen*, *soutaen*, *soutaessa*, *soutaisi*, *soutaisimme*, *soutaisin*, *soutaisit*, *soutaisitte*, *soutaisivat*, *soutakaa*, *soutakaamme*, *soutako*, *soutakoon*, *soutakoot*, *soutama*, *soutamaan*, *soutamaisillaan*, *soutamalla*, *soutaman*, *soutamassa*, *soutamasta*, *soutamatta*, *soutaminen*, *soutamista*, *soutane*, *soutanee*, *soutaneet*, *soutanemme*, *soutanen*, *soutanet*, *soutanette*, *soutanevat*, *soutanut*, *soutava*, *soutavat*, *souti*, *soutivat*} — (55) 65 words;

{*kaiva*, *kaivaa*, *kaivaisi*, *kaivaisimme*, *kaivaisin*, *kaivaisit*, *kaivaisitte*, *kaivaisivat*, *kaivakaa*, *kaivakaamme*, *kaivako*, *kaivakoon*, *kaivakoot*, *kaivamme*, *kaivan*, *kaivaneet*, *kaivanut*, *kaivat*, *kaivatte*, *kaivavat*, *kaiveta*, *kaivetaan*, *kaivettaisi*, *kaivettaisiin*, *kaivettako*, *kaivettakoon*, *kaivettiin*, *kaivettu*, *kaivoi*, *kaivoimme*, *kaivoin*, *kaivoit*, *kaivoitte*, *kaivoivat*} — (56) 34 words;

{*saartamaton*} — (57) 1 word;

{*saarra*, *saarramme*, *saarran*, *saarrat*, *saarratte*, *saarreta*, *saarretaan*, *saarrettaessa*, *saarrettaisi*, *saarrettaisiin*, *saarrettako*, *saarrettakoon*, *saarrettaman*, *saarrettane*, *saarrettaneen*, *saarrettava*, *saarrettiin*, *saarrettu*, *saarroimme*, *saarroin*, *saarroit*, *saarroitte*, *saarsi*, *saarsimme*, *saarsin*, *saarsit*, *saarsitte*, *saarsivat*, *saartaa*, *saartaakseen*, *saartaen*, *saartaessa*, *saartaisi*, *saartaisimme*, *saartaisin*, *saartaisit*, *saartaisitte*, *saartaisivat*, *saartakaa*, *saartakaamme*, *saartako*, *saartakoon*, *saartakoot*, *saartama*, *saartamaan*, *saartamaisillaan*, *saartamalla*, *saartaman*, *saartamassa*, *saartamasta*, *saartamatta*, *saartaminen*, *saartamista*, *saartane*, *saartanee*, *saartaneet*, *saartanemme*, *saartanen*, *saartanet*, *saartanette*, *saartanevat*, *saartanut*, *saartava*, *saartavat*, *saartoi*, *saartoivat*} — (57) 66 words;

{*laskematon*} — (58) 1 word;

{*laske*, *laskea*, *laskeakseen*, *laskee*, *laskekaa*, *laskekaamme*, *laskeko*, *laskekoon*, *laskekoot*, *laskema*, *laskemaan*, *laskemaisillaan*, *laskemalla*, *laskeman*, *laskemassa*, *laskemasta*, *laskematta*, *laskeminen*, *laskemista*, *laskemme*, *lasken*, *laskene*, *laskenee*, *laskeneet*, *laskenemme*, *laskenen*, *laskenet*, *laskenette*, *laskenevat*, *laskenut*, *lasket*, *lasketa*, *lasketaan*, *laskettaessa*, *laskettaisi*, *laskettaisiin*, *laskettako*, *laskettakoon*, *laskettaman*, *laskettane*, *laskettaneen*, *laskettava*, *laskette*, *laskettiin*, *laskettu*, *laskeva*, *laskevat*, *laski*, *laskien*, *laskiessa*, *laskimme*, *laskin*, *laskisi*, *laskisimme*, *laskisin*, *laskisit*, *laskisitte*, *laskisivat*, *laskit*, *laskitte*, *laskivat*} — (58) 61 words;

{*tuntematon*} — (59) 1 word;

{*tunne*, *tunnemme*, *tunnen*, *tunnet*, *tunneta*, *tunnetaan*, *tunnettaessa*, *tunnettaisi*, *tunnettaisiin*, *tunnettako*, *tunnettakoon*, *tunnettaman*, *tunnettane*, *tunnettaneen*, *tunnettava*, *tunnette*, *tunnettiin*, *tunnettu*, *tunsi*, *tunsimme*, *tunsin*, *tunsit*, *tunsitte*, *tunsivat*, *tuntea*, *tunteakseen*, *tuntee*, *tuntekaa*, *tuntekaamme*, *tunteko*, *tuntekoon*, *tuntekoot*, *tuntema*, *tuntemaan*, *tuntemaisillaan*, *tuntemalla*, *tunteman*, *tuntemassa*, *tuntemasta*, *tuntematta*, *tunteminen*, *tuntemista*, *tuntene*, *tuntenee*, *tunteneet*, *tuntenemme*, *tuntenen*, *tuntenet*, *tuntenette*, *tuntenevat*, *tuntenut*, *tunteva*, *tuntevat*, *tuntien*, *tuntiessa*, *tuntisi*, *tuntisimme*, *tuntisin*, *tuntisit*, *tuntisitte*, *tuntisivat*} — (59) 61 words;

{*lähtemätön*} — (60) 1 word;

{*lähde*, *lähdemme*, *lähden*, *lähdet*, *lähdette*, *lähdettiin*, *lähdetty*, *lähdettäessä*, *lähdettäisi*, *lähdettäisiin*, *lähdettäkö*, *lähdettäköön*, *lähdettämän*, *lähdettäne*, *lähdettäneen*, *lähdettävä*, *lähdetä*, *lähdetään*, *lähdimme*, *lähdin*, *lähdit*, *lähditte*, *lähtee*, *lähtekää*, *lähtekäämme*, *lähtekö*, *lähteköön*, *lähteköot*, *lähteminen*, *lähtemistä*, *lähtemä*, *lähtemäisillään*, *lähtemällä*, *lähtemän*, *lähtemässä*, *lähtemästä*, *lähtemättä*, *lähtemään*, *lähtene*, *lähtenee*, *lähteneet*, *lähtenemme*, *lähtenen*, *lähtenet*, *lähtenette*, *lähtenevät*, *lähtenyt*, *lähtevä*, *lähtevät*, *lähteä*, *lähteäkseen*, *lähti*, *lähtien*, *lähtiessä*, *lähtisi*, *lähtisimme*, *lähtisin*, *lähtisit*, *lähtisitte*, *lähtisivät*, *lähtivät*} — (60) 61 words;

{*sallimaton*} — (61) 1 word;

{*salli*, *sallia*, *salliakseen*, *sallien*, *salliessa*, *sallii*, *sallikaa*, *sallikaamme*, *salliko*, *sallikoon*, *sallikoot*, *sallima*, *sallimaan*, *sallimaisillaan*, *sallimalla*, *salliman*, *sallimassa*, *sallimasta*, *sallimatta*, *salliminen*, *sallimista*, *sallimme*, *sallin*, *salline*, *sallinee*, *sallineet*, *sallinemme*, *sallinen*, *sallinet*, *sallinette*, *sallinevat*, *sallinut*, *sallisi*, *sallisimme*, *sallisin*, *sallisit*, *sallisitte*, *sallisivat*, *sallit*, *sallita*, *sallitaan*, *sallittaessa*, *sallittaisi*, *sallittaisiin*, *sallittako*, *sallittakoon*, *sallittaman*, *sallittane*, *sallittaneen*, *sallittava*, *sallitte*, *sallittiin*, *sallittu*, *salliva*, *sallivat*} — (61) 55 words;

{*voimaton*} — (62) 1 word;

{*voi*, *voida*, *voidaan*, *voidakseen*, *voiden*, *voidessa*, *voikaa*, *voikaamme*, *voiko*, *voikoon*, *voikoot*, *voima*, *voimaan*, *voimaisillaan*, *voimalla*, *voiman*, *voimassa*, *voimasta*, *voimatta*, *voiminen*, *voimista*, *voimme*, *voin*, *voine*, *voinee*, *voineet*, *voinemme*, *voinen*, *voinet*, *voinette*, *voinevat*, *voinut*, *voisi*, *voisimme*, *voisin*, *voisit*, *voisitte*, *voisivat*, *voit*, *voitaessa*, *voitaisi*, *voitaisiin*, *voitako*, *voitakoon*, *voitaman*, *voitane*, *voitaneen*, *voitava*, *voitiin*, *voitte*, *voitu*, *voiva*, *voivat*} — (62) 53 words;

{*saamaton*} — (63) 1 word;

{*saataisiin*, *saataman*, *saatane*, *saataneen*, *saatiin*, *saatu*} — (63) 6 words;

{*saa*, *saada*, *saadaan*, *saadakseen*, *saaden*, *saadessa*, *saakaa*, *saakaamme*, *saako*, *saakoon*, *saakoot*, *saama*, *saamaan*, *saamaisillaan*, *saamalla*, *saaman*, *saamassa*, *saamasta*, *saamatta*, *saaminen*, *saamista*, *saamme*, *saan*, *saane*, *saanee*, *saaneet*, *saanemme*, *saanen*, *saanet*, *saanette*, *saanevat*, *saanut*, *saat*, *saataessa*, *saataisi*, *saatako*, *saatakoon*, *saatava*, *saatte*, *saava*, *saavat*} — (63) 41 words;

{*juomaton*} — (64) 1 word;

{*juo*, *juoda*, *juodaan*, *juodakseen*, *juoden*, *juodessa*, *juokaa*, *juokaamme*, *juoko*, *juokoon*, *juokoot*, *juoma*, *juomaan*, *juomaisillaan*, *juomalla*, *juoman*, *juomassa*, *juomasta*, *juomatta*, *juominen*, *juomista*, *juomme*, *juon*, *juone*, *juonee*, *juoneet*, *juonemme*, *juonen*, *juonet*, *juonette*, *juonevat*, *juonut*, *juot*, *juotaessa*, *juotaisi*, *juotaisiin*, *juotako*, *juotakoon*, *juotaman*, *juotane*, *juotaneen*, *juotava*, *juotiin*, *juotte*, *juotu*, *juova*, *juovat*} — (64) 47 words;

{*käymätön*} — (65) 1 word;

{*kävi*, *kävimme*, *kävin*, *kävisi*, *kävisimme*, *kävisin*, *kävisit*, *kävisitte*, *kävisivät*, *kävit*, *kävitte*, *kävivät*, *käy*, *käyden*, *käydessä*, *käydä*, *käydäkseen*, *käydään*, *käykää*, *käykäämme*, *käykö*, *käyköön*, *käyköot*, *käyminen*, *käymistä*, *käymme*, *käymä*, *käymäisillään*, *käymällä*, *käymän*, *käymässä*, *käymästä*, *käymättä*, *käymään*, *käyn*, *käyne*, *käynee*, *käyneet*, *käynemme*, *käynen*, *käynet*, *käynette*, *käynevät*, *käynyt*, *käyt*, *käytiin*, *käytte*, *käyty*, *käytäessä*, *käytäisi*, *käytäisiin*, *käytäkö*, *käytäköön*, *käytämän*, *käytäne*, *käytäneen*, *käytävä*, *käyvä*, *käyvät*} — (65) 59 words;

{*rohkaisematon*} — (66) 1 word;

{*rohkaise*, *rohkaisee*, *rohkaisema*, *rohkaisemaan*, *rohkaisemaisillaan*, *rohkaisemalla*, *rohkaiseman*, *rohkaisemassa*, *rohkaisemasta*, *rohkaisematta*, *rohkaiseminen*, *rohkaisemista*, *rohkaisemme*, *rohkaisen*, *rohkaiset*, *rohkaisette*, *rohkaiseva*, *rohkaisevat*, *rohkaisi*, *rohkaisimme*, *rohkaisin*, *rohkaisisi*, *rohkaisisimme*, *rohkaisisin*, *rohkaisisit*, *rohkaisisitte*, *rohkaisisivat*, *rohkaisit*, *rohkaisitte*, *rohkaisivat*, *rohkaiskaa*, *rohkaiskaamme*, *rohkaisko*, *rohkaiskoon*, *rohkaiskoot*, *rohkaisse*, *rohkaissee*, *rohkaisseet*, *rohkaissemme*, *rohkaissen*, *rohkaisset*, *rohkaissette*, *rohkaissevat*, *rohkaissut*, *rohkaista*, *rohkaistaan*, *rohkaistaessa*, *rohkaistaisi*, *rohkaistaisiin*, *rohkaistako*, *rohkaistakoon*, *rohkaistakseen*, *rohkaistaman*, *rohkaistane*, *rohkaistaneen*, *rohkaistava*, *rohkaisten*, *rohkaistessa*, *rohkaistiin*, *rohkaistu*} — (66) 60 words;

{*tulematon*} — (67) 1 word;

{*tule*, *tulee*, *tulema*, *tulemaan*, *tulemaisillaan*, *tulemalla*, *tuleman*, *tulemassa*, *tulemasta*, *tulematta*, *tuleminen*, *tulemista*, *tulemme*, *tulen*, *tulet*, *tulette*, *tuleva*, *tulevat*, *tuli*, *tulimme*, *tulin*, *tulisi*, *tulisimme*, *tulisin*, *tulisit*, *tulisitte*, *tulisivat*, *tulit*, *tulitte*, *tulivat*, *tulkaa*, *tulkaamme*, *tulko*, *tulkoon*, *tulkoot*, *tulla*, *tullaan*, *tullakseen*, *tulle*, *tullee*, *tulleet*, *tullemme*, *tullen*, *tullessa*, *tullet*, *tullette*, *tullevat*, *tullut*, *tultaessa*, *tultaisi*, *tultaisiin*, *tultako*, *tultakoon*, *tultaman*, *tultane*, *tultaneen*, *tultava*, *tultiin*, *tultu*} — (67) 59 words;

{*tupakoimaton*, *tupakoitsematon*} — (68) 2 words;

{*tupakoi*, *tupakoida*, *tupakoidaan*, *tupakoidakseen*, *tupakoiden*, *tupakoidessa*, *tupakoikaa*, *tupakoikaamme*, *tupakoiko*, *tupakoikoon*, *tupakoikoot*, *tupakoima*, *tupakoimaan*, *tupakoimaisillaan*, *tupakoimalla*, *tupakoiman*, *tupakoimassa*, *tupakoimasta*, *tupakoimatta*, *tupakoiminen*, *tupakoimista*, *tupakoimme*, *tupakoin*, *tupakoine*, *tupakoinee*, *tupakoineet*, *tupakoinemme*, *tupakoinen*, *tupakoinet*, *tupakoinette*, *tupakoinevat*, *tupakoinut*, *tupakoisi*, *tupakoisimme*, *tupakoisin*, *tupakoisit*, *tupakoisitte*, *tupakoisivat*, *tupakoit*, *tupakoitaessa*, *tupakoitaisi*, *tupakoitaisiin*, *tupakoitako*, *tupakoitakoon*, *tupakoitaman*, *tupakoitane*, *tupakoitaneen*, *tupakoitava*, *tupakoitiin*, *tupakoitse*, *tupakoitsee*, *tupakoitsema*, *tupakoitsemaan*, *tupakoitsemaisillaan*, *tupakoitsemalla*, *tupakoitseman*, *tupakoitsemassa*, *tupakoitsemasta*, *tupakoitsematta*, *tupakoitseminen*, *tupakoitsemista*, *tupakoitsemme*, *tupakoitsen*, *tupakoitset*, *tupakoitsette*, *tupakoitseva*, *tupakoitsevat*, *tupakoitsi*, *tupakoitsimme*, *tupakoitsin*, *tupakoitsisi*, *tupakoitsisimme*, *tupakoitsisin*, *tupakoitsisit*, *tupakoitsisitte*, *tupakoitsisivat*, *tupakoitsit*, *tupakoitsitte*, *tupakoitsivat*, *tupakoitte*, *tupakoitu*, *tupakoiva*, *tupakoivat*} — (68) 83 words;

{*valitsematon*} — (69) 1 word;

{*valinne*, *valinnee*, *valinneet*, *valinnemme*, *valinnen*, *valinnet*, *valinnette*, *valinnevat*, *valinnut*, *valita*, *valitaan*, *valitakseen*, *valiten*, *valitessa*, *valitkaa*, *valitkaamme*, *valitko*, *valitkoon*, *valitkoot*, *valitse*, *valitsee*, *valitsema*, *valitsemaan*, *valitsemaisillaan*, *valitsemalla*, *valitseman*, *valitsemassa*, *valitsemasta*, *valitsematta*, *valitseminen*, *valitsemista*, *valitsemme*, *valitsen*, *valitset*, *valitsette*, *valitseva*, *valitsevat*, *valitsi*, *valitsimme*, *valitsin*, *valitsisi*, *valitsisimme*, *valitsisin*, *valitsisit*, *valitsisitte*, *valitsisivat*, *valitsit*, *valitsitte*, *valitsivat*, *valittaessa*, *valittaisi*, *valittaisiin*, *valittako*, *valittakoon*, *valittaman*, *valittane*, *valittaneen*, *valittava*, *valittiin*, *valittu*} — (69) 60 words;

{*juoksematon*} — (70) 1 word;

{*juokse*, *juoksee*, *juoksema*, *juoksemaan*, *juoksemaisillaan*, *juoksemalla*, *juokseman*, *juoksemassa*, *juoksemasta*, *juoksematta*, *juokseminen*, *juoksemista*, *juoksemme*, *juoksen*, *juokset*, *juoksette*, *juokseva*, *juoksevat*, *juoksi*, *juoksimme*, *juoksin*, *juoksisi*, *juoksisimme*, *juoksisin*, *juoksisit*, *juoksisitte*, *juoksisivat*, *juoksit*, *juoksitte*, *juoksivat*, *juoskaa*, *juoskaamme*, *juosko*, *juoskoon*, *juoskoot*, *juosse*, *juossee*, *juosseet*, *juossemme*, *juossen*, *juosset*, *juossette*, *juossevat*, *juossut*, *juosta*, *juostaan*, *juostaessa*, *juostaisi*, *juostaisiin*, *juostako*, *juostakoon*, *juostakseen*, *juostaman*, *juostane*, *juostaneen*, *juostava*, *juosten*, *juostessa*, *juostiin*, *juostu*} — (70) 60 words;

{*näkemätön*} — (71) 1 word;

{*näe*, *näemme*, *näen*, *näet*, *näette*, *nähden*, *nähdessä*, *nähdä*, *nähdäkseen*, *nähdään*, *nähkää*, *nähkäämme*, *nähkö*, *nähköön*, *nähkööt*, *nähne*, *nähnee*, *nähneet*, *nähnemme*, *nähnen*, *nähnet*, *nähnette*, *nähnevät*, *nähnyt*, *nähtiin*, *nähty*, *nähtäessä*, *nähtäisi*, *nähtäisiin*, *nähtäkö*, *nähtäköön*, *nähtämän*, *nähtäne*, *nähtäneen*, *nähtävä*, *näimme*, *näin*, *näit*, *näitte*, *näkee*, *näkeminen*, *näkemistä*, *näkemä*, *näkemäisillään*, *näkemällä*, *näkemän*, *näkemässä*, *näkemästä*, *näkemättä*, *näkemään*, *näkevä*, *näkevät*, *näki*, *näkisi*, *näkisimme*, *näkisin*, *näkisit*, *näkisitte*, *näkisivät*, *näkivät*} — (71) 60 words;

{*vanhenematon*} — (72) 1 word;

{*vanhene*, *vanhenee*, *vanhenema*, *vanhenemaan*, *vanhenemaisillaan*, *vanhenemalla*, *vanheneman*, *vanhenemassa*, *vanhenemasta*, *vanhenematta*, *vanheneminen*, *vanhenemista*, *vanhenemme*, *vanhenen*, *vanhenet*, *vanhenette*, *vanheneva*, *vanhenevat*, *vanheni*, *vanhenimme*, *vanhenin*, *vanhenisi*, *vanhenisimme*, *vanhenisin*, *vanhenisit*, *vanhenisitte*, *vanhenisivat*, *vanhenit*, *vanhenitte*, *vanhenivat*, *vanhenne*, *vanhennee*, *vanhenneet*, *vanhennemme*, *vanhennen*, *vanhennet*, *vanhennette*, *vanhennevat*, *vanhennut*, *vanheta*, *vanhetaan*, *vanhetakseen*, *vanheten*, *vanhetessa*, *vanhetkaa*, *vanhetkaamme*, *vanhetko*, *vanhetkoon*, *vanhetkoot*, *vanhettaessa*, *vanhettaisi*, *vanhettaisiin*, *vanhettako*, *vanhettakoon*, *vanhettaman*, *vanhettane*, *vanhettaneen*, *vanhettava*, *vanhettiin*, *vanhettu*} — (72) 60 words;

{*salaamaton*} — (73) 1 word;

{*salaa*, *salaama*, *salaamaan*, *salaamaisillaan*, *salaamalla*, *salaaman*, *salaamassa*, *salaamasta*, *salaamatta*, *salaaminen*, *salaamista*, *salaamme*, *salaan*, *salaat*, *salaatte*, *salaava*, *salaavat*, *salaisi*, *salaisimme*, *salaisin*, *salaisit*, *salaisitte*, *salaisivat*, *salanne*, *salannee*, *salanneet*, *salannemme*, *salannen*, *salannet*, *salannette*, *salannevat*, *salannut*, *salasi*, *salasimme*, *salasin*, *salasit*, *salasitte*, *salasivat*, *salata*, *salataan*, *salatakseen*, *salaten*, *salatessa*, *salatkaa*, *salatkaamme*, *salatko*, *salatkoon*, *salatkoot*, *salattaessa*, *salattaisi*, *salattaisiin*, *salattako*, *salattakoon*, *salattaman*, *salattane*, *salattaneen*, *salattava*, *salattiin*, *salattu*} — (73) 59 words;

{*katkeamaton*} — (74) 1 word;

{*katkea*, *katkeaa*, *katkeaisi*, *katkeaisimme*, *katkeaisin*, *katkeaisit*, *katkeaisitte*, *katkeaisivat*, *katkeama*, *katkeamaan*, *katkeamaisillaan*, *katkeamalla*, *katkeaman*, *katkeamassa*, *katkeamasta*, *katkeamatta*, *katkeaminen*, *katkeamista*, *katkeamme*, *katkean*, *katkeat*, *katkeatte*, *katkeava*, *katkeavat*, *katkeisi*, *katkeisimme*, *katkeisin*, *katkeisit*, *katkeisitte*, *katkeisivat*, *katkenne*, *katkennee*, *katkenneet*, *katkennemme*, *katkennen*, *katkennet*, *katkennette*, *katkennevat*, *katkennut*, *katkesi*, *katkesimme*, *katkesin*, *katkesit*, *katkesitte*, *katkesivat*, *katketa*, *katketaan*, *katketakseen*, *katketen*, *katketessa*, *katketkaa*, *katketkaamme*, *katketko*, *katketkoon*, *katketkoot*, *katkettaessa*, *katkettaisi*, *katkettaisiin*, *katkettako*, *katkettakoon*, *katkettaman*, *katkettane*, *katkettaneen*, *katkettava*, *katkettiin*, *katkettu*} — (74) 66 words;

{*selviämätön*} — (75) 1 word;

{*selvinne*, *selvinnee*, *selvinneet*, *selvinnemme*, *selvinnen*, *selvinnet*, *selvinnette*, *selvinnevät*, *selvinnyt*, *selvisi*, *selvisimme*, *selvisin*, *selvisit*, *selvisitte*, *selvisivät*, *selviten*, *selvitessä*, *selvitkää*, *selvitkäämme*, *selvitkö*, *selvitköön*, *selvitköot*, *selvittiin*, *selvitty*, *selvittäessä*, *selvittäisi*, *selvittäisiin*, *selvittäkö*, *selvittäköön*, *selvittämän*, *selvittäne*, *selvittäneen*, *selvittävä*, *selvitä*, *selvitäkseen*, *selvitään*, *selviä*, *selviäisi*, *selviäisimme*, *selviäisin*, *selviäisit*, *selviäisitte*, *selviäisivät*, *selviäminen*, *selviämistä*, *selviämme*, *selviämä*, *selviämäisillään*, *selviämällä*, *selviämän*, *selviämässä*, *selviämästä*, *selviämättä*, *selviämään*, *selviän*, *selviät*, *selviätte*, *selviävä*, *selviävät*, *selviää*} — (75) 60 words;

{*taitamaton*} — (76) 1 word;

{*taida*, *taidamme*, *taidan*, *taidat*, *taidatte*, *taideta*, *taidetaan*, *taidettaessa*, *taidettaisi*, *taidettaisiin*, *taidettako*, *taidettakoon*, *taidettaman*, *taidettane*, *taidettaneen*, *taidettava*, *taidettiin*, *taidettu*, *tainne*, *tainnee*, *tainneet*, *tainnemme*, *tainnen*, *tainnet*, *tainnette*, *tainnevat*, *tainnut*, *taisi*, *taisimme*, *taisin*, *taisit*, *taisitte*, *taisivat*, *taitaa*, *taitaakseen*, *taitaen*, *taitaessa*, *taitaisi*, *taitaisimme*, *taitaisin*, *taitaisit*, *taitaisitte*, *taitaisivat*, *taitakaa*, *taitakaamme*, *taitako*, *taitakoon*, *taitakoot*, *taitama*, *taitamaan*, *taitamaisillaan*, *taitamalla*, *taitaman*, *taitamassa*, *taitamasta*, *taitamatta*, *taitaminen*, *taitamista*, *taitane*, *taitanee*, *taitaneet*, *taitanemme*, *taitanen*, *taitanet*, *taitanette*, *taitanevat*, *taitanut*, *taitava*, *taitavat*} — (76) 69 words;

{*kumajaa*, *kumajaisi*, *kumaji*} — (77) 3 words;

{*kaikaa*, *kaikaavat*, *kaikaisi*, *kaikaisivat*} — (78) 4 words.

Here, words ending in *-maton* and *mätön* are negative participles, which roughly translate into English words with prefixes *il-*, *im-*, *in-*, *ir-*, *non-*, *un-*, or suffix *-less*. To be consistent with our clustering policy for English words, we have separated negative participles from other inflected forms of Finnish verbs.

8.1.3 Heuristic detection of compounds

The following algorithm for heuristic detection of Finnish compounds differs from the Danish version (Algorithm 5.12) only in some specific details. To make the context clear, we still state the algorithm in full. (In what follows, the string minus operation $\hat{\beta} \ominus \hat{\alpha}$ is prescribed by Definition 5.11.)

Algorithm 8.13 (Heuristic identification of Finnish binary compounds)**.** *Let $\Lambda^{\hat{\rho}} = \{\hat{\rho}_1, \ldots, \hat{\rho}_Q\}$ be a list of distinct Finnish essential roots (without vowel blotting) that contain at least one instance of* **V** *(in either upper or lower case) and DO NOT match the following string patterns:*

$$(\mathbf{CV}ij|\mathbf{VC}|ko|mi|te|tu|(ant|ei|ent|nut|nyt|to|vat|väl|vät)\mathbf{X}).$$

The output of the function $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is obtained through the following procedures:*

(1) *Construct a list $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$ where $\lambda_q^{\hat{\rho}} = \{\hat{\rho}_{(q,1)}, \ldots, \hat{\rho}_{(q,n_q)}\}$ is a subset of $\Lambda^{\hat{\rho}}$ whose members all match the string pattern $\hat{\rho}_q\sim$, for $q \in \mathbb{Z} \cap [1, Q]$.*

(2) *Expand the aforementioned entry $(\hat{\rho}_q, \lambda_q^{\hat{\rho}})$ into a list of triplets $\{(\hat{\rho}_{(q,1)}, \hat{\rho}_q, \hat{\rho}_{(q,1)} \ominus \hat{\rho}_q), \ldots, (\hat{\rho}_{(q,n_q)}, \hat{\rho}_q, \hat{\rho}_{(q,n_q)} \ominus \hat{\rho}_q)\}$ for every $q \in \mathbb{Z} \cap [1, Q]$ such that $\lambda_q^{\hat{\rho}} \neq \varnothing$. Collect all these triplets as one runs through the list $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$. The list of these triplets $\{(\hat{\rho}_{\langle 1\rangle}, \hat{\eta}_{\langle 1\rangle}, \hat{\rho}_{\langle 1\rangle} \ominus \hat{\eta}_{\langle 1\rangle}), \ldots, (\hat{\rho}_{\langle Q'\rangle}, \hat{\eta}_{\langle Q'\rangle}, \hat{\rho}_{\langle Q'\rangle} \ominus \hat{\eta}_{\langle Q'\rangle})\}$ contains potentially valid decompositions of compounds.*

(3) *Screen the aforementioned list of triplets as follows: for every $q' \in \mathbb{Z} \cap [1, Q']$, if $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \hat{\tau}_{\langle q'\rangle} = \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle})$ satisfies*

$$\ell(\hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle}) \geq 2 \quad \textbf{AND} \quad \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle} = \mathbf{X}_1\mathbf{V}\mathbf{X}_2,$$

then construct $\hat{\tau}^*_{\langle q'\rangle}$ by performing $(e|i)ksi\sim\!\!\rightarrow \varnothing$, $(en|En|et|i|ta|ten|tä)\sim\!\!\rightarrow \varnothing$, $(i|I|ie)n\sim\!\!\rightarrow \varnothing$, $e(L|lt|n|S|T)(a|e|ä)\sim\!\!\rightarrow \varnothing$, $i(n|S|T)(a|e|ä)\sim\!\!\rightarrow \varnothing$ on $\hat{\tau}_{\langle q'\rangle}$, before generating a list $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ by members of $\Lambda^{\hat{\rho}}$ that match the pattern $(\underline{\hat{\tau}_{\langle q'\rangle}|\hat{\tau}^*_{\langle q'\rangle}})$; otherwise, set $\lambda^{\hat{\tau}}_{\langle q'\rangle} = \varnothing$.

(4) *Collect all the triplets $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \lambda^{\hat{\tau}}_{\langle q'\rangle})$ where $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ is non-void and $\hat{\tau}_{\langle q'\rangle}$ DOES NOT match the following string patterns:*

$$(Tom|ton|Töm|tön|\tau\nu\tau)\mathbf{X}.$$

Trim the tags in the triplets by doing $\sim^{\mathbf{X}\in}(\mathbf{VC})\mathbf{V}_m \rightarrow \mathbf{X}$ on the strings. This resulting list of triplets $\mathrm{CpdDet}(\Lambda_{\hat{\rho}})$ *contains the heuristic decompositions of all the identified binary compounds.*

Algorithm 8.14 (Approximate clustering of Finnish words with heuristic detection of compounds)**.** *The approximate clustering of a list of Finnish words $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ respecting compounding is completed in four stages:*

(1) *Do as in Algorithm 8.12(1).*

(2) *Do as in Algorithm 8.12(2). Save both the tagged clusters $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$ and the list of word clusters $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ for further use.*

(3) *Construct a tagged list of word clusters $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$ where $\Lambda_k^{\hat{\rho}}$ is the union of all Finnish essential roots (without vowel blotting) available to Γ_k, for $k \in \mathbb{Z} \cap [1, K]$. Set $\Lambda^{\hat{\rho}} = \Lambda_1^{\hat{\rho}} \cup \cdots \cup \Lambda_K^{\hat{\rho}}$, and evaluate* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$.

(4) *The first component $\hat{\rho}_{\langle q''\rangle}$ of each triplet $(\hat{\rho}_{\langle q''\rangle}, \hat{\eta}_{\langle q''\rangle}, \lambda^{\hat{\tau}}_{\langle q''\rangle})$ in* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is called a "dissolvable compound", the second component $\hat{\eta}_{\langle q''\rangle}$ a "heuristic head", and the first member in the third component $\lambda^{\hat{\tau}}_{\langle q''\rangle}$ a "heuristic tail". In the tagged list $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$, every entry containing a "dissolvable compound" is removed, and regrouped with the entries matching its "heuristic head" and "heuristic tail". Finally, remove all tags.*

Example 8.14.1. If we have the inflected compound words from the aforementioned KOTUS-51 sample (containing 27 words), as well as the words *nuori* "young" and *pari* "couple", then our heuristic compound detection generates the following output:

> {*nuoreenpariin*, *nuoreksipariksi*, *nuorellaparilla*, *nuorelleparille*, *nuoreltaparilta*, *nuorenaparina*, *nuorenparin*, *nuoressaparissa*, *nuorestaparista*, *nuoretparit*, *nuorettaparitta*, *nuorienparien*, *nuoriinpareihin*, *nuoriksipareiksi*, *nuorinapareina*, *nuorinepareineen*, *nuorinparein*, *nuoripari*, *nuorissapareissa*, *nuoristapareista*, *nuorittapareitta*, *nuortaparia*, *nuortenparien*, *pari*} — 24 words;

> {*nuoreenpariin*, *nuoreksipariksi*, *nuorellaparilla*, *nuorelleparille*, *nuoreltaparilta*, *nuorenaparina*, *nuorenparin*, *nuoressaparissa*, *nuorestaparista*, *nuoretparit*, *nuorettaparitta*, *nuori*, *nuoriapareja*, *nuorienparien*, *nuoriinpareihin*, *nuoriksipareiksi*, *nuorillapareilla*, *nuorillepareille*, *nuoriltapareilta*, *nuorinapareina*, *nuorinepareineen*, *nuorinparein*, *nuoripari*, *nuorissapareissa*, *nuoristapareista*, *nuorittapareitta*, *nuortaparia*, *nuortenparien*} — 28 words.

This is partially satisfactory: a small fraction of the compounds were not decomposed properly (component *pari* undetected), but were simply regarded as derivatives of the word *nuori*. If we have *nuorin* "youngest" instead of *nuori* "young" in our input, then we get something slightly worse:

{*nuoriapareja*} — 1 word;

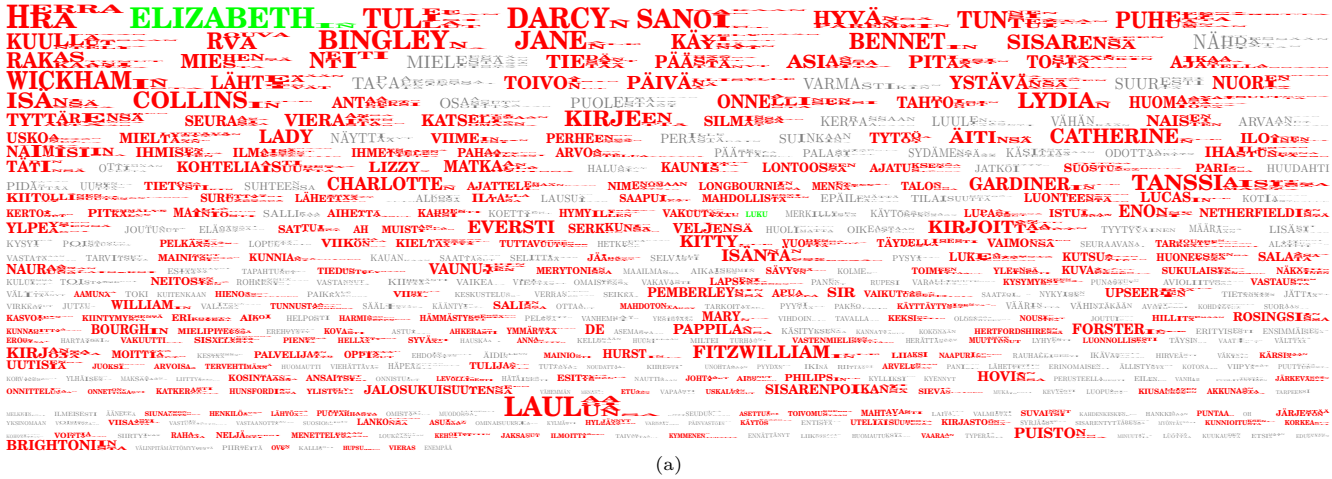
{*nuoreenpariin*, *nuoreksipariksi*, *nuorellaparilla*, *nuorelleparille*, *nuoreltaparilta*, *nuorenaparina*, *nuorenparin*, *nuoressaparissa*, *nuorestaparista*, *nuoretparit*, *nuorettaparitta*, *nuorienparien*, *nuoriinpareihin*, *nuoriksipareiksi*, *nuorinepareineen*, *nuorinparein*, *nuoripari*, *nuorissapareissa*, *nuoristapareista*, *nuorittapareitta*, *nuortaparia*, *nuortenparien*, *pari*} — 23 words;

{*nuoreenpariin*, *nuoreksipariksi*, *nuorellaparilla*, *nuorelleparille*, *nuoreltaparilta*, *nuorenaparina*, *nuorenparin*, *nuoressaparissa*, *nuorestaparista*, *nuoretparit*, *nuorettaparitta*, *nuorienparien*, *nuoriinpareihin*, *nuoriksipareiksi*, *nuorillapareilla*, *nuorillepareille*, *nuoriltapareilta*, *nuorin*, *nuorinapareina*, *nuorinepareineen*, *nuorinparein*, *nuoripari*, *nuorissapareissa*, *nuoristapareista*, *nuorittapareitta*, *nuortaparia*, *nuortenparien*} — 27 words.

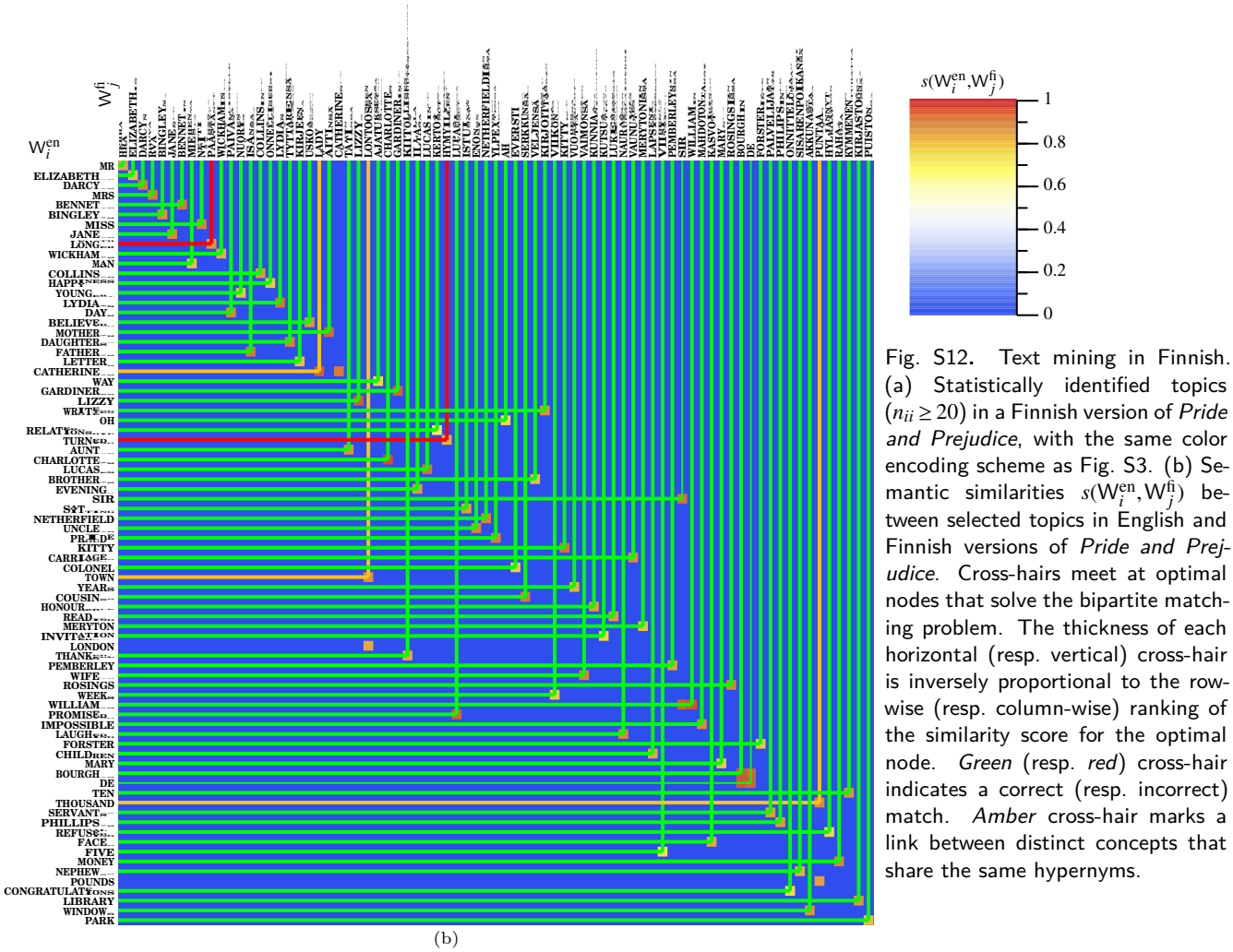
Example 8.14.2. In Fig. S12, we further test the aforementioned word clustering algorithm (with heuristic detection of compounds) on topic extraction and machine translation (see Table S1 for text sources).

Peni(n)kulma is an archaic measure of length used in Finland. Although this is not numerically equivalent to the English *mile*, we consider these length units exact matches in Fig. S12b′ and b′′.

In Darwin's *Origin of Species*, the word "fresh" is generally used in the context of *fresh water* and the word "reciprocal" is mostly used in the phrase *reciprocal cross*. Therefore, we consider *suolattoman* (genitive singular of *suolaton* "saltless") an exact match to *fresh* and *vastavuoroinen* "reciprocal" a close match to *graft* in Fig. S12b′′. Similarly, we regard *siitoselimistö* "breeding organ" as a near synonym to *reciprocal*.



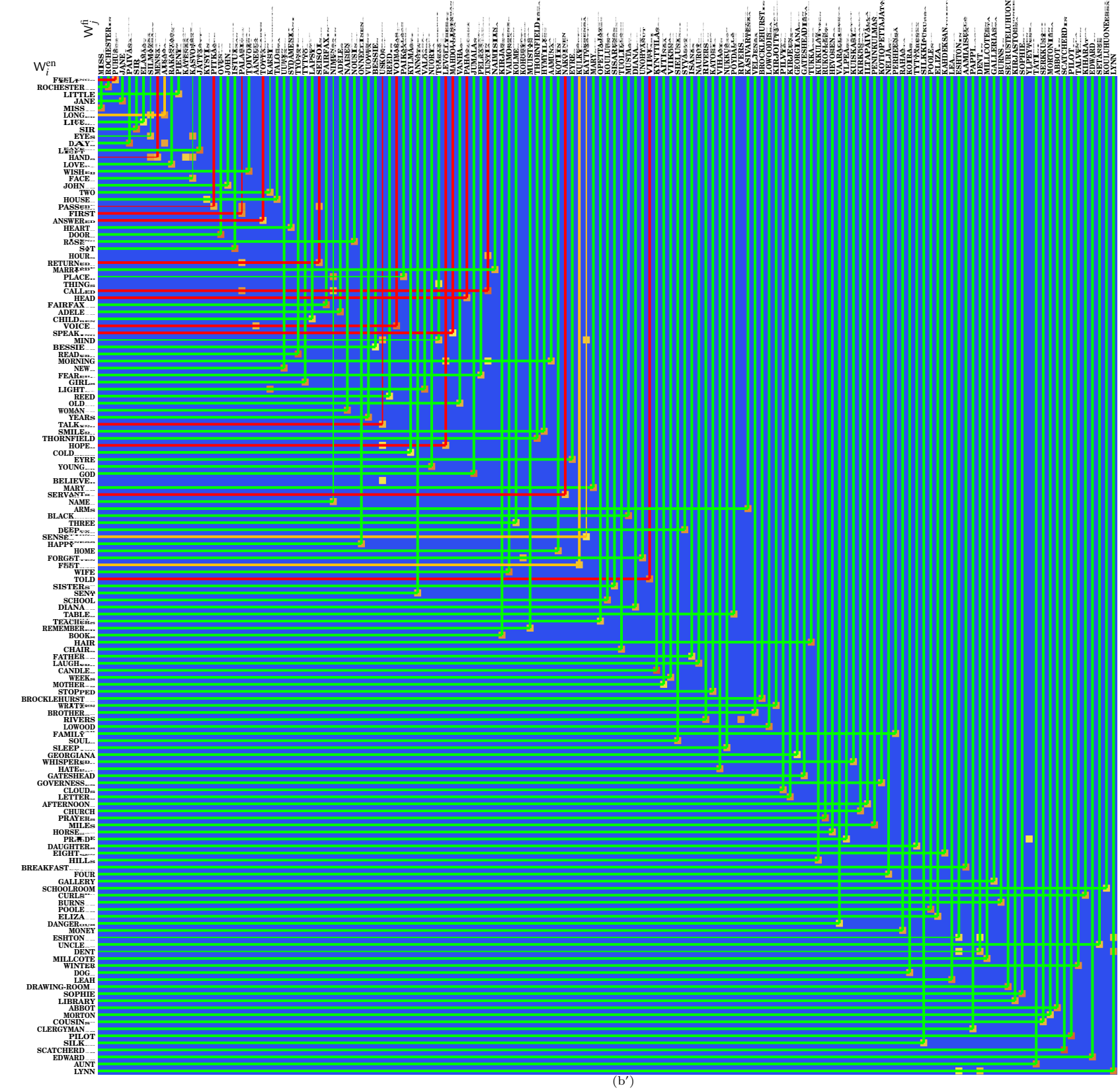
(a)



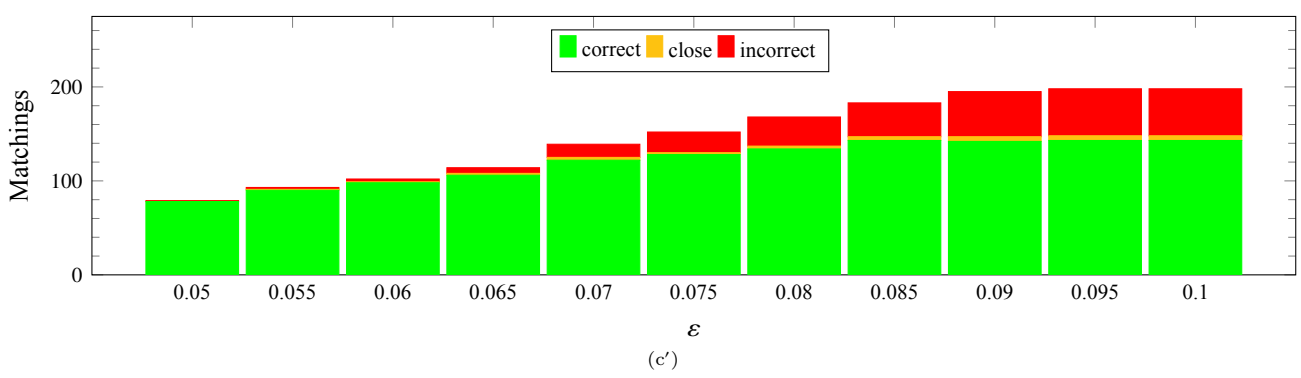
(b)

Fig. S12. Text mining in Finnish. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Finnish version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{fi})$ between selected topics in English and Finnish versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.

(a′)



(b′)



(c′)

Fig. S12. Text mining in Finnish. (*Continued*) (a′) Statistically identified topics ($n_{ii} \geq 20$) in a Finnish version of *Jane Eyre*, with the same color encoding scheme as Fig. S3. (b′) Semantic similarities $s(W_i^{en}, W_j^{fi})$ between selected topics in English and Finnish versions of *Jane Eyre*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c′) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

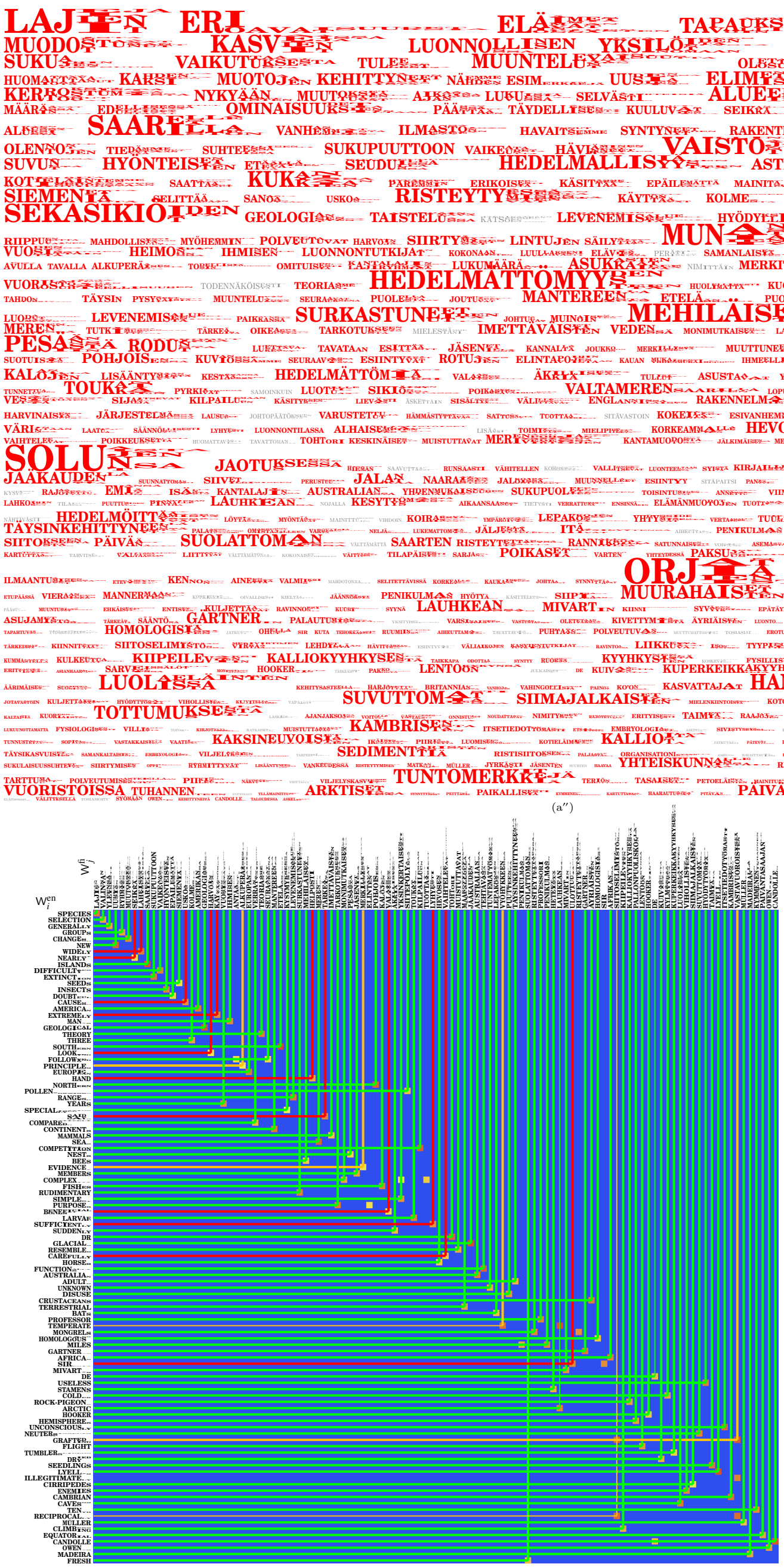


Fig. S12. Text mining in Finnish. (*Continued*) (a″) Statistically identified topics ($n_{ii} \geq 20$) in a Finnish version of *Origin of Species*, with the same color encoding scheme as Fig. S3. (b″) Semantic similarities $s(W_i^{en}, W_j^{fi})$ between selected topics in English and Finnish versions of *Origin of Species*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c″) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

8.2 Modified Porter stemming algorithm for Hungarian

In our presentation below, we will disable hyphenation for all the Hungarian words. There are two reasons for doing so: (1) The hyphenation rules for Hungarian are both complex and counterintuitive, which may baffle non-Hungarian readers of this document. (2) The Hungarian language option in the LaTeX `babel` package is incompatible with many other LaTeX packages, which discourages us from loading it.

Definition 8.15 (Hungarian stop words)**.** If a word, up to removal of *~-e* (hyphen and the letter *e*, attached at the end[100]), belongs to the following list[101]:

a, *abban*, *ahhoz*, *ahogy*, *ahogyan*, *ahol*, *ahonnan*, *ahonnét*, *ahova*, *akár*, *akárhol*, *akárhonnan*, *akárhonnét*, *akárhova*, *aki*, *akik*, *akkor*, *aközben*, *alá*, *alád*, *alájuk*, *alám*, *alánk*, *alátok*, *alatt*, *alatta*, *alattad*, *alattam*, *alattatok*, *alattuk*, *alattunk*, *alig*, *aligha*, *alighanem*, *alighogy*, *alól*, *alóla*, *alólad*, *alólam*, *alólatok*, *alóluk*, *alólunk*, *által*, *általa*, *általában*, *általad*, *általam*, *általatok*, *általuk*, *általunk*, *alul*, *amely*, *amelyek*, *amelyekben*, *amelyeket*, *amelyet*, *amelynek*, *ami*, *amíg*, *amikor*, *amint*, *amit*, *amolyan*, *annak*, *annyi*, *annyiba*, *annyiban*, *annyiból*, *annyiért*, *annyihoz*, *annyiig*, *annyiként*, *annyin*, *annyinak*, *annyinál*, *annyira*, *annyiról*, *annyiszor*, *annyit*, *annyitól*, *annyivá*, *annyival*, *arra*, *arról*, *át*, *az*, *azaz*, *azért*, *azok*, *azon*, *azonban*, *azt*, *aztán*, *azután*, *azzal*, *azzonal*, *bár*, *bárcsak*, *bárhogyan*, *bárhol*, *bárhonnan*, *bárhonnét*, *bárhova*, *bármikor*, *be*, *belé*, *beléd*, *beléjük*, *belém*, *belénk*, *belétek*, *beljebb*, *belőle*, *belőled*, *belőlem*, *belőletek*, *belőlük*, *belőlünk*, *belül*, *benn*, *benne*, *benned*, *bennem*, *bennetek*, *benneteket*, *bennük*, *bennünk*, *bennünket*, *bent*, *bizony*, *cikk*, *cikkek*, *cikkeket*, *csak*, *csaknem*, *csinál*, *csinálandó*, *csináld*, *csinálhat*, *csinálj*, *csinálja*, *csináljad*, *csináljak*, *csinálják*, *csináljál*, *csináljalak*, *csináljam*, *csináljanak*, *csináljatok*, *csináljátok*, *csináljon*, *csináljuk*, *csináljunk*, *csinállak*, *csinálna*, *csinálná*, *csinálnád*, *csinálnak*, *csinálnák*, *csinálnál*, *csinálnálak*, *csinálnám*, *csinálnának*, *csinálnánk*, *csinálnátok*, *csinálnék*, *csinálni*, *csinálnia*, *csinálniuk*, *csinálnod*, *csinálnom*, *csinálnotok*, *csinálnunk*, *csináló*, *csinálod*, *csinálok*, *csinálom*, *csinálsz*, *csinált*, *csinálta*, *csináltad*, *csináltak*, *csinálták*, *csináltál*, *csináltalak*, *csináltam*, *csináltatok*, *csináltátok*, *csináltok*, *csináltuk*, *csináltunk*, *csinálunk*, *csinálva*, *de*, *dehát*, *dehogy*, *e*, *ebben*, *eddig*, *egész*, *egy*, *egybe*, *egyben*, *egyből*, *egye*, *egyéb*, *egyébként*, *egyed*, *egyem*, *egyen*, *egyért*, *egyes*, *egyet*, *egyetek*, *egyetlen*, *eggyé*, *eggyel*, *egyhez*, *egyig*, *egyik*, *egyikbe*, *egyikben*, *egyikből*, *egyike*, *egyiked*, *egyikek*, *egyikekbe*, *egyikekben*, *egyikekből*, *egyikeken*, *egyikekért*, *egyikeket*, *egyikekhez*, *egyikekig*, *egyikekké*, *egyikekkel*, *egyikekként*, *egyikeknek*, *egyikeknél*, *egyikekre*, *egyikekről*, *egyikektől*, *egyikem*, *egyiken*, *egyikért*, *egyiket*, *egyiketek*, *egyikhez*, *egyikig*, *egyikké*, *egyikkel*, *egyikként*, *egyiknek*, *egyiknél*, *egyikre*, *egyikről*, *egyiktől*, *egyiktük*, *egyikünk*, *egyként*, *egymás*, *egymásba*, *egymásban*, *egymásból*, *egymásért*, *egymáshoz*, *egymásig*, *egymásként*, *egymásnak*, *egymásnál*, *egymáson*, *egymásra*, *egymásról*, *egymássá*, *egymással*, *egymást*, *egymástól*, *egynek*, *egynél*, *egyre*, *egyről*, *egytől*, *együk*, *együnk*, *együtt*, *ehhez*, *ekkor*, *eközben*, *el*, *elé*, *eléd*, *elég*, *eléjük*, *elém*, *elénk*, *elétek*, *ellen*, *ellenben*, *ellene*, *ellened*, *ellenem*, *ellenére*, *ellenetek*, *ellenük*, *ellenünk*, *elő*, *elől*, *előle*, *előled*, *előlem*, *előletek*, *előlük*, *előlünk*, *először*, *előtt*, *előtte*, *előtted*, *előttem*, *előttetek*, *előttük*, *előttünk*, *első*, *emilyen*, *én*, *énbelőlem*, *engem*, *ennek*, *éppen*, *erre*, *érte*, *érted*, *értem*, *értetek*, *értük*, *értünk*, *és*, *esetében*, *esetedben*, *esetemben*, *esetén*, *esetetekben*, *esetleg*, *esetükben*, *esetünkben*, *ez*, *ezek*, *ezen*, *ezért*, *ezt*, *ezzel*, *fel*, *fele*, *felé*, *feléd*, *feléjük*, *felém*, *felénk*, *felétek*, *felett*, *felől*, *felőle*, *felőled*, *felőlem*, *felőletek*, *felőlük*, *felőlünk*, *felül*, *fenn*, *fent*, *fenti*, *fentibe*, *fentiben*, *fentiből*, *fentiek*, *fentiekbe*, *fentiekben*, *fentiekből*, *fentieken*, *fentiekért*, *fentieket*, *fentiekhez*, *fentiekig*, *fentiekké*, *fentiekkel*, *fentiekként*, *fentieknek*, *fentieknél*, *fentiekre*, *fentiekről*, *fentiektől*, *fentiért*, *fentihez*, *fentiig*, *fentiként*, *fentin*, *fentinek*, *fentinél*, *fentire*, *fentiről*, *fentit*, *fentitől*, *fentivé*, *fentivel*, *fog*, *fogja*, *fogják*, *fogjátok*, *fogjuk*, *fognak*, *fogod*, *fogok*, *fogom*, *fogsz*, *fogtok*, *fogunk*, *fogva*, *folyamán*, *folytán*, *föl*, *fölé*, *föléd*, *föléjük*, *fölém*, *fölénk*, *fölétek*, *följebb*, *fölött*, *fölötte*, *fölötted*, *fölöttem*, *fölöttetek*, *fölöttük*, *fölöttünk*, *fölül*, *fölüle*, *fölüled*, *fölülem*, *fölületek*, *fölülük*, *fölülünk*, *gyanánt*, *ha*, *hadd*, *hanem*, *hány*, *hasonlóan*, *hát*, *hátha*, *helyett*, *helyette*, *helyetted*, *helyettem*, *helyettetek*, *helyettük*, *helyettünk*, *hiszen*, *hogy*, *hogyan*, *hogyha*, *hogyne*, *hol*, *holott*, *honnan*, *honnét*, *hosszat*, *hova*, *hozzá*, *hozzád*, *hozzájuk*, *hozzám*, *hozzánk*, *hozzátok*, *ide*, *időközben*, *igen*, *így*, *ill*, *ill.*, *illetve*, *ily*, *ilyen*, *ilyenkor*, *innen*, *iránt*, *iránta*, *irántad*, *irántam*, *iránttatok*, *irántuk*, *iránttunk*, *is*, *ismét*, *ison*, *itt*, *jóvoltából*, *kedvéért*, *kegyed*, *kegyedbe*, *kegyedben*, *kegyedből*, *kegyeddé*, *kegyeddel*, *kegyeden*, *kegyedért*, *kegyedet*, *kegyedhez*, *kegyedig*, *kegyedként*, *kegyednek*, *kegyednél*, *kegyedre*, *kegyedről*, *kegyedtől*, *kell*, *kellett*, *képest*, *keressünk*, *keresztül*, *kezdve*, *ki*, *kijjebb*, *kinn*, *kint*, *kivéve*, *kívül*, *köré*, *köréd*, *köréjük*, *körém*, *körénk*, *körétek*, *körül*, *körülötte*, *körülötted*, *körülöttem*, *körülöttetek*, *körülöttük*, *körülöttünk*, *következtében*, *közben*, *közbeni*, *közé*, *közéd*, *közéjük*, *közém*, *közénk*, *közétek*, *között*, *közötte*, *közötted*, *közöttem*, *közöttetek*, *közöttük*, *közöttünk*, *közt*, *közte*, *közted*, *köztem*, *köztetek*, *köztük*, *köztünk*, *közül*, *közüle*, *közüled*, *közülem*, *közületek*, *közülük*, *közülünk*, *különben*, *le*, *leendő*, *legalább*, *légy*, *legyek*, *legyél*, *legyen*, *legyenek*, *legyetek*, *legyünk*, *lehet*, *lehetett*, *lejjebb*, *lenn*, *lenne*, *lenned*, *lennék*, *lennél*, *lennem*, *lennének*, *lennénk*, *lennetek*,

[100] Word final *~-e* occurs in Hungarian tag questions.

[101] Our list of Hungarian stop words is based on `http://snowball.tartarus.org/algorithms/hungarian/stop.txt` (prepared by Anna Tordai), with extensive additions to roughly match their counterparts in English. Although the words *jó* "good", *jobban* "better", *jól* "well", *nagy* "big", *nagyobb* "bigger", *új* "new", *újabb* "newer" showed up in Anna Tordai's list of Hungarian stop words, they do not appear in our list here, for the sake of consistency with other languages.

lennétek, *lenni*, *lennie*, *lenniük*, *lennünk*, *lent*, *lesz*, *leszek*, *leszel*, *lesznek*, *lesztek*, *leszünk*, *lett*, *lettek*, *lettél*, *lettem*, *lettetek*, *lettünk*, *léve*, *lévén*, *levő*, *lévő*, *maga*, *magát*, *maguk*, *magukat*, *majd*, *majdhogynem*, *majdnem*, *már*, *máris*, *más*, *mása*, *másai*, *másaid*, *másaik*, *másaim*, *másaink*, *másaitok*, *másba*, *másban*, *másból*, *másért*, *másfelől*, *máshol*, *máshonnan*, *máshonnét*, *máshova*, *máshoz*, *másig*, *másik*, *másként*, *másnak*, *másnál*, *másod*, *mások*, *másokat*, *másokba*, *másokban*, *másokból*, *másokért*, *másokhoz*, *másokig*, *másokká*, *másokkal*, *másokként*, *másoknak*, *másoknál*, *másokon*, *másokra*, *másokról*, *másoktól*, *másom*, *máson*, *másotok*, *másra*, *másról*, *mássá*, *mással*, *mást*, *mástól*, *másuk*, *másunk*, *meg*, *még*, *megint*, *mégis*, *mellé*, *melléd*, *melléjük*, *mellém*, *mellénk*, *mellétek*, *mellett*, *mellette*, *melletted*, *mellettem*, *mellettetek*, *mellettük*, *mellettünk*, *mellől*, *mellőle*, *mellőled*, *mellőlem*, *mellőletek*, *mellőlük*, *mellőlünk*, *mely*, *melyek*, *mentén*, *mert*, *merthogy*, *mi*, *miatt*, *miatta*, *miattad*, *miattam*, *miattatok*, *miattuk*, *miattunk*, *mibelőlünk*, *micsoda*, *mielőtt*, *miért*, *míg*, *mihelyt*, *mikor*, *miközben*, *milyen*, *mind*, *mindbe*, *mindben*, *mindből*, *minddé*, *minddel*, *mindegyik*, *mindegyikbe*, *mindegyikben*, *mindegyikből*, *mindegyike*, *mindegyiked*, *mindegyikek*, *mindegyikekbe*, *mindegyikekben*, *mindegyikekből*, *mindegyikeken*, *mindegyikekért*, *mindegyikeket*, *mindegyikekhez*, *mindegyikekig*, *mindegyikekké*, *mindegyikekkel*, *mindegyikekként*, *mindegyikeknek*, *mindegyikeknél*, *mindegyikekre*, *mindegyikekről*, *mindegyikektől*, *mindegyikem*, *mindegyiken*, *mindegyikért*, *mindegyiket*, *mindegyiketek*, *mindegyikhez*, *mindegyikig*, *mindegyikké*, *mindegyikkel*, *mindegyikként*, *mindegyiknek*, *mindegyiknél*, *mindegyikre*, *mindegyikről*, *mindegyiktől*, *mindegyikük*, *mindegyikünk*, *minden*, *mindenhol*, *mindenhonnan*, *mindenhonnét*, *mindenhova*, *mindenki*, *mindent*, *mindért*, *mindet*, *mindhez*, *mindig*, *mindként*, *mindnek*, *mindnél*, *mindre*, *mindről*, *mindtől*, *minket*, *mint*, *mintha*, *mit*, *miután*, *mivel*, *most*, *mögé*, *mögéd*, *mögéjük*, *mögém*, *mögénk*, *mögétek*, *mögött*, *mögötte*, *mögötted*, *mögöttem*, *mögöttetek*, *mögöttük*, *mögöttünk*, *mögül*, *mögüle*, *mögüled*, *mögülem*, *mögületek*, *mögülük*, *mögülünk*, *múlva*, *nagyon*, *nahát*, *nála*, *nálad*, *nálam*, *nálatok*, *náluk*, *nálunk*, *ne*, *néha*, *néhány*, *néhányat*, *néhányba*, *néhányban*, *néhányból*, *néhányért*, *néhányhoz*, *néhányig*, *néhányként*, *néhánynak*, *néhánynál*, *néhánnyá*, *néhánnyal*, *néhányon*, *néhányra*, *néhányról*, *néhánytól*, *néhol*, *neked*, *nekem*, *neki*, *nekik*, *nektek*, *nekünk*, *nélkül*, *nélküle*, *nélküled*, *nélkülem*, *nélkületek*, *nélkülük*, *nélkülünk*, *nem*, *nemcsak*, *némely*, *némelybe*, *némelyben*, *némelyből*, *némelyek*, *némelyekbe*, *némelyekben*, *némelyekből*, *némelyeken*, *némelyekért*, *némelyeket*, *némelyekhez*, *némelyekig*, *némelyekké*, *némelyekkel*, *némelyekként*, *némelyeknek*, *némelyeknél*, *némelyekre*, *némelyekről*, *némelyektől*, *némelyen*, *némelyért*, *némelyet*, *némelyhez*, *némelyig*, *némelyik*, *némelyikbe*, *némelyikben*, *némelyikből*, *némelyiken*, *némelyikért*, *némelyiket*, *némelyikhez*, *némelyikig*, *némelyikké*, *némelyikkel*, *némelyikként*, *némelyiknek*, *némelyiknél*, *némelyikre*, *némelyikről*, *némelyiktől*, *némelyként*, *némellyé*, *némellyel*, *némelynek*, *némelynél*, *némelyre*, *némelyről*, *némelytől*, *némi*, *némiképp*, *némiképpen*, *némileg*, *nézve*, *nincs*, *nincsen*, *nincsenek*, *noha*, *oda*, *oly*, *olyan*, *olyat*, *olykor*, *óta*, *ott*, *ő*, *őbelőle*, *őbelőlük*, *ők*, *őket*, *ön*, *önbe*, *önben*, *önből*, *önért*, *önhöz*, *önig*, *önként*, *önné*, *önnek*, *önnel*, *önnél*, *önök*, *önökbe*, *önökben*, *önökből*, *önökért*, *önöket*, *önökhöz*, *önökig*, *önökké*, *önökkel*, *önökként*, *önöknek*, *önöknél*, *önökön*, *önökre*, *önökről*, *önöktől*, *önön*, *önre*, *önről*, *önt*, *öntől*, *össze*, *összes*, *összesen*, *őt*, *pedig*, *persze*, *rá*, *rád*, *rajta*, *rajtad*, *rajtam*, *rajtatok*, *rajtuk*, *rajtunk*, *rájuk*, *rám*, *ránk*, *rátok*, *részedre*, *részedről*, *részemre*, *részemről*, *részére*, *részéről*, *részetekre*, *részetekről*, *részükre*, *részükről*, *részünkre*, *részünkről*, *réveden*, *révemen*, *révén*, *réveteken*, *révükön*, *révünkön*, *róla*, *rólad*, *rólam*, *rólatok*, *róluk*, *rólunk*, *rögtön*, *s*, *saját*, *se*, *segítségeddel*, *segítségemmel*, *segítségetekkel*, *segítségével*, *segítségükkel*, *segítségünkkel*, *sehol*, *sehonnan*, *sehonnét*, *sehova*, *sem*, *semmi*, *sincs*, *sincsen*, *sincsenek*, *soha*, *sohasem*, *sok*, *sokat*, *sokféle*, *sokféleképp*, *sokféleképpen*, *sokkal*, *során*, *sőt*, *számára*, *számodra*, *számomra*, *számotokra*, *számukra*, *számunkra*, *szemben*, *szerint*, *szerinte*, *szerinted*, *szerintem*, *szerintetek*, *szerintük*, *szerintünk*, *szinte*, *táján*, *tájban*, *tájt*, *talán*, *te*, *tebelőled*, *tedd*, *teendő*, *téged*, *tegye*, *tegyed*, *tegyek*, *tegyék*, *tegyél*, *tegyelek*, *tegyem*, *tegyen*, *tegyenek*, *tegyetek*, *tegyétek*, *tegyük*, *tegyünk*, *tehát*, *tehet*, *teljes*, *tenne*, *tenné*, *tenned*, *tennéd*, *tennék*, *tennél*, *tennélek*, *tennem*, *tenném*, *tennének*, *tennénk*, *tennetek*, *tennétek*, *tenni*, *tennie*, *tenniük*, *tennünk*, *tesz*, *teszed*, *teszek*, *teszel*, *teszem*, *teszi*, *teszik*, *teszitek*, *teszlek*, *tesznek*, *tesszük*, *tesztek*, *teszünk*, *tett*, *tette*, *tetted*, *tettek*, *tették*, *tettél*, *tettelek*, *tettem*, *tettetek*, *tettétek*, *tettük*, *tettünk*, *tevő*, *ti*, *tibelőletek*, *titeket*, *tovább*, *továbbá*, *több*, *tőle*, *tőled*, *tőlem*, *tőletek*, *tőlük*, *tőlünk*, *túl*, *úgy*, *ugyan*, *ugyanis*, *ugye*, *úgyis*, *újból*, *újra*, *után*, *utána*, *utánad*, *utánam*, *utánatok*, *utánuk*, *utánunk*, *útján*, *utolsó*, *vagy*, *vagyis*, *vagyok*, *vagytok*, *vagyunk*, *valaha*, *valahányszor*, *valahogyan*, *valahol*, *valahonnan*, *valahonnét*, *valahova*, *valaki*, *valamerre*, *valami*, *valamihez*, *valamiképpen*, *valamikor*, *valamin*, *valaminek*, *valaminél*, *valamint*, *valamire*, *valamit*, *valamitől*, *valamivel*, *való*, *van*, *vannak*, *végett*, *vele*, *veled*, *velem*, *veletek*, *velük*, *velünk*, *viszont*, *vissza*, *volna*, *volnál*, *volnának*, *volnánk*, *volnátok*, *volnék*, *volt*, *voltak*, *voltál*, *voltam*, *voltatok*, *voltunk*,

or matches one of the following string patterns:

(∅|*jó*|*ön*)*mag*(*a*|*á*|*u*(∅|*n*)*k*)(*d*|*m*|*tok*)(∅|*at*|*ba*|*ban*|*ból*|*dá*|*dal*|*ért*|*hoz*|*ig*|*ká*|*kal*|*ként*|*má*|*mal*|*n*|*nak*|*nál*|*on*|*ra*|*ról*|*t*|*tól*|*va*|*vál*),

(*af*|*ef*|*mi*|*ki*|*semmi*)*fél*(*e*|*é*)(∅|*k*)(∅|*be*|*ben*|*ből*|*en*|*ént*|*ért*|*et*|*hez*|*ig*|*ké*|*kel*|*ként*|*n*|*nek*|*nél*|*re*|*ről*|*t*|*től*|*vé*|*vel*),

(∅|*a*|*akár*|*bár*|*más*|*minden*|*mint*|*ne*|*né*|*se*|*vala*)*ho*(*gy*|*l*|*va*|*vá*),

(∅|*mind*|*ugyan*)*e*(*bbe*|*bben*|*bből*|*ddig*|*hhez*|*kként*|*nnek*|*nnél*|*rre*|*rről*|*ttől*|*z*|*zek*|*zekbe*|*zekben*|*zekből*|*zeken*|*zekért*|*zeket*|*zekhez*|*zekig*|*zekké*|*zekkel*|*zekként*|*zeknek*|*zeknél*|*zekre*|*zekről*|*zektől*|*zen*|*zért*|*zt*|*zzé*|*zzel*),

(∅|*mind*|*ugyan*)*a*(*bba*|*bban*|*bból*|*ddig*|*hhoz*|*kként*|*nnak*|*nnál*|*rra*|*rról*|*ttól*|*z*|*zért*|*zok*|*zokat*|*zokba*|*zokban*|*zokból*|*zokért*|*zokhoz*|*zokig*|*zokká*|*zokkal*|*zokként*|*zoknak*|*zoknál*|*zokon*|*zokra*|*zokról*|*zoktól*|*zon*|*zt*|*zzá*|*zzal*),

(∅|*akárm*|*am*|*bárm*|*m*|*né*|*sem*|*ugyan*|*valam*)(*annyi*|*elyik*|*ennyi*|*ilyen*)(∅|*e*)(∅|*d*|*je*|*jei*|*jeid*|*jeik*|*jeim*|*jeink*|*jeitek*|*jük*|*m*|*nk*|*tek*)(∅|*e*)(∅|*be*|*ben*|*ből*|*dé*|*del*|*ért*|*hez*|*ig*|*k*|*kbe*|*kben*|*kből*|*ké*|*kel*|*ken*|*ként*|*kért*|*ket*|*khez*|*kig*|*kké*|*kkel*|*kként*|*knek*|*knél*|*kre*|*kről*|*ktől*|*mé*|*mel*|*n*|*nek*|*nél*|*re*|*ről*|*t*|*től*|*vé*|*vel*),

(∅|*a*|*akár*|*bár*|*minden*|*sem*|*sen*|*vala*)(*k*|*m*)*i*(∅|*d*|*é*|*je*|*jei*|*jeid*|*jeik*|*jeim*|*jeink*|*jeitek*|*jük*|*m*|*nk*|*tek*)(∅|*e*)(∅|*be*|*ben*|*ből*|*dé*|*del*|*ért*|*hez*|*ig*|*k*|*kbe*|*kben*|*kből*|*ké*|*kel*|*ken*|*ként*|*kért*|*ket*|*khez*|*kig*|*kké*|*kkel*|*kként*|*knek*|*knél*|*kre*|*kről*|*ktől*|*mé*|*mel*|*n*|*nek*|*nél*|*re*|*ről*|*t*|*től*|*vé*|*vel*),

(*enyéim*|*enyém*|*magáé*|*magáéi*|*maguké*|*magukéi*|*mieink*|*mienk*|*miénk*|*öné*|*önéi*|*önöké*|*önökéi*|*övé*|*övéi*|*övéik*|*övék*|*tied*|*tiéd*|*tieid*|*tieitek*|*tietek*|*tiétek*)(∅|*be*|*ben*|*ből*|*dé*|*del*|*ek*|*ekbe*|*ekben*|*ekből*|*eken*|*ekért*|*eket*|*ekhez*|*ekig*|*ekké*|*ekkel*|*ekként*|*eknek*|*eknél*|*ekre*|*ekről*|*ektől*|*en*|*ért*|*et*|*hez*|*ig*|*ké*|*kel*|*ként*|*mé*|*mel*|*nek*|*nél*|*re*|*ről*|*től*|*vé*|*vel*),

(*ahány*|*akárhány*|*néhány*|*sok*|*valahány*)(∅|*ak*|*akat*|*akba*|*akban*|*akból*|*akért*|*akhoz*|*akig*|*akká*|*akkal*|*akként*|*aknak*|*aknál*|*akon*|*akra*|*akról*|*aktól*|*at*|*ba*|*ban*|*ból*|*ért*|*hoz*|*ig*|*ká*|*kal*|*ként*|*nak*|*nál*|*on*|*ra*|*ról*|*tól*),

(*kevés*|*kevesebb*|*legkevesebb*|*legtöbb*|*több*)(∅|*ed*|*em*|*etek*|*je*|*jei*|*jeid*|*jeik*|*jeim*|*jeink*|*jeitek*|*jük*|*ünk*)(∅|*d*|*k*|*m*|*v*)(∅|*e*|*é*|*ek*|*ekbe*|*ekben*|*ekből*|*eken*|*ekért*|*eket*|*ekhez*|*ekig*|*ekké*|*ekkel*|*ekként*|*eknek*|*eknél*|*ekre*|*ekről*|*ektől*|*el*|*en*|*en*|*ért*|*et*|*hez*|*ig*|*ként*|*nek*|*nél*|*ől*|*re*|*ről*|*től*),

(∅|*akár*|*bár*|*se*)(*ak*|*ek*|*jó*|*mek*)*kor*(*a*|*ába*|*ában*|*ából*|*áért*|*ához*|*áig*|*ák*|*ákat*|*ákba*|*ákban*|*ákból*|*aként*|*ákért*|*ákhoz*|*ákig*|*ákká*|*ákkal*|*ákként*|*áknak*|*áknál*|*ákon*|*ákra*|*ákról*|*áktól*|*án*|*ának*|*ánál*|*ára*|*áról*|*át*|*ától*|*ává*|*ával*),

(∅|*a*)*h*(*á*|*a*)*nyadik*(∅|*ak*|*akat*|*akba*|*akban*|*akból*|*akért*|*akhoz*|*akig*|*akká*|*akkal*|*akként*|*aknak*|*aknál*|*akon*|*akra*|*akról*|*aktól*|*at*|*ba*|*ban*|*ból*|*ért*|*hoz*|*ig*|*ká*|*kal*|*ként*|*nak*|*nál*|*ok*|*okat*|*okba*|*okban*|*okból*|*okért*|*okhoz*|*okig*|*okká*|*okkal*|*okként*|*oknak*|*oknál*|*okon*|*okra*|*okról*|*októl*|*on*|*ot*|*ra*|*ról*|*tól*),

(∅|*épp*|*ugyan*)*olyan*(∅|*ja*|*jai*|*jaid*|*jaik*|*jaim*|*jaink*|*jaitok*|*juk*|*od*|*om*|*otok*|*unk*)(∅|*ba*|*ban*|*ból*|*dá*|*dal*|*ért*|*hoz*|*ig*|*ká*|*kal*|*ként*|*má*|*mal*|*ná*|*nak*|*nal*|*nál*|*ok*|*okat*|*okba*|*okban*|*okból*|*okért*|*okhoz*|*okig*|*okká*|*okkal*|*okként*|*oknak*|*oknál*|*okon*|*okra*|*okról*|*októl*|*on*|*ra*|*ról*|*t*|*tól*),

(∅|*a*|*bár*|*né*|*vala*)*mely*(∅|*be*|*ben*|*ből*|*ek*|*ekbe*|*ekben*|*ekből*|*eken*|*ekért*|*eket*|*ekhez*|*ekig*|*ekké*|*ekkel*|*ekként*|*eknek*|*eknél*|*ekre*|*ekről*|*ektől*|*en*|*ért*|*et*|*hez*|*ig*|*ként*|*nek*|*nél*|*re*|*ről*|*től*),

then we consider it a Hungarian stop word. All the Hungarian stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list.

The Hungarian verbs for "buy", "carry", "drink" and "eat" have highly irregular conjugations (see Example 8.24.2 below for full lists of their conjugated forms), which also resemble other unrelated words (to untrained eyes and naïve algorithms). We are not going to include these verbs in our list of stop words, but will use their full conjugation tables to define word patterns **buyHungarian**, **carryHungarian**, **drinkHungarian** and **eatHungarian** in order to handle them separately in the word clustering algorithm below. □

Up to some modifications, our method is modeled after Anna Tordai's stemming algorithm for Hungarian (`http://snowball.tartarus.org/algorithms/hungarian/stemmer.html`). Our modifications of Tordai's stemming algorithm aim at two goals: to fully take care of verb conjugations and to fully accommodate to vowel alternations and fleeting vowels in declensions.

8.2.1 Effective spelling and essential root

It is assumed that all Hungarian words are converted to lowercase before going through any of the procedures below.

We note that Hungarian vowel lengths are sometimes used to distinguish words of unrelated etymologies, such as *agy* "brain" vs. *ágy* "bed", *örül* "rejoices" vs. *őrül* "goes crazy", and *vall* "confesses" vs. *váll* "shoulder". Meanwhile, many other words may alter between short *a* and long *á* (or short *ö* and long *ő*) during inflections. We will allow vowel length variation as a general rule, and accommodate to special cases by adding exceptions.

The pronunciations of Hungarian *s* (like English *sh*) and *sz* (like English *s*) are exactly the opposite to the Polish practices. However, for convenience, we will still replace the Hungarian diagraph *sz* by *š* for clustering purposes. Likewise, in what follows, the replacements of *gy* and *ty* by the Icelandic letters *ð* and *þ* are not motivated by phonological resemblances.

Definition 8.16 (Hungarian Vowels)**.** Hereafter in §8.2, the symbol **V** stands for any member from the list of Hungarian vowels {*a*, *á*, *e*, *é*, *i*, *í*, *o*, *ó*, *ö*, *ő*, *u*, *ú*, *ü*, *ű*}. In line with the multiplicity notations introduced in Definition 3.3, the symbol $\mathbf{V}_m$ stands for a text string formed by consecutive appearance of one or more (not necessarily identical or distinct) members from the set of Hungarian vowels.

Dual to the notations above, the symbol **C** stands for any character that does not belong to the list {*a*, *á*, *e*, *é*, *i*, *í*, *o*, *ó*, *ö*, *ő*, *u*, *ú*, *ü*, *ű*}, and $\mathbf{C}_{m_0}$ stands for a text string formed by consecutive appearance of zero or more (not necessarily identical or distinct) characters that do not belong to the same list. □

The following protected range of Hungarian words is defined according to Anna Tordai's Hungarian stemming algorithm (`http://snowball.tartarus.org/algorithms/hungarian/stemmer.html`). Some subsequent steps in our clustering algorithms are also inspired by hers. We note, however, that Anna Tordai's algorithm only treats the declensions of Hungarian nouns and adjectives (without vowel alternations and fleeting vowels), and does not fully consider the verb conjugations. Some cosmetic augmentations are needed to make the clustering algorithm work for all the test words.

Definition 8.17 (Hungarian protected range)**.** Let $\hat{\sigma}$ be a Hungarian text string, then its protected range $\mathrm{ProtRg}(\hat{\sigma}) = \max\{\lambda_1(\hat{\sigma}), \lambda_2(\hat{\sigma})\}$ is specified through the following procedures:

- Look for the string pattern (∅|*be*|*fel*|*ki*|*le*|*meg*)$\mathbf{V}_m\mathbf{C}$~ in the string $\hat{\sigma}$;
- If such a string pattern is found, its last occupying position defines $\lambda_1(\hat{\sigma})$; otherwise, set $\lambda_1(\hat{\sigma}) = 0$;
- Look for the string pattern (∅|*be*|*fel*|*ki*|*le*|*meg*)$\mathbf{C}_m\mathbf{V}$~ in the string $\hat{\sigma}$;
- If such a string pattern is found, its last occupying position defines $\lambda_2(\hat{\sigma})$; otherwise, set $\lambda_2(\hat{\sigma}) = 0$. □

Algorithm 8.18 (Hungarian effective spelling)**.** *For a Hungarian word $\hat{\sigma}$ (converted to lower case form), its effective spelling* $\mathrm{EffSpell}(\hat{\sigma})$ *is constructed in sequential steps:*

(1) *Convert to lowercase, do* -**X** → ∅ *(i.e. delete everything after hyphen), dr*~ → *δr, tr*~ → *τr,* <u>*visszük*</u> → *viszik, ~nyig* → *ny, ~r*(*a*|*e*|*o*)*st*(*u*|*ü*)*l* → *r, before replacing*[102]

(∅|*be*)*vall*~	(∅|*édes*)*any*(∅|*j*)	(∅|*édes*)*ap*(*a*|*á*)	(∅|*édes*)*apj*
αδμτ	*any*	*χφαδρ*	*χφαδρ*

(∅|*leg*|*legesleg*)*finom*~	(∅|*leg*|*legesleg*)*gazdag*~	(∅|*leg*|*legesleg*)*hideg*~
φiῖvm	*ρiξg*	*κολδg*

(∅|*leg*|*legesleg*)*magá*(∅|*n*)*ny*~	(∅|*leg*|*legesleg*)*meleg*~	(∅|*leg*|*legesleg*)*messz*~	(∅|*leg*|*legesleg*)*sim*~
πριβ	*ωαρμg*	*φαρssz*	*σiμ*

(∅|*leg*|*legesleg*)*szórak*~	(∅|*leg*|*legesleg*)*vidám*~	(∅|*meg*)*ért*~	(∅|*meg*)*eskü*(*d*|*sz*)~
eῖvjok	*τζiραm*	*ῖvαστῖvδ*	*σεωρüd*

(∅|*nagy*)*nén*~	(*báty*|*bátty*|*fivér*|*fivérr*|*öcs*|*öccs*)~	(*leg*|*legesleg*)*kisebb*~	*agg*~
αuῖvτ	*βρατ*	*kis*	*ωρix*

ágy	*alak*~	*áll*~	*általános*	*anyag*~	*apu*~	*ár*~	*asztal*~	*bácsi*~	*báj*	*bál*~	*bámul*	*bán*
βeδgy	*αλακk*	*ααλl*	*γeῖvεραλs*	*ματg*	*χφαδρ*	*αar*	*ταβl*	*uῖvκλi*	*τχαρμj*	*βάλ*	*γαζl*	*μiῖvδn*

bát(*or*|*ra*)(∅|*bb*|*k***X**)	*bennet***X**~	*bessie*~	*beteg*	*charles*	*család*	*csoport*	*dönt*
βαρανvr	*βeῖvῖv*	*βεσθιe*	*ιiλg*	*καρλ*	*φad*	*γυρπt*	*δikδt*

egész~	*eliz*(*á*|*a*)**X**~	*élő*$\widehat{(s)}$**X**	*érkez*	*esküvő*~	*est*~	*ezredes*	*faj*	*fej*	*fejlett*	*férj*
iῖvτsz	*eλiζ*	*él*	*ραrβz*	*ωeδő*	*eεβ*	*ωβερστ*	*σεπj*	*ηeαδj*	*δevt*	*χuσβj*

fiatal	*font*	*form*(*a*|*á*)	*fosszil*	*gyermek*~	*haj*	*hál*(*a*|*á*)~	*hasz*(∅|*o*)*n*	*hat*~	*hiáb*
γul	*πuῖvδt*	*φορμa*	*φοσσiλ*	*gyerek*	*ηaiρj*	*θaγῖvκa*	*uσεn*	*6at*	*hiú*

hiány	*hív*	*hivat*	*hónap*~	*hossz*	*idom*~	*indul*	*jaj*~	*játék*	*jones*	*kemény*	*kezd*
λacκny	*καλv*	*σeῖvδet*	*μoῖvθp*	*λoš*	*φiγρm*	*βeγῖvl*	*ωoεj*	*τoηk*	*γjoneσ*	*χarδny*	*βeγin*

kisasszon(∅|*n*)*y*	*kíván*	*könyv*	*könyvtár*	*kösz*	*lá*(*sd*|*ss*|*t*)	*lament*~	*leah*~	*lev*(*e*|*é*)*l*	*lyhet*
μisny	*ωisχon*	*βkuv*	*λiβραr*	*θaῖvκ*	*σeεt*	*λaμeῖvτ*	*λeaη*	*λetρl*	*lyhek*

[102] We are being Anglocentric here: *báty* "elder brother", *fivér* "male sibling", *öcs* "younger brother" are merged into one class; the same applies to *nővér* "elder sister" and *húg* "younger sister". This is more a temporary compromise than a generally healthy practice. In fact, many Asian languages (including Chinese, Japanese, Korean, Khmer, Malay, Mongolian, Tamil, Telugu, Thai, Vietnamese etc.) possess distinct words for siblings (of given gender) depending on seniority.

mad(á|a)r	*madam*	*mam(a|á)*	*mari~*	*maso~*	*más~*	*méh*	*mesél~*	*mrs~*	*mulat~*	*munk*$^{\mathbf{X}\in}$*(a|á)~*	*muszl~*
βιρδr	*μαδαμm*	*anya*	*μααρi*	*μασο*	*οθρs*	*βεεh*	*τελl*	*φυρx*	*φut*	*ωορκ***X**	*μυσλ*

nagybá(csi|ty|tty)~	*nem~*	*nővér(∅|r)*	*ook*	*oor*	*orr~*	*ow*	*őrül*	*örül*	*ösztön(∅|ös)(∅|s***V***)*	*ösztönöz*	*palánt(a|á)*
uṽκli	*ṽeμ*	*húg*	*oωκ*	*ωορ*	*ṽeζ*	*oω*	*μaδl*	*ρjoil*	*iṽστiṽ*	*iṽστiṽzi*	*πλaṽτa*

papír	*pár~*	*poh(á|a)r*	*puszt*	*ree~*	*rég~*	*remé(l|ny|nny)~*	*rend*	*rossz*	*rovar*	*sajn~*	*szabad*
παπiρ	*παaρ*	*κοππr*	*δεστρ*	*ρεε*	*oλδg*	*χοπl*	*ορδd*	*βaš*	*iṽσεκτ*	*σορn*	*φριd*

száj~	*szalon~*	*szám*	*szem~*	*szer*	*szeré(∅|n)ny~*	*szív*	*szob(a|á)*	*tábor*	*táj~*	*táv*	
μuṽδj	*ζaλωn*	*száμm*	*eγεm*	*szrer*	*μοδστt*	*κορδαv*	*ρομ*	*καμπor*	*ρεγṽj*	*φárv*	

tavasz~	*te(a|á)~*	*te(ss|tsz|tssz)*	*tekint*	*terem*	*terhet*	*term*	*terv*	*tetszés*	*téved*
leβρsz	*τζaji*	*λikγnδ*	*nézt*	*τερem*	*terhek*	*τερm*	*πλanv*	*σατφis*	*εορed*

tit(∅|o)k~	*tojás*	*törvény~*	*vad*	*vála(szd|ssz|szt)***X**	*váll*	*való*	*vált*	*változ***X**	*város*	
σεκρk	*eγγs*	*λaωny*	*ωiλd*	*σελεκt*	*σohδl*	*iṽδεεδó*	*ξaṽγεt*	*ξaṽγεt*	*κiτης*	

véd	*vétk*	*világ*	*vissz~*	*zavar*	*zs(á|é)g*	*cs*	*jane*	*miss*	*sir*	*tv*	*~kénti*
προκτod	*σiṽik*	*ωελτeg*	*βκ*	*κοṽφσr*	*z*	*γuρπt*	*γοiṽṽα*	*μiσx*	*σiρx*	*λaωny*	*∅*

(2) *Do αart~→ ηυρt,* (*∅|el|meg*)*vár→ ωαatρ,* (*∅|meg*)**eatHungarian**→*aεατ,* **drinkHungarian**→*εδριṽκ, mé*(*gy|sz*)→*menni,* ~**buyHungarian** →*hβuηh,* ~**carryHungarian** →*hκρταρh*

(3) *Replace*

(*∅|fel|meg*)*kér~*	(*atlan|etlen|talan|telen|tlan|tlen*)							(*két|kettő*)*~*	*ablak*	*ajt~*	*álm*	*almo~*		
kér	*zτ*							*2két*	*abλaκ*	*ajτ*	*αulm*	*aλmo*		

álom	*alom~*	*barát~*	*beszé(d|l)*	*egy~*	*el~*	*év~*	*gyerek~*	*hall~*	*hib*	*ifjú~*	*igy~*	*jó~*	*kedv*	*kert~*
αulom	*aλom*	*βarát*	*beszéλ*	*zgy*	*λl*	*yév*	*gyereκ*	*χalλ*	*hiβ*	*ifja*	*qgy*	*jo*	*keδβ*	*kerτ*

kicsi~	*leány~*	*lép*	*lev~*	*ond*	*orvos*	*szebb*	*szó~*	*tér~*	*tiszt*$^{\mathbf{X}\in}$*(a|á|i|o)*	*tud*	*új~*	*úr~*	*vagyon~*	*mr*	*~l(a|e)g*	*~th*
kis	*lány*	*léπ*	*lev*	*onδ*	*δοκτ*	*szépbb*	*szav*	*tép*	*τiszτ***X**	*tuδ*	*ωúj*	*ur*	*wagyon*	*xηερ*	*∅*	*t*

(4) *Do tiszt→tiszτ,* (*∅|leg|legesleg*)**X***bb* → **X′**, *where one constructs* **X′** *by doing ~á→a, ~é→e on* **X**.

(5) *Replace*

ccs	*ggy*	*lly*	*nny*	*ssz*	*tty*	*zzs*
čč	*ðð*	*łł*	*ññ*	*šš*	*þþ*	*žž*

(6) *Replace*

cs	*gy*	*ly*	*ny*	*sz*	*ty*	*zs*
č	*ð*	*ł*	*ñ*	*š*	*þ*	*ž*

and call the result $\hat{\sigma}'$.

(7) *Break down* $\hat{\sigma}' = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma}')]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$, *where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma}')$ *is equal to the protected range of* $\hat{\sigma}'$.

(8) *On* $\hat{\sigma}_1$, *do* (*be|fel|le|meg*)*~→∅, ki*$_{\hat{\chi}\notin}$(*s*)*~→*$\hat{\chi}$, *and call the result* $\hat{\sigma}'_1$.

(9) *On* $\hat{\sigma}_2$, *perform the following substitutions in a sequel:*

(9.1) *Do* ~(*ad|al*(*∅|o*)*m|áš|beli|dal*(*∅|o*)*m|del*(*∅|e*)*m|ék*(*e|o*)*ñ|el*(*∅|e*)*m|éš|h*(*a|e*)*t|ist*(*a|á*)*|ít|ñi|omás|s*(*á|é*)*g|šerű*)**X** → *∅*, ~(*∅|m|v*)(*á|é*)*ñ***X** → *∅*, ~*izá*(*ció|l*)**X**→ *∅*, ~(*st*(*u|ü*)*l*)**X** → *∅*, ~(*anak|enek|j*(*á|é*)*k|n*(*á|é*)(*k|m*)) → *∅*, ~((*á|e|é|o*)*s|k*(*e|o|ö*)(*d|z*))**X** → *∅*, ~*t*(*ál|am|él|em*) → *∅*, ~*ni*(*∅|a|e*) → *∅*, ~$\hat{\chi}_{\times 2}$(*a|e*)*l* → $\hat{\chi}$, ~$\hat{\chi}_{\hat{\chi}\notin}$(*ł|r*)*h*(*a|e*)*t* → $\hat{\chi}$, ~(*∅|and*)*ó* → *∅*, ~(*∅|e|en*)(*∅|d*)*ő* → *∅*, ~*l*(*a|e*)*k* → *∅*, ~**V***tt* → *∅*.

(9.2) *Do* ~$^{\mathbf{X}\in}$(*∅|á|é*)((*a|e|o|ö*)*n*(*∅|ként*)|(*a|e|o|ö*)*t|b*(*a|e*)(*∅|n*)|*b*(*ó|ő*)*l|ért|h*(*e|o|ö*)*z|ig|ként|képp*(*∅|en*)) → **X′**, ~$^{\mathbf{X}\in}$(*∅|á|é*)(*n|n*(*a|e*)*k|n*(*á|é*)*l|r*(*a|e*)|(*r|t*)(*ó|ő*)*l|t*|(*u|ü*)*l|v*(*á|é*)*|v*(*a|e*)*l*) → **X′**, *where one constructs* **X′** *by doing ~á→a, ~é→e on* **X**.

(9.3) *Do ~án*(*∅|ként*) → *a, ~énként→ e, ~lom→ lm,* ~(*n*(*a|á|ának|ánk|e|é|ének|énk*)|*t*(*á|é*)*k*) → *∅*.

(9.4) *Do* ~*ass***X** → ∅, ~*šš***X** →*š*, ~(*a*|*e*|*o*|*ö*)*st* → ∅, ~*ást* → *a*, ~*ést* → *e*.

(9.5) *Do* ~$\hat{\chi}_{\times 2}$(*á*|*é*) → $\hat{\chi}$, ~**C**(*d*|*i*|*j*) → **C**, ~*n*(*e*|*o*)*d* → ∅.

The result after these five steps of operations is called $\hat{\sigma}'_2$.

(10) *Concatenate* $\hat{\sigma}'_1$ *and* $\hat{\sigma}'_2$.

Algorithm 8.19 (Hungarian essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Hungarian word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do* ~$\hat{\chi}_{\times 2}$(*á*|*é*)(∅|*l*) → $\hat{\chi}$, ~(*e*|*o*|*ö*)*l* → ∅, *and call the result* $\hat{\sigma}'$.

(2) *Break down* $\hat{\sigma}' = \hat{\sigma}_1\hat{\sigma}_2$ *into the concatenation of two strings* $\hat{\sigma}_1 = \hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma}')]}$ *(see the notation in Definition 3.1) and* $\hat{\sigma}_2$*, where the length of the first string* $\ell(\hat{\sigma}_1) = \mathrm{ProtRg}(\hat{\sigma}')$ *is equal to the protected range of* $\hat{\sigma}'$.

(3) *On* $\hat{\sigma}_2$*, perform the following substitutions in a sequel:*

(3.1) *Do* ~*dd* → ∅, ~(∅|*a*|*e*|*o*|*ö*)*ké* → ∅, ~*é*(∅|*i*) → ∅, ~*á*(*éi*|*ké*) → *a*, ~*é*(*é*|*éi*|*ké*) → *e;*

(3.2) *Do* ~(∅|*a*|*e*|*o*|*ö*)*d* → ∅, ~(∅|*an*|*n*)(∅|*a*|*e*|*o*)*m* → ∅, ~(∅|*j*)(*a*|*e*) → ∅, ~(∅|*j*|*ni*|*t*)(*u*|*ü*)*k* → ∅, ~(∅|*u*|*ü*)*nk* → ∅, ~*á*(∅|*d*|*juk*|*m*|*nk*) → *a*, ~*é*(∅|*d*|*jük*|*m*|*nk*) → *e*, ~*o* → ∅*;*

(3.3) *Do* ~(∅|(∅|*j*)(*ai*|*o*))*tok* → ∅, ~(∅|*j*)(∅|*e*)(∅|*i*)*tek* → ∅, ~(∅|*j*)(*a*|*e*)(*i*|*id*|*ik*|*im*|*ink*) → ∅, ~*á*(*i*|*id*|*ik*|*im*|*ink*|*tok*) → *a*, ~*é*(*i*|*id*|*ik*|*im*|*ink*|*tek*) → *e*, ~*tök* → ∅*;*

(3.4) *Do* ~*ák* → *a*, ~*ék* → *e*, ~(*a*|*e*|*i*|*o*|*ö*)*k* → ∅.

The result after these four steps of operations is called $\hat{\sigma}'_2$.

(4) *Concatenate* $\hat{\sigma}_1$ *and* $\hat{\sigma}'_2$.

(5) *Do* ~*tt*(∅|**V**) → ∅.

(6) *Do* ~$\hat{\chi}$(*j*|*n*|*š*|*t*|*v*)(∅|**V**)(∅|*l*) → ∅.

(7) *Do* $\hat{\chi}_{\times 2} \to \hat{\chi}$, ~$\hat{\chi}^{\in}$(*ĺ*|*r*)*h* → *h*$\hat{\chi}$.

(8) *Replace*

kis~	(*ðe*|*ðer*|*jö*|*jő*|*jöh*|*jöj*|*jös*)	**X**$^{\in}$(∅|*h*|*v*)*i*(∅|*ð*|*ðe*|*ðél*|*h*|*š*)	**X**$^{\in}$(*l*|*v*)(*é*|*e*|*eð*|*éð*|*eðél*|*eh*)	*me*(∅|*ð*|*h*)	*ve*(*ðe*|*š*)
κισ	*jön*	**X***in*	**X***en*	*men*	*ven*

For clustering purposes, we will also define an operation to shorten the last long vowel in the effective spelling of a Hungarian word.

Algorithm 8.20 (Length reduction of final vowel)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Hungarian word, then* RdFinLV($\hat{\sigma}$) *is defined by the following procedure:*

- *Look for the pattern* **V** *in* $\hat{\sigma}$.
- *If the pattern is not found, do nothing. Otherwise, do* (*á*|*o*|*ó*) → *a*, *é* → *e*, *ő* → *ö on its last occupying position.*

8.2.2 Admissible mutation and approximate clustering

Algorithm 8.21 (Simple heredity test)**.** *Construct* $\hat{\alpha}'$ *by doing* ~(*d*|*e*|*i*|*n*|*t*|*v*) → ∅ *on* $\hat{\alpha}$*. Define* $\hat{\beta}'$ *similarly. The Boolean-valued function* SimpHrdTest($\hat{\alpha}$,$\hat{\beta}$) *returns* **TRUE** *if the lowercase form of* $\hat{\alpha}$ *contains at least one instance of* {*a, á, e, é, i, í, o, ó, ö, ő, u, ú, ü, ű*} **AND** *at least one of the following three conditions holds*:[103]

(i) $\hat{\alpha} = \hat{\beta}$*;*

(ii) $\hat{\beta} = \hat{\alpha}$(*d*|*m*|*k*|*v*)*;*

(iii) $\ell(\hat{\beta}) > \ell(\hat{\alpha}) \geq \frac{\ell(\hat{\beta})}{2}$ **AND** $\hat{\beta} = \hat{\alpha}$**VX**.

In what follows, we define SuffixNW($\hat{\alpha}$,$\hat{\beta}$), RootNW($\hat{\alpha}$,$\hat{\beta}$), NW*($\hat{\alpha}$,$\hat{\beta}$) and SuffixSW($\hat{\alpha}$,$\hat{\beta}$), SuffixSW($\hat{\alpha}$,$\hat{\beta}$), SW*($\hat{\alpha}$,$\hat{\beta}$) as what is done in the Danish case (Algorithm 5.7), which was also applied to two other Germanic languages treated in §5.

[103] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

Algorithm 8.22 (Admissible suffix mismatch and vowel alternation)**.** *For two strings $\hat{\alpha}$ and $\hat{\beta}$, the Boolean-valued function*

$$\text{AdmMut}(\text{RootNW}(\hat{\alpha},\hat{\beta}),\text{SuffixNW}(\hat{\alpha},\hat{\beta}),\text{NW}^*(\hat{\alpha},\hat{\beta}))$$

returns **TRUE** *if at least one of the following two conditions holds:*

(i) $\text{SuffixNW}(\hat{\alpha},\hat{\beta}) = [\varnothing|\mathbf{V}k, \varnothing|\mathbf{V}k]$ **AND** ($\text{NW}^*(\hat{\alpha},\hat{\beta}) = [\varnothing,\ e]||[\varnothing,\ o]||[\varnothing,\ ö]||[a,\ á]||[e,\ é]||[o,\ ó]$ **OR** $\text{NW}^*(\hat{\beta},\hat{\alpha}) = [\varnothing,\ e]||[\varnothing,\ o]||[\varnothing,\ ö]||[a,\ á]||[e,\ é]||[o,\ ó]$ **OR** ($\text{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$ **AND** *the lowercase form of* $\text{RootNW}(\hat{\alpha},\hat{\beta})$ *contains at least one instance of* **V**)*;*

(ii) $\text{NW}^*(\hat{\alpha},\hat{\beta}) = \varnothing$ **AND** $\text{SuffixNW}'(\hat{\alpha},\hat{\beta}) = [a,\ ó]||[ava,\ ó]||[ö,\ ő]||[öve,\ ő]$*, where one obtains* $\text{SuffixNW}'(\hat{\alpha},\hat{\beta})$ *by doing* $\sim(k|st|v) \to \varnothing$ *on both components of* $\text{SuffixNW}(\hat{\alpha},\hat{\beta})$*.*

Similarly, one can evaluate another Boolean-valued function

$$\text{AdmMut}(\text{RootSW}(\hat{\alpha},\hat{\beta}),\text{SuffixSW}(\hat{\alpha},\hat{\beta}),\text{SW}^*(\hat{\alpha},\hat{\beta}))$$

by trading all the occurrences of NW *in the statements above with* SW*.*

Algorithm 8.23 (Heredity test function)**.** *The structure of the Hungarian heredity test function* $\text{HrdTest}(\hat{\alpha},\hat{\beta})$ *is identical to the German version (Algorithm 8.1.2), except that the functions* SimpHrdTest*,* RootNW*,* SuffixNW*,* NW^**,* RootSW*,* SuffixSW*,* SW^* *must follow the Hungarian rules stated above.*

The structure of the following clustering algorithm differs from the Finnish counterpart only in cosmetic details.

Algorithm 8.24 (Approximate clustering of Hungarian words)**.** *The approximate clustering of a list of Hungarian words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1,\text{EffSpell}(\hat{\alpha}_1),\text{RdFinLV}(\text{EffSpell}(\hat{\alpha}_1))), \ldots, (\hat{\alpha}_N,\text{EffSpell}(\hat{\alpha}_N),\text{RdFinLV}(\text{EffSpell}(\hat{\alpha}_N)))\}$ *alphabetically according to the third component (with higher priority) and the second component (with lower priority). If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)},\text{EffSpell}(\hat{\alpha}_{(1)}),\text{RdFinLV}(\text{EffSpell}(\hat{\alpha}_{(1)}))), \ldots, (\hat{\alpha}_{(N)},\ \text{EffSpell}(\hat{\alpha}_{(N)}),\ \text{RdFinLV}(\text{EffSpell}(\hat{\alpha}_{(N)})))\}$ *satisfies both* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n)}),\text{EffSpell}(\hat{\alpha}_{(n+1)})) = \textbf{FALSE}$ *and* $\text{HrdTest}(\text{EffSpell}(\hat{\alpha}_{(n+1)}),\ \text{EffSpell}(\hat{\alpha}_{(n)})) = \textbf{FALSE}$*, where $n \in \mathbb{Z} \cap [1, N)$, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words with tags:* $\{g_1 = \{(\hat{\alpha}_{(1,1)},\text{EffSpell}(\hat{\alpha}_{(1,1)}),\ldots), \ldots, (\hat{\alpha}_{(1,n_1)},\ \text{EffSpell}(\hat{\alpha}_{(1,n_1)}),\ldots)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)},\text{EffSpell}(\hat{\alpha}_{(M,1)}),\ldots), \ldots, (\hat{\alpha}_{(M,n_M)},\text{EffSpell}(\hat{\alpha}_{(M,n_M)}),\ldots)\}\}$*, where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$*.*

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)},\text{EffSpell}(\hat{\alpha}_{(m,1)})), \ldots, (\hat{\alpha}_{(m,n_m)},\text{EffSpell}(\hat{\alpha}_{(m,n_m)}))\}$ *where $m \in \mathbb{Z} \cap [1, M]$, we augment it into a tagged entry* $G_m = (g_m,\text{EffSpell}(\hat{\alpha}_{(m,1)}),\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)})), T_m)$*, where*

$$T_m = \text{RdFinLV}(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))).$$

The list $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* T_m *(with highest priority),* $\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with medium priority), and* $\text{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lowest priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)},\hat{\gamma}'_{(1)},\hat{\gamma}''_{(1)},\hat{\gamma}'''_{(1)}), \ldots, G_{(M)} = (g_{(M)},\hat{\gamma}'_{(M)},\hat{\gamma}''_{(M)},\hat{\gamma}'''_{(M)})\}$ *satisfy*

$$\text{HrdTest}(\hat{\gamma}''_{(m)},\hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)},\hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, then add a demarcation line between these two entries. In this manner, the alphabetized list $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, generate a list of word clusters* $\{\check{\Gamma}_1 = (\check{G}_{(1,1)},\ldots), \ldots, \check{\Gamma}_K = (\check{G}_{(K,1)},\ldots)\}$ *by discarding all the tags.*

Example 8.24.1. The following samples of regular nouns (declined in all cases, in singular and plural) are extracted from [48, Appendix 2].

> *álmaitok*, *álmaitokat*, *álmaitokba*, *álmaitokban*, *álmaitokból*, *álmaitokért*, *álmaitokhoz*, *álmaitokig*, *álmaitokká*, *álmaitokkal*, *álmaitoknak*, *álmaitoknál*, *álmaitokon*, *álmaitokra*, *álmaitokról*, *álmaitoktól*, *álmotok*, *álmotokat*, *álmotokba*, *álmotokban*, *álmotokból*, *álmotokért*, *álmotokhoz*, *álmotokig*, *álmotokká*, *álmotokkal*, *álmotokként*, *álmotoknak*, *álmotoknál*, *álmotokon*, *álmotokra*, *álmotokról*, *álmotoktól*, *álom* — "your dream";

bokor, *bokorba*, *bokorban*, *bokorból*, *bokorért*, *bokorhoz*, *bokorig*, *bokorként*, *bokornak*, *bokornál*, *bokorra*, *bokorrá*, *bokorral*, *bokorról*, *bokortól*, *bokrok*, *bokrokat*, *bokrokba*, *bokrokban*, *bokrokból*, *bokrokért*, *bokrokhoz*, *bokrokig*, *bokrokká*, *bokrokkal*, *bokrokként*, *bokroknak*, *bokroknál*, *bokrokon*, *bokrokra*, *bokrokról*, *bokroktól*, *bokron*, *bokronként*, *bokrostul*, *bokrot* — “bush”;

dinnye, *dinnyébe*, *dinnyében*, *dinnyéből*, *dinnyéért*, *dinnyéhez*, *dinnyéig*, *dinnyék*, *dinnyékbe*, *dinnyékben*, *dinnyékből*, *dinnyéken*, *dinnyeként*, *dinnyékért*, *dinnyéket*, *dinnyékhez*, *dinnyékig*, *dinnyékké*, *dinnyékkel*, *dinnyékként*, *dinnyéknek*, *dinnyéknél*, *dinnyékre*, *dinnyékről*, *dinnyéktől*, *dinnyén*, *dinnyének*, *dinnyénél*, *dinnyénként*, *dinnyére*, *dinnyéről*, *dinnyéstül*, *dinnyét*, *dinnyétől*, *dinnyévé*, *dinnyével* — “melon”;

ház, *házak*, *házakat*, *házakba*, *házakban*, *házakból*, *házakért*, *házakhoz*, *házakig*, *házakká*, *házakkal*, *házakként*, *házaknak*, *házaknál*, *házakon*, *házakra*, *házakról*, *házaktól*, *házanként*, *házastul*, *házat*, *házba*, *házban*, *házból*, *házért*, *házhoz*, *házig*, *házként*, *háznak*, *háznál*, *házon*, *házra*, *házról*, *háztól*, *házzá*, *házzal* — “house”;

ház, *házaim*, *házaimat*, *házaimba*, *házaimban*, *házaimból*, *házaimért*, *házaimhoz*, *házaimig*, *házaimmá*, *házaimmal*, *házaimnak*, *házaimnál*, *házaimon*, *házaimra*, *házaimról*, *házaimtól*, *házam*, *házamat*, *házamba*, *házamban*, *házamból*, *házamért*, *házamhoz*, *házamig*, *házamként*, *házammá*, *házammal*, *házamnak*, *házamnál*, *házamon*, *házamra*, *házamról*, *házamtól* — “my house”;

iker, *ikerbe*, *ikerben*, *ikerből*, *ikerért*, *ikerhez*, *ikerig*, *ikerként*, *ikernek*, *ikernél*, *ikerre*, *ikerré*, *ikerrel*, *ikerről*, *ikertől*, *ikrek*, *ikrekbe*, *ikrekben*, *ikrekből*, *ikreken*, *ikrekért*, *ikreket*, *ikrekhez*, *ikrekig*, *ikrekké*, *ikrekkel*, *ikrekként*, *ikreknek*, *ikreknél*, *ikrekre*, *ikrekről*, *ikrektől*, *ikren*, *ikrenként*, *ikrestül*, *ikret* — “twin”;

jutalmaink, *jutalmainkat*, *jutalmainkba*, *jutalmainkban*, *jutalmainkból*, *jutalmainkért*, *jutalmainkhoz*, *jutalmainkig*, *jutalmainkká*, *jutalmainkkal*, *jutalmainknak*, *jutalmainknál*, *jutalmainkon*, *jutalmainkra*, *jutalmainkról*, *jutalmainktól*, *jutalmunk*, *jutalmunkat*, *jutalmunkba*, *jutalmunkban*, *jutalmunkból*, *jutalmunkért*, *jutalmunkhoz*, *jutalmunkig*, *jutalmunkká*, *jutalmunkkal*, *jutalmunkként*, *jutalmunknak*, *jutalmunknál*, *jutalmunkon*, *jutalmunkra*, *jutalmunkról*, *jutalmunktól*, *jutalom* — “our reward”;

kép, *képeik*, *képeikbe*, *képeikben*, *képeikből*, *képeiken*, *képeikért*, *képeiket*, *képeikhez*, *képeikig*, *képeikké*, *képeikkel*, *képeiknek*, *képeiknél*, *képeikre*, *képeikről*, *képeiktől*, *képük*, *képükbe*, *képükben*, *képükből*, *képükért*, *képüket*, *képükhöz*, *képükig*, *képükké*, *képükkel*, *képükként*, *képüknek*, *képüknél*, *képükön*, *képükre*, *képükről*, *képüktől* — “their picture”;

kéz, *kézbe*, *kézben*, *kézből*, *kezek*, *kezekbe*, *kezekben*, *kezekből*, *kezeken*, *kezekért*, *kezeket*, *kezekhez*, *kezekig*, *kezekké*, *kezekkel*, *kezekként*, *kezeknek*, *kezeknél*, *kezekre*, *kezekről*, *kezektől*, *kézen*, *kezenként*, *kézért*, *kezestül*, *kezet*, *kézhez*, *kézig*, *kézként*, *kéznek*, *kéznél*, *kézre*, *kézről*, *kéztől*, *kézzé*, *kézzel* — “hand”;

kő, *kőbe*, *kőben*, *kőből*, *kőért*, *kőhöz*, *kőig*, *kőként*, *kőnek*, *kőnél*, *kőre*, *kőről*, *kőtől*, *kővé*, *kövek*, *kövekbe*, *kövekben*, *kövekből*, *köveken*, *kövekért*, *köveket*, *kövekhez*, *kövekig*, *kövekké*, *kövekkel*, *kövekként*, *köveknek*, *köveknél*, *kövekre*, *kövekről*, *kövektől*, *kővel*, *kövenként*, *kövestül*, *követ*, *kövön* — “stone”;

könyv, *könyvbe*, *könyvben*, *könyvből*, *könyvek*, *könyvekbe*, *könyvekben*, *könyvekből*, *könyveken*, *könyvekért*, *könyveket*, *könyvekhez*, *könyvekig*, *könyvekké*, *könyvekkel*, *könyvekként*, *könyveknek*, *könyveknél*, *könyvekre*, *könyvekről*, *könyvektől*, *könyvenként*, *könyvért*, *könyvestül*, *könyvet*, *könyvhöz*, *könyvig*, *könyvként*, *könyvnek*, *könyvnél*, *könyvön*, *könyvre*, *könyvről*, *könyvtől*, *könyvvé*, *könyvvel* — “book”;

lány, *lányastul*, *lányba*, *lányban*, *lányból*, *lányért*, *lányhoz*, *lányig*, *lányként*, *lánynak*, *lánynál*, *lánnyá*, *lánnyal*, *lányok*, *lányokat*, *lányokba*, *lányokban*, *lányokból*, *lányokért*, *lányokhoz*, *lányokig*, *lányokká*, *lányokkal*, *lányokként*, *lányoknak*, *lányoknál*, *lányokon*, *lányokra*, *lányokról*, *lányoktól*, *lányon*, *lányonként*, *lányra*, *lányról*, *lányt*, *lánytól* — “girl”;

levél, *leveled*, *leveledbe*, *leveledben*, *leveledből*, *leveleddé*, *leveleddel*, *leveleden*, *leveledért*, *leveledet*, *leveledhez*, *leveledig*, *leveledként*, *levelednek*, *levelednél*, *leveledre*, *leveledről*, *leveledtől*, *leveleid*, *leveleidbe*, *leveleidben*, *leveleidből*, *leveleiddé*, *leveleiddel*, *leveleiden*, *leveleidért*, *leveleidet*, *leveleidhez*, *leveleidig*, *leveleidnek*, *leveleidnél*, *leveleidre*, *leveleidről*, *leveleidtől* — “your letter”;

lova, *lovába*, *lovában*, *lovából*, *lováért*, *lovához*, *lovai*, *lovaiba*, *lovaiban*, *lovaiból*, *lovaiért*, *lováig*, *lovaihoz*, *lovaiig*, *lovaiként*, *lovain*, *lovainak*, *lovainál*, *lovaira*, *lovairól*, *lovait*, *lovaitól*, *lovaivá*, *lovaival*, *lovaként*, *lován*, *lovának*, *lovánál*, *lovára*, *lováról*, *lovát*, *lovától*, *lovává*, *lovával* — “his/her horse”;

őr, *őrbe*, *őrben*, *őrből*, *őrért*, *őrestül*, *őrhöz*, *őrig*, *őrként*, *őrnek*, *őrnél*, *őrök*, *őrökbe*, *őrökben*, *őrökből*, *őrökért*, *őröket*, *őrökhöz*, *őrökig*, *őrökké*, *őrökkel*, *őrökként*, *őröknek*, *őröknél*, *őrökön*, *őrökre*, *őrökről*, *őröktől*, *őrön*, *őrönként*, *őrre*, *őrré*, *őrrel*, *őrről*, *őrt*, *őrtől* — "guard";

pohár, *poharak*, *poharakat*, *poharakba*, *poharakban*, *poharakból*, *poharakért*, *poharakhoz*, *poharakig*, *poharakká*, *poharakkal*, *poharakként*, *poharaknak*, *poharaknál*, *poharakon*, *poharakra*, *poharakról*, *poharaktól*, *poharanként*, *poharastul*, *poharat*, *pohárba*, *pohárban*, *pohárból*, *pohárért*, *pohárhoz*, *pohárig*, *pohárként*, *pohárnak*, *pohárnál*, *poháron*, *pohárra*, *pohárrá*, *pohárral*, *pohárról*, *pohártól* — "glass";

szavak, *szavakat*, *szavakba*, *szavakban*, *szavakból*, *szavakért*, *szavakhoz*, *szavakig*, *szavakká*, *szavakkal*, *szavakként*, *szavaknak*, *szavaknál*, *szavakon*, *szavakra*, *szavakról*, *szavaktól*, *szavanként*, *szavastul*, *szavon*, *szó*, *szóba*, *szóban*, *szóból*, *szóért*, *szóhoz*, *szóig*, *szók*, *szóként*, *szónak*, *szónál*, *szóra*, *szóról*, *szót*, *szótól*, *szóvá*, *szóval* — "word";

szék, *székbe*, *székben*, *székből*, *székek*, *székekbe*, *székekben*, *székekből*, *székeken*, *székekért*, *székeket*, *székekhez*, *székekig*, *székekké*, *székekkel*, *székekként*, *székeknek*, *székeknél*, *székekre*, *székekről*, *székektől*, *széken*, *székenként*, *székért*, *székestül*, *széket*, *székhez*, *székig*, *székké*, *székkel*, *székként*, *széknek*, *széknél*, *székre*, *székről*, *széktől* — "chair";

táska, *táskába*, *táskában*, *táskából*, *táskáért*, *táskához*, *táskáig*, *táskák*, *táskákat*, *táskákba*, *táskákban*, *táskákból*, *táskaként*, *táskákért*, *táskákhoz*, *táskákig*, *táskákká*, *táskákkal*, *táskákként*, *táskáknak*, *táskáknál*, *táskákon*, *táskákra*, *táskákról*, *táskáktól*, *táskán*, *táskának*, *táskánál*, *táskánként*, *táskára*, *táskáról*, *táskástul*, *táskát*, *táskától*, *táskává*, *táskával* — "bag";

tavak, *tavakat*, *tavakba*, *tavakban*, *tavakból*, *tavakért*, *tavakhoz*, *tavakig*, *tavakká*, *tavakkal*, *tavakként*, *tavaknak*, *tavaknál*, *tavakon*, *tavakra*, *tavakról*, *tavaktól*, *tavanként*, *tavastul*, *tavat*, *tavon*, *tó*, *tóba*, *tóban*, *tóból*, *tóért*, *tóhoz*, *tóig*, *tóként*, *tónak*, *tónál*, *tóra*, *tóról*, *tótól*, *tóvá*, *tóval* — "lake";

tükör, *tükörbe*, *tükörben*, *tükörből*, *tükörért*, *tükörhöz*, *tükörig*, *tükörként*, *tükörnek*, *tükörnél*, *tükörre*, *tükörré*, *tükörrel*, *tükörről*, *tükörtől*, *tükrök*, *tükrökbe*, *tükrökben*, *tükrökből*, *tükrökért*, *tükröket*, *tükrökhöz*, *tükrökig*, *tükrökké*, *tükrökkel*, *tükrökként*, *tükröknek*, *tükröknél*, *tükrökön*, *tükrökre*, *tükrökről*, *tükröktől*, *tükrön*, *tükrönként*, *tükröstül*, *tükröt* — "mirror".

These regular nouns are clustered correctly by our algorithm.

We then select three nouns listed in [48, Appendix 3, Section "Metathesis"], as their declensions deviate significantly from the regular patterns covered above.

kehely, *kehelybe*, *kehelyben*, *kehelyből*, *kehelyért*, *kehelyhez*, *kehelyig*, *kehelyként*, *kehellyé*, *kehellyel*, *kehelynek*, *kehelynél*, *kehelyre*, *kehelyről*, *kehelytől*, *kelyhek*, *kelyhekbe*, *kelyhekben*, *kelyhekből*, *kelyheken*, *kelyhekért*, *kelyheket*, *kelyhekhez*, *kelyhekig*, *kelyhekké*, *kelyhekkel*, *kelyhekként*, *kelyheknek*, *kelyheknél*, *kelyhekre*, *kelyhekről*, *kelyhektől*, *kelyhen*, *kelyhet* — "chalice";

pehely, *pehelybe*, *pehelyben*, *pehelyből*, *pehelyért*, *pehelyhez*, *pehelyig*, *pehelyként*, *pehellyé*, *pehellyel*, *pehelynek*, *pehelynél*, *pehelyre*, *pehelyről*, *pehelytől*, *pelyhek*, *pelyhekbe*, *pelyhekben*, *pelyhekből*, *pelyheken*, *pelyhekért*, *pelyheket*, *pelyhekhez*, *pelyhekig*, *pelyhekké*, *pelyhekkel*, *pelyhekként*, *pelyheknek*, *pelyheknél*, *pelyhekre*, *pelyhekről*, *pelyhektől*, *pelyhen*, *pelyhet* — "flake";

teher, *teherbe*, *teherben*, *teherből*, *teherért*, *teherhez*, *teherig*, *teherként*, *tehernek*, *tehernél*, *teherre*, *teherré*, *teherrel*, *teherről*, *tehertől*, *terhek*, *terhekbe*, *terhekben*, *terhekből*, *terheken*, *terhekért*, *terheket*, *terhekhez*, *terhekig*, *terhekké*, *terhekkel*, *terhekként*, *terheknek*, *terheknél*, *terhekre*, *terhekről*, *terhektől*, *terhen*, *terhet* — "load".

They are automatically clustered by our algorithm, as expected.

The declension paradigms of Hungarian adjectives are essentially identical to those of Hungarian nouns. In the following, we only list various adjectives in their positive and comparative forms (in the nominative case, singular number), based on [48, §10.3]. Superlatives are not listed below, as they are consistently formed with a *leg~* prefix to the comparative forms.

bátor, *bátrabb* — "brave";

bő, *bővebb* — "abundant";

derék, *derekabb* — "decent";

drága, *drágább* — "expensive";

édes, *édesebb* — “sweet”;

érthető, *érthetőbb* — “understandable”;

fekete, *feketébb* — “black”;

hosszabb, *hosszú* — “long”;

hű, *hűbb* — “faithful”;

ifjabb, *ifjú* — “young”;

jó, *jobb* — “good”;

keserű, *keserűbb* — “bitter”;

kevés, *kevesebb* — “few”;

kicsi, *kisebb* — “small”;

könnyebb, *könnyű* — “easy”;

különös, *különösebb* — “special”;

lassabb, *lassú* — “slow”;

nehéz, *nehezebb* — “difficult”;

olcsó, *olcsóbb* — “cheap”;

piros, *pirosabb* — “red”;

régi, *régibb* — “old”;

szebb, *szép* — “beautiful”;

szomorú, *szomorúbb* — “sad”;

szörnyebb, *szörnyű* — “awful”.

Our clustering algorithm groups these adjectives correctly.

Example 8.24.2. Our sample of regular Hungarian verbs are chosen from a representative subset of [48, §4.1], so as to cover all the possible regular verb typologies.

cselekedd, *cselekeded*, *cselekedek*, *cselekedem*, *cselekedendő*, *cselekedett*, *cselekedhet*, *cselekedi*, *cselekedik*, *cselekeditek*, *cselekedj*, *cselekedje*, *cselekedjed*, *cselekedjek*, *cselekedjék*, *cselekedjél*, *cselekedjelek*, *cselekedjem*, *cselekedjen*, *cselekedjenek*, *cselekedjetek*, *cselekedjétek*, *cselekedjük*, *cselekedjünk*, *cselekedlek*, *cselekedne*, *cselekedné*, *cselekedned*, *cselekednéd*, *cselekednek*, *cselekednék*, *cselekednél*, *cselekednélek*, *cselekednem*, *cselekedném*, *cselekednének*, *cselekednénk*, *cselekednetek*, *cselekednétek*, *cselekedni*, *cselekednie*, *cselekedniük*, *cselekednünk*, *cselekedő*, *cselekedsz*, *cselekedte*, *cselekedted*, *cselekedtek*, *cselekedték*, *cselekedtél*, *cselekedtelek*, *cselekedtem*, *cselekedtetek*, *cselekedtétek*, *cselekedtük*, *cselekedtünk*, *cselekedünk*, *cselekedve*, *cselekszed*, *cselekszel*, *cselekszem*, *cselekszenek*, *cselekszetek*, *cselekszi*, *cselekszik*, *cselekszitek*, *cselekszünk*, *cselekvő* — “act”;

emlékezek, *emlékezel*, *emlékezett*, *emlékezhet*, *emlékezne*, *emlékezned*, *emlékeznek*, *emlékeznék*, *emlékeznél*, *emlékeznem*, *emlékeznének*, *emlékeznénk*, *emlékeznetek*, *emlékeznétek*, *emlékezni*, *emlékeznie*, *emlékezniük*, *emlékeznünk*, *emlékező*, *emlékeztek*, *emlékeztél*, *emlékeztem*, *emlékeztetek*, *emlékeztünk*, *emlékezünk*, *emlékezve*, *emlékezz*, *emlékezzek*, *emlékezzél*, *emlékezzen*, *emlékezzenek*, *emlékezzetek*, *emlékezzünk*, *emlékszel*, *emlékszem*, *emlékszenek*, *emlékszetek*, *emlékszik*, *emlékszünk* — “remember”;

gyarapodhat, *gyarapodik*, *gyarapodj*, *gyarapodjak*, *gyarapodjál*, *gyarapodjanak*, *gyarapodjatok*, *gyarapodjon*, *gyarapodjunk*, *gyarapodna*, *gyarapodnak*, *gyarapodnál*, *gyarapodnának*, *gyarapodnánk*, *gyarapodnátok*, *gyarapodnék*, *gyarapodni*, *gyarapodnia*, *gyarapodniuk*, *gyarapodnod*, *gyarapodnom*, *gyarapodnotok*, *gyarapodnunk*, *gyarapodó*, *gyarapodok*, *gyarapodott*, *gyarapodsz*, *gyarapodtak*, *gyarapodtál*, *gyarapodtam*, *gyarapodtatok*, *gyarapodtok*, *gyarapodtunk*, *gyarapodunk*, *gyarapodva* — “increase”;

igyekezek, *igyekezel*, *igyekezett*, *igyekezhet*, *igyekezne*, *igyekezned*, *igyekeznek*, *igyekeznék*, *igyekeznél*, *igyekeznem*, *igyekeznének*, *igyekeznénk*, *igyekeznetek*, *igyekeznétek*, *igyekezni*, *igyekeznie*, *igyekezniük*, *igyekeznünk*, *igyekező*, *igyekeztek*, *igyekeztél*, *igyekeztem*, *igyekeztetek*, *igyekeztünk*, *igyekezünk*, *igyekezve*, *igyekezz*, *igyekezzek*, *igyekezzél*, *igyekezzen*, *igyekezzenek*, *igyekezzetek*, *igyekezzünk*, *igyekszel*, *igyekszem*, *igyekszenek*, *igyekszetek*, *igyekszik*, *igyekszünk*, *igyekvő* — "strive";

mosolygandó, *mosolygó*, *mosolygod*, *mosolygok*, *mosolygom*, *mosolygott*, *mosolygunk*, *mosolyog*, *mosolyogd*, *mosolyoghat*, *mosolyogj*, *mosolyogja*, *mosolyogjad*, *mosolyogjak*, *mosolyogják*, *mosolyogjál*, *mosolyogjalak*, *mosolyogjam*, *mosolyogjanak*, *mosolyogjatok*, *mosolyogjátok*, *mosolyogjon*, *mosolyogjuk*, *mosolyogjunk*, *mosolyoglak*, *mosolyogna*, *mosolyogná*, *mosolyognád*, *mosolyognak*, *mosolyognák*, *mosolyognál*, *mosolyognálak*, *mosolyognám*, *mosolyognának*, *mosolyognánk*, *mosolyognátok*, *mosolyognék*, *mosolyogni*, *mosolyognia*, *mosolyogniuk*, *mosolyognod*, *mosolyognom*, *mosolyognotok*, *mosolyognunk*, *mosolyogsz*, *mosolyogta*, *mosolyogtad*, *mosolyogtak*, *mosolyogták*, *mosolyogtál*, *mosolyogtalak*, *mosolyogtam*, *mosolyogtatok*, *mosolyogtátok*, *mosolyogtok*, *mosolyogtuk*, *mosolyogtunk*, *mosolyogva* — "smile";

sző, *sződd*, *szőj*, *szője*, *szőjed*, *szőjek*, *szőjék*, *szőjél*, *szőjelek*, *szőjem*, *szőjenek*, *szőjetek*, *szőjétek*, *szőjön*, *szőjük*, *szőjünk*, *szőlek*, *szőne*, *szőné*, *szőnéd*, *szőnek*, *szőnék*, *szőnél*, *szőnélek*, *szőném*, *szőnének*, *szőnénk*, *szőnétek*, *szőni*, *szőnie*, *szőniük*, *szőnöd*, *szőnöm*, *szőnötök*, *szőnünk*, *szősz*, *szőtt*, *szőtte*, *szőtted*, *szőttek*, *szőtték*, *szőttél*, *szőttelek*, *szőttem*, *szőttetek*, *szőttétek*, *szőttök*, *szőttük*, *szőttünk*, *szőve*, *szövendő*, *szövi*, *szövik*, *szövitek*, *szövő*, *szövöd*, *szövök*, *szövöm*, *szövünk* — "weave";

utazhat, *utazik*, *utazna*, *utaznak*, *utaznál*, *utaznának*, *utaznánk*, *utaznátok*, *utaznék*, *utazni*, *utaznia*, *utazniuk*, *utaznod*, *utaznom*, *utaznotok*, *utaznunk*, *utazó*, *utazok*, *utazol*, *utazom*, *utazott*, *utaztak*, *utaztál*, *utaztam*, *utaztatok*, *utaztok*, *utaztunk*, *utazunk*, *utazva*, *utazz*, *utazzak*, *utazzál*, *utazzanak*, *utazzatok*, *utazzon*, *utazzunk* — "travel".

These verbs are clustered correctly by our algorithm.
Our selection of irregular Hungarian verbs are based on [48, Appendix 1].

edd, *egye*, *egyed*, *egyek*, *egyék*, *egyél*, *egyelek*, *egyem*, *egyen*, *egyenek*, *egyetek*, *egyétek*, *együk*, *együnk*, *ehet*, *enne*, *enné*, *enned*, *ennéd*, *ennék*, *ennél*, *ennélek*, *ennem*, *enném*, *ennének*, *ennénk*, *ennetek*, *ennétek*, *enni*, *ennie*, *enniük*, *ennünk*, *eszed*, *eszek*, *eszel*, *eszem*, *eszi*, *eszik*, *eszitek*, *eszlek*, *esznek*, *esszük*, *esztek*, *eszünk*, *ette*, *etted*, *ettek*, *ették*, *ettél*, *ettelek*, *ettem*, *ettetek*, *ettétek*, *ettük*, *ettünk*, *evett*, *evő* — "eat";

gyere, *gyertek*, *gyerünk*, *jöhet*, *jöjj*, *jöjjek*, *jöjjél*, *jöjjenek*, *jöjjetek*, *jöjjön*, *jöjjünk*, *jön*, *jönne*, *jönnek*, *jönnék*, *jönnél*, *jönnének*, *jönnénk*, *jönnétek*, *jönni*, *jönnie*, *jönniük*, *jönnöd*, *jönnöm*, *jönnötök*, *jönnünk*, *jössz*, *jösztök*, *jött*, *jöttek*, *jöttél*, *jöttem*, *jöttetek*, *jöttök*, *jöttünk*, *jőve*, *jövendő*, *jövő*, *jövök*, *jövünk* — "come";

hidd, *hiendő*, *higgy*, *higgye*, *higgyed*, *higgyek*, *higgyék*, *higgyél*, *higgyelek*, *higgyem*, *higgyen*, *higgyenek*, *higgyetek*, *higgyétek*, *higgyük*, *higgyünk*, *hihet*, *hinne*, *hinné*, *hinned*, *hinnéd*, *hinnék*, *hinnél*, *hinnélek*, *hinnem*, *hinném*, *hinnének*, *hinnénk*, *hinnetek*, *hinnétek*, *hinni*, *hinnie*, *hinniük*, *hinnünk*, *hisz*, *hiszed*, *hiszek*, *hiszel*, *hiszem*, *hiszi*, *hiszik*, *hiszitek*, *hiszlek*, *hisznek*, *hisszük*, *hisztek*, *hiszünk*, *hitt*, *hitte*, *hitted*, *hittek*, *hitték*, *hittél*, *hittelek*, *hittem*, *hittetek*, *hittétek*, *hittük*, *hittünk* — "believe";

idd, *igya*, *igyad*, *igyak*, *igyák*, *igyál*, *igyalak*, *igyam*, *igyanak*, *igyatok*, *igyátok*, *igyon*, *igyuk*, *igyunk*, *ihat*, *inna*, *inná*, *innád*, *innák*, *innál*, *innálak*, *innám*, *innának*, *innánk*, *innátok*, *innék*, *inni*, *innia*, *inniuk*, *innod*, *innom*, *innotok*, *innunk*, *iszik*, *iszlak*, *isznak*, *iszod*, *iszok*, *iszol*, *iszom*, *issza*, *isszák*, *isszátok*, *isszuk*, *isztok*, *iszunk*, *itta*, *ittad*, *ittak*, *itták*, *ittál*, *ittalak*, *ittam*, *ittatok*, *ittátok*, *ittuk*, *ittunk*, *ivandó*, *ivó*, *ivott* — "drink";

megy, *megyek*, *megyünk*, *mehet*, *menj*, *menjek*, *menjél*, *menjen*, *menjenek*, *menjetek*, *menjünk*, *menne*, *menned*, *mennek*, *mennék*, *mennél*, *mennem*, *mennének*, *mennénk*, *mennetek*, *mennétek*, *menni*, *mennie*, *menniük*, *mennünk*, *menő*, *ment*, *mentek*, *mentél*, *mentem*, *mentetek*, *mentünk*, *menve* — "go";

vedd, *veendő*, *végy*, *vegye*, *vegyed*, *vegyek*, *vegyék*, *vegyél*, *vegyelek*, *vegyem*, *vegyen*, *vegyenek*, *vegyetek*, *vegyétek*, *vegyük*, *vegyünk*, *vehet*, *venne*, *venné*, *venned*, *vennéd*, *vennék*, *vennél*, *vennélek*, *vennem*, *venném*, *vennének*, *vennénk*, *vennetek*, *vennétek*, *venni*, *vennie*, *venniük*, *vennünk*, *vesz*, *veszed*, *veszek*, *veszel*, *veszem*, *veszi*, *veszik*, *veszitek*, *veszlek*, *vesznek*, *vesszük*, *vesztek*, *veszünk*, *vett*, *vette*, *vetted*, *vettek*, *vették*, *vettél*, *vettelek*, *vettem*, *vettetek*, *vettétek*, *vettük*, *vettünk*, *véve*, *vevő* — "buy";

vidd, *viendő*, *vigye*, *vigyed*, *vigyek*, *vigyék*, *vigyél*, *vigyelek*, *vigyem*, *vigyen*, *vigyenek*, *vigyetek*, *vigyétek*, *vigyük*, *vigyünk*, *vihet*, *vinne*, *vinné*, *vinned*, *vinnéd*, *vinnék*, *vinnél*, *vinnélek*, *vinnem*, *vinném*, *vinnének*, *vinnénk*, *vinnetek*, *vinnétek*, *vinni*, *vinnie*, *vinniük*, *vinnünk*, *visz*, *viszed*, *viszek*, *viszel*, *viszem*, *viszi*, *viszik*, *viszitek*, *viszlek*, *visznek*, *visszük*, *visztek*, *viszünk*, *vitt*, *vitte*, *vitted*, *vittek*, *vitték*, *vittél*, *vittelek*, *vittem*, *vittetek*, *vittétek*, *vittük*, *vittünk*, *vive*, *vivő* — "carry".

Sending them to our algorithm, we receive the correct clustering results.

8.2.3 Heuristic detection of compounds

The following algorithm for heuristic detection of Hungarian compounds differs from the Danish version (Algorithm 5.12) only in some specific details. To make the context clear, we still state the algorithm in full. (In what follows, the string minus operation $\hat{\beta} \ominus \hat{\alpha}$ is prescribed by Definition 5.11.)

Algorithm 8.25 (Heuristic identification of Hungarian binary compounds)**.** *Let $\Lambda^{\hat{\rho}} = \{\hat{\rho}_1, \ldots, \hat{\rho}_Q\}$ be a list of distinct Hungarian essential roots (without vowel blotting) that are at two characters long, contain at least one instance of* **V** *in the first two letters, and DO NOT match either*

$$(ag|ar|az|ba|bo|cel|či|de|dö|dul|em|ha|he|hi|ig|is|ka|ke|ko|kö|li|mas|mi|mu|ne|od|ős|rá|re|ta|te|tor|tü|va|vá|vas)$$

or

$$(\beta\kappa a|(a|e|o|ö)(s|z)|al|am|di|eg|er|ér|fe|fog|hi|kor|mag|meg|ol|or|ot|š\mathbf{V}|ter|ve)\mathbf{X}.$$

The output of the function $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is obtained through the following procedures:*

(1) *Construct a list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$ *where* $\lambda_q^{\hat{\rho}} = \{\hat{\rho}_{(q,1)}, \ldots, \hat{\rho}_{(q,n_q)}\}$ *is a subset of* $\Lambda^{\hat{\rho}}$ *whose members all match the string pattern* $\hat{\rho}_q\sim$*, for* $q \in \mathbb{Z} \cap [1, Q]$*.*

(2) *Expand the aforementioned entry* $(\hat{\rho}_q, \lambda_q^{\hat{\rho}})$ *into a list of triplets* $\{(\hat{\rho}_{(q,1)}, \hat{\rho}_q, \hat{\rho}_{(q,1)} \ominus \hat{\rho}_q), \ldots, (\hat{\rho}_{(q,n_q)}, \hat{\rho}_q, \hat{\rho}_{(q,n_q)} \ominus \hat{\rho}_q)\}$ *for every* $q \in \mathbb{Z} \cap [1, Q]$ *such that* $\lambda_q^{\hat{\rho}} \neq \varnothing$*. Collect all these triplets as one runs through the list* $\{(\hat{\rho}_1, \lambda_1^{\hat{\rho}}), \ldots, (\hat{\rho}_Q, \lambda_Q^{\hat{\rho}})\}$*. The list of these triplets* $\{(\hat{\rho}_{\langle 1\rangle}, \hat{\eta}_{\langle 1\rangle}, \hat{\rho}_{\langle 1\rangle} \ominus \hat{\eta}_{\langle 1\rangle}), \ldots, (\hat{\rho}_{\langle Q'\rangle}, \hat{\eta}_{\langle Q'\rangle}, \hat{\rho}_{\langle Q'\rangle} \ominus \hat{\eta}_{\langle Q'\rangle})\}$ *contains potentially valid decompositions of compounds.*

(3) *Screen the aforementioned list of triplets as follows: for every* $q' \in \mathbb{Z} \cap [1, Q']$*, if* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \hat{\tau}_{\langle q'\rangle} = \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle})$ *satisfies*

$$\ell(\hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle}) \geq 2 \quad \textbf{AND} \quad \hat{\rho}_{\langle q'\rangle} \ominus \hat{\eta}_{\langle q'\rangle} = \mathbf{X}_1\mathbf{V}\mathbf{X}_2,$$

then construct $\hat{\tau}^*_{\langle q'\rangle}$ *by performing* $(\varnothing|a|e)(\varnothing|i|ñi|s|t|ú|ű)\sim\!\!\rightarrow \varnothing$ *and*

$$(\lambda l(\varnothing|e|ő)|a|á|að|ala|be|f(\varnothing|e|el|él)|h(á|eł|oz)|id|k(ere|i|ör|öz)|le(\varnothing|n)|meg|neki|od|öš|ra|rá|šé|šer|to|túl|u|vég|vin)\sim\!\!\rightarrow \varnothing$$

on $\hat{\tau}_{\langle q'\rangle}$*, before generating a list* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *by members of* $\Lambda^{\hat{\rho}}$ *that match the pattern* $\underline{(\hat{\tau}_{\langle q'\rangle}|\hat{\tau}^*_{\langle q'\rangle})}$*; otherwise, set* $\lambda^{\hat{\tau}}_{\langle q'\rangle} = \varnothing$*.*

(4) *Collect all the triplets* $(\hat{\rho}_{\langle q'\rangle}, \hat{\eta}_{\langle q'\rangle}, \lambda^{\hat{\tau}}_{\langle q'\rangle})$ *where* $\lambda^{\hat{\tau}}_{\langle q'\rangle}$ *is non-void and* $\hat{\tau}_{\langle q'\rangle}$ *DOES NOT match* $z\tau\mathbf{X}$*. This list of triplets* $\mathrm{CpdDet}(\Lambda_{\hat{\rho}})$ *contains the heuristic decompositions of all the identified binary compounds.*

Algorithm 8.26 (Approximate clustering of Hungarian words with heuristic detection of compounds)**.** *The approximate clustering of a list of Hungarian words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *respecting compounding is completed in four stages:*

(1) *Do as in Algorithm 8.24(1).*

(2) *Do as in Algorithm 8.24(2). Save both the tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$ *and the list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *for further use.*

(3) *Construct a tagged list of word clusters* $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$ *where* $\Lambda_k^{\hat{\rho}}$ *is the union of all Hungarian essential roots (without vowel blotting) available to* Γ_k*, for* $k \in \mathbb{Z} \cap [1, K]$*. Set* $\Lambda^{\hat{\rho}} = \Lambda_1^{\hat{\rho}} \cup \cdots \cup \Lambda_K^{\hat{\rho}}$*, and evaluate* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$*.*

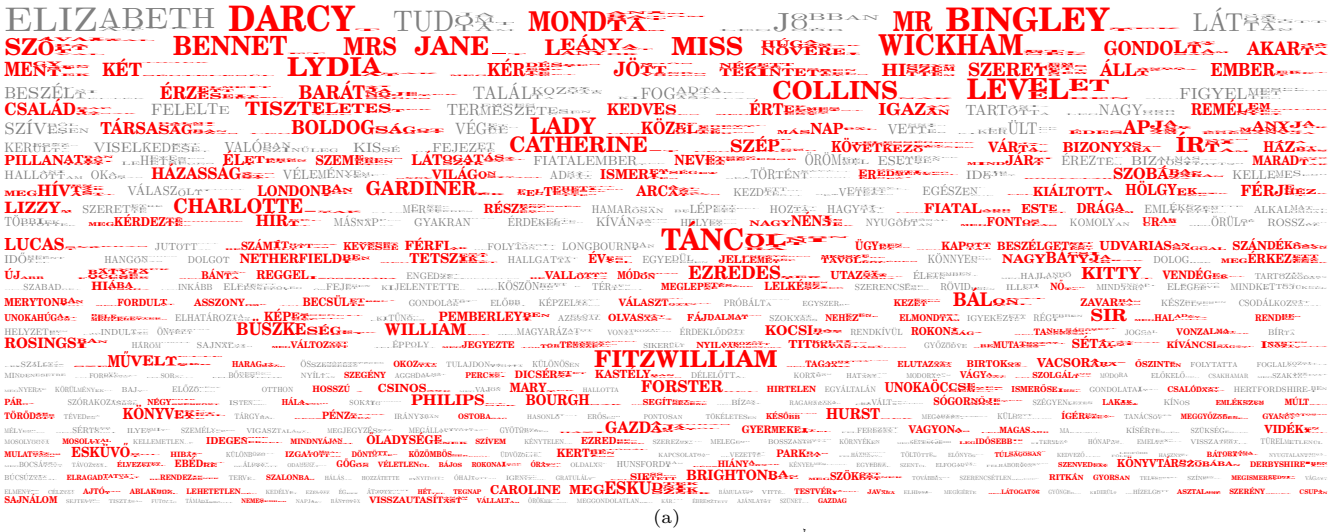
(4) *The first component* $\hat{\rho}_{\langle q''\rangle}$ *of each triplet* $(\hat{\rho}_{\langle q''\rangle}, \hat{\eta}_{\langle q''\rangle}, \lambda^{\hat{\tau}}_{\langle q''\rangle})$ *in* $\mathrm{CpdDet}(\Lambda^{\hat{\rho}})$ *is called a "dissolvable compound", the second component* $\hat{\eta}_{\langle q''\rangle}$ *a "heuristic head", and the first member in the third component* $\lambda^{\hat{\tau}}_{\langle q''\rangle}$ *a "heuristic tail". In the tagged list* $\{(\check{\Gamma}_1, \Lambda_1^{\hat{\rho}}), \ldots, (\check{\Gamma}_K, \Lambda_K^{\hat{\rho}})\}$*, every entry containing a "dissolvable compound" is removed, and regrouped with the entries matching its "heuristic head" and "heuristic tail". Finally, remove all tags.*

Example 8.26.1. In Fig. S13, we apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text sources).

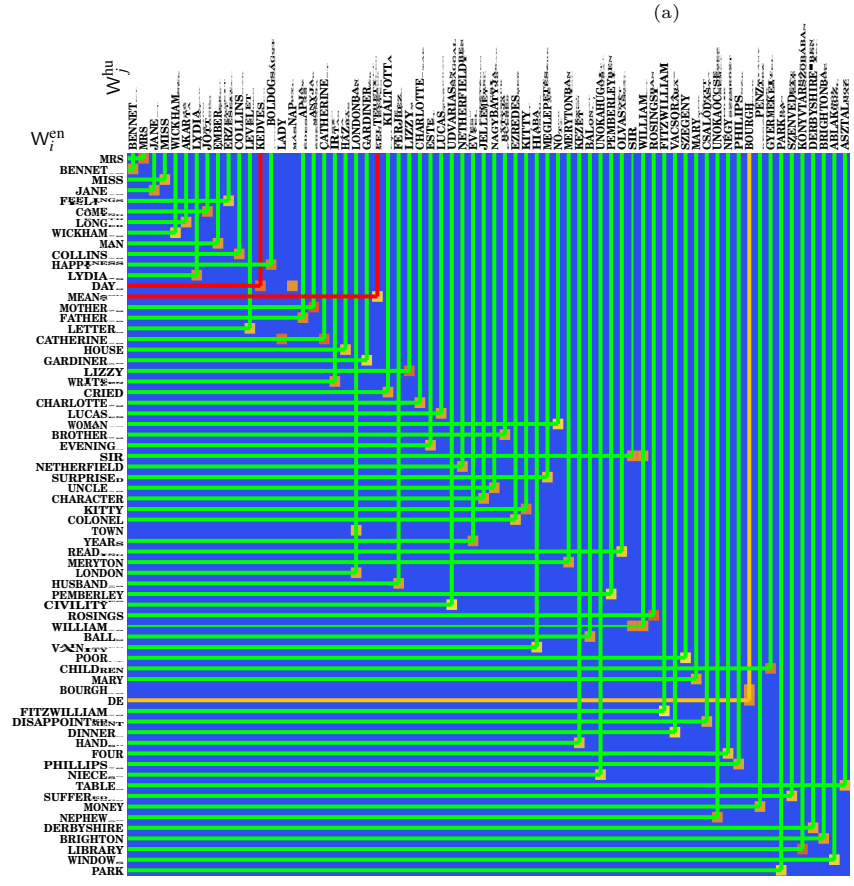
According to Hungarian orthography, vowel lengths (*a* vs. *á*, *ö* vs. *ő* etc.) are ignored during alphabetization. If implemented as is, stacked vowels in word clusters may become illegible at times. As a compromise, we always sort long vowels after short vowels in Fig. S13.

In Hungarian, *de* "but" is a stop word, so it is understandable that our algorithm does not provide a perfect translation of the proper name *de Bourgh* in Fig. S13b.

In Darwin's *Origin of Species*, the word "reciprocal" is mostly used in the phrase *reciprocal cross*. Therefore, we consider "reciprocal" and "graft" near synonyms in Fig. S13b″. Furthermore, in Fig. S13b″, sometimes and English word may correspond to a Hungarian phrase or vice versa: we consider English *fresh* an exact match to Hungarian *édesvizi* "fresh water" (literally "sweet water"), and similarly for English *rock-pigeon* and Hungarian *szirti* (the latter being part of the phrase *szirti galamb* that translates *rock-pigeon* into Hungarian).



(a)



(b)

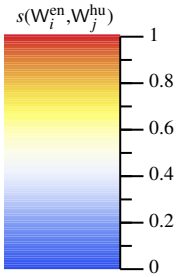


Fig. S13. Text mining in Hungarian. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Hungarian version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{hu})$ between selected topics in English and Hungarian versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.

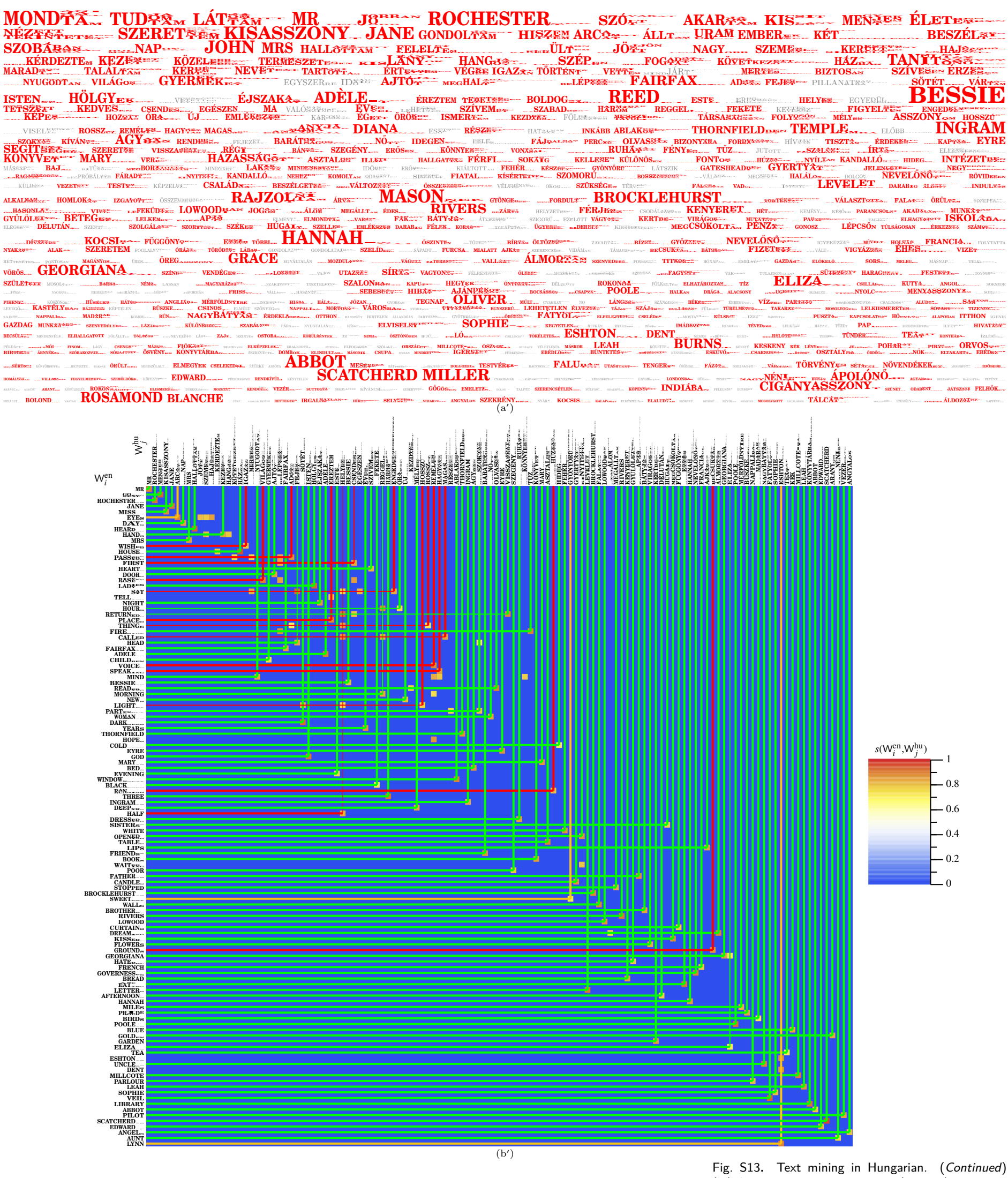


Fig. S13. Text mining in Hungarian. (*Continued*) (a′) Statistically identified topics ($n_{ii} \geq 20$) in a Hungarian version of *Jane Eyre*, with the same color encoding scheme as Fig. S3. (b′) Semantic similarities $s(W_i^{en}, W_j^{hu})$ between selected topics in English and Hungarian versions of *Jane Eyre*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c′) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

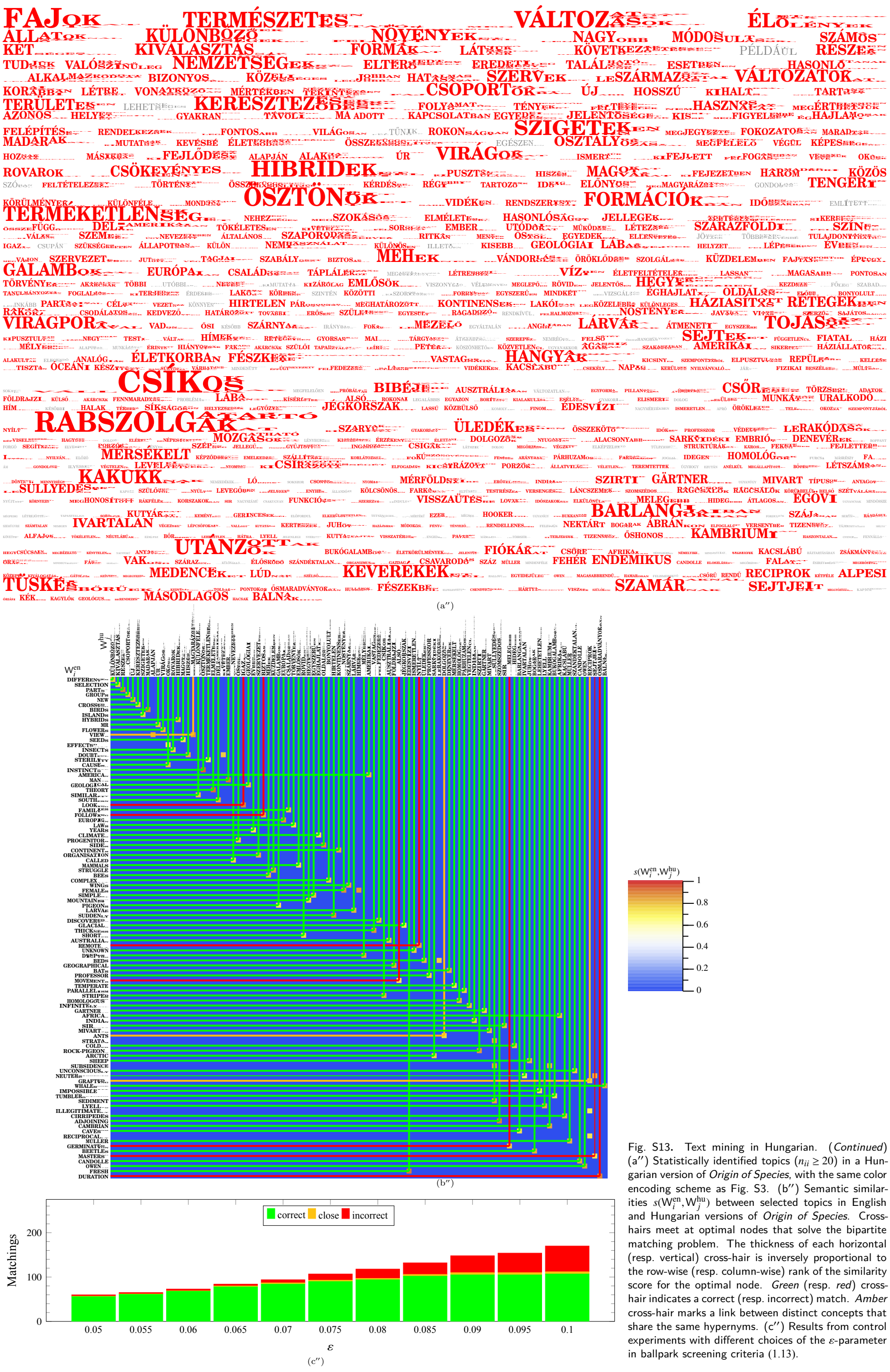


Fig. S13. Text mining in Hungarian. (*Continued*) (a″) Statistically identified topics ($n_{ii} \geq 20$) in a Hungarian version of *Origin of Species*, with the same color encoding scheme as Fig. S3. (b″) Semantic similarities $s(\mathrm{W}_i^{\mathrm{en}}, \mathrm{W}_j^{\mathrm{hu}})$ between selected topics in English and Hungarian versions of *Origin of Species*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms. (c″) Results from control experiments with different choices of the ε-parameter in ballpark screening criteria (1.13).

9 Approximate word clustering in various agglutinative languages

Basque, Korean and Turkish are three agglutinative SOV languages with different linguistic ancestors,[104] but they do share a few typological and morphological features:

- All these languages are rich in suffixes, but have almost no prefixes. A typical word in these languages assumes the form of

- Typically, the word stem is one or two syllables long.[105] However, due to strong and persisted foreign influences on these languages (French/Spanish on Basque, Chinese/English/Japanese on Korean, Arabic/Persian on Turkish), there are certain exceptions to this estimate. Sometimes, stems of loanwords may be confused with native suffixes, if not handled properly.
- There are hardly any irregular verbs or nouns. Unlike Finnish, the suffixes in these languages (usually) do not cause sound changes upon interactions.
- These languages have more sophisticated systems for kinship terms than typical Indo-European languages. There could be different words for a sibling of a given gender, depending on seniority (Korean) or the gender of the possessor (Basque and Korean); there could be different words for extended family members depending on maternal/paternal sides (Korean and Turkish). Our word clustering algorithms will accommodate to such cultural differences, as much as possible.

[104] Some Altaicists believe that both Korean and Turkish descend from the same Proto-Altaic language [49].

[105] Since we do not know, *a priori*, whether the second syllable belongs to the stem or the suffix, a non-trivial stemming algorithm is still necessary.

9.1 Modified Porter stemming algorithm for Basque

Our Basque stemming algorithm is adapted from the method of Mikel Otxandorena (`http://snowball.tartarus.org/algorithms/basque/stemmer.html`). Our modest contributions include the following aspects:

- We add special treatments for the only irregular adjective in Basque: *on* "good" [50, §3.1.3.3].
- We add special routines to cluster finite forms[106] of Basque verbs [50, §3.5.2.1].
- We carefully handle loanwords that begin with *des~*, *err~*, *irr~* and so on.
- We carefully treat the negative suffixes *~gabe* and *~gabea*, which carry similar functions as the English prefix *un~* and suffix *~less*.
- We group Basque kinship terms according to (roughly) their English equivalents.

Basque stop words can carry various agglutinative suffixes. Instead of enumerating all of them in a single list, we will define several string patterns that match the criterion for stop words.

Definition 9.1 (Basque stop words)**.** Define the pattern **BasqueStopRoot** as any one from the following list:

> *aitzin*, *albo*, *alde*, *arte*, *asko*, *atze*, *aurre*, *azken*, *azpi*, *barru*, *batega*, *batek*, *baten*, *batera*, *batet*, *behe*, *beza*, *buruzko*, *dela*, *denaren*, *denet*, *ditu*, *dizu*, *eduki*, *egote*, *erdi*, *ere*, *gain*, *garen*, *genieza*, *genitu*, *genu*, *geure*, *gibel*, *hain*, *heure*, *hieza*, *hitu*, *huen*, *inguru*, *izan*, *izate*, *lizki*, *neure*, *nieza*, *nitu*, *nuen*, *ondo*, *zela*, *zenieza*, *zenitu*, *zenu*, *zeuen*, *zeure*, *zieza*, *zire*, *zitu*, *zizki*, *zuen*.

Define the pattern **BasqueStopString** as any one from the following list:

> *abar*, *agertutako*, *agian*, *ahal*, *ahaztuak*, *al*, *ala*, *alde*, *aldera*, *aplikatuta*, *arabera*, *ari*, *arren*, *arte*, *artean*, *aski*, *asko*, *askotan*, *atzean*, *atzera*, *aurka*, *aurrean*, *aurrera*, *aurrerantzean*, *aurretik*, *azpian*, *ba*, *bada*, *bagengozkie*, *bagengozkik*, *bagengozkin*, *bagengozkio*, *bagengozkizu*, *bagengozkizue*, *bagenie*, *bagenik*, *bagenin*, *bagenio*, *bagenizu*, *bagenizue*, *bagenkizkie*, *bagenkizkik*, *bagenkizkio*, *bagenkizkizu*, *bagenkizkizue*, *bageunde*, *bagina*, *bagintez*, *bagintzaizkie*, *bagintzaizkik*, *bagintzaizkin*, *bagintzaizkio*, *bagintzaizkizu*, *bagintzaizkizue*, *bahendi*, *bahengo*, *bahengokie*, *bahengokigu*, *bahengokio*, *bahengokit*, *bahenkie*, *bahenkigu*, *bahenkio*, *bahenkit*, *bahie*, *bahigu*, *bahintz*, *bahintzaie*, *bahintzaigu*, *bahintzaio*, *bahintzait*, *bahio*, *bahit*, *bai*, *baina*, *baino*, *baitzen*, *baizik*, *bakarrik*, *bakoitza*, *baldin*, *baledi*, *balego*, *balegokie*, *balegokigu*, *balegokik*, *balegokin*, *balegokio*, *balegokit*, *balegokizu*, *balegokizue*, *balegozkie*, *balegozkigu*, *balegozkik*, *balegozkin*, *balegozkio*, *balegozkit*, *balegozkizu*, *balegozkizue*, *balekie*, *balekigu*, *balekik*, *balekio*, *balekit*, *balekizkie*, *balekizkigu*, *balekizkik*, *balekizkio*, *balekizkit*, *balekizkizu*, *balekizkizue*, *balekizu*, *balekizue*, *baleude*, *baliabideak*, *baliate*, *balidate*, *balie*, *baliete*, *baligu*, *baligute*, *balik*, *balin*, *balinate*, *balio*, *baliote*, *balira*, *balit*, *balitez*, *balitz*, *balitzaie*, *balitzaigu*, *balitzaik*, *balitzain*, *balitzaio*, *balitzait*, *balitzaizkie*, *balitzaizkigu*, *balitzaizkik*, *balitzaizkin*, *balitzaizkio*, *balitzaizkit*, *balitzaizkizu*, *balitzaizkizue*, *balitzaizu*, *balitzaizue*, *balizu*, *balizue*, *balizuete*, *balizute*, *banendi*, *banengo*, *banengokie*, *banengokik*, *banengokin*, *banengokio*, *banengokizu*, *banengokizue*, *banenkie*, *banenkik*, *banenkio*, *banenkizu*, *banenkizue*, *banie*, *banik*, *banin*, *banintz*, *banintzaie*, *banintzaik*, *banintzain*, *banintzaio*, *banintzaizu*, *banintzaizue*, *banio*, *banizu*, *banizue*, *barruan*, *bat*, *batean*, *batera*, *batez*, *batzuetan*, *bazengozkidate*, *bazengozkie*, *bazengozkiete*, *bazengozkigu*, *bazengozkigute*, *bazengozkio*, *bazengozkiote*, *bazengozkit*, *bazenidate*, *bazenie*, *bazeniete*, *bazenigu*, *bazenigute*, *bazenio*, *bazeniote*, *bazenit*, *bazenkizkidate*, *bazenkizkie*, *bazenkizkiete*, *bazenkizkigu*, *bazenkizkigute*, *bazenkizkio*, *bazenkizkiote*, *bazenkizkit*, *bazeunde*, *bazeundete*, *bazina*, *bazinete*, *bazintez*, *bazintezte*, *bazintzaizkidate*, *bazintzaizkie*, *bazintzaizkiete*, *bazintzaizkigu*, *bazintzaizkigute*, *bazintzaizkio*, *bazintzaizkiote*, *bazintzaizkit*, *bedi*, *begira*, *bego*, *begozkie*, *begozkigu*, *begozkik*, *begozkin*, *begozkio*, *begozkit*, *begozkizu*, *begozkizue*, *behar*, *beharbada*, *beharko*, *beharra*, *beharrean*, *behean*, *behera*, *behin*, *bengokie*, *bengokigu*, *bengokik*, *bengokin*, *bengokio*, *bengokit*, *bengokizu*, *bengokizue*, *beraz*, *berea*, *berean*, *berez*, *berri*, *berriro*, *berriz*, *bertan*, *bestalde*, *beste*, *bestela*, *besterik*, *betetzera*, *beti*, *beude*, *bezala*, *biak*, *bidean*, *bidez*, *bihurtu*, *bihurtuz*, *bihurtzen*, *bila*, *biok*, *bitartean*, *bitez*, *borondatea*, *burua*, *buruz*, *da*, *dadila*, *dadin*, *dago*, *dagoela*, *dagoeneko*, *dagoke*, *dagokiake*, *dagokidake*, *dagokie*, *dagokieke*, *dagokigu*, *dagokiguke*, *dagokik*, *dagokin*, *dagokinake*, *dagokio*, *dagokioke*, *dagokit*, *dagokizu*, *dagokizue*, *dagokizueke*, *dagokizuke*, *dagozkiake*, *dagozkidake*, *dagozkie*, *dagozkieke*, *dagozkigu*, *dagozkiguke*, *dagozkik*, *dagozkin*, *dagozkinake*, *dagozkio*, *dagozkioke*, *dagozkit*, *dagozkizu*, *dagozkizue*, *dagozkizueke*, *dagozkizuke*, *daiteke*, *daitez*, *daitezela*, *daitezke*, *dakiake*, *dakian*, *dakidake*, *dakidan*, *dakieke*, *dakien*, *dakiguke*, *dakigun*, *dakioke*, *dakion*, *dakizkiake*, *dakizkian*, *dakizkidake*, *dakizkidan*, *dakizkieke*, *dakizkien*, *dakizkiguke*, *dakizkigun*, *dakizkioke*, *dakizkion*, *dakizkizueke*, *dakizkizuen*, *dakizkizuke*, *dakizkizun*,

[106] All but a handful verbs (like "come", "go", "say" and "walk") in Basque can only exist in non-finite forms, which do not involve personal prefixes.

dakizueke, *dakizuen*, *dakizuke*, *dakizun*, *dateke*, *daude*, *daudeke*, *daudela*, *Dekretu*, *delako*, *den*, *dena*, *denak*, *denean*, *denek*, *denok*, *dezagun*, *dira*, *direteke*, *dire*, *direla*, *diren*, *direnak*, *dituk*, *ditun*, *dizkio*, *du*, *dua*, *duela*, *dugu*, *dugun*, *duk*, *dun*, *dut*, *dute*, *dutela*, *duzu*, *duzue*, *duzuela*, *duzun*, *ea*, *edo*, *egin*, *egingo*, *egiten*, *egon*, *egongo*, *egote*, *egotea*, *egoten*, *elkarrekin*, *eman*, *emana*, *ematen*, *erdian*, *eske*, *esker*, *eskuratzerakoan*, *eta*, *euren*, *ez*, *ezean*, *ezen*, *ezik*, *ezin*, *ezinbestekoa*, *ezta*, *gabe*, *gagozkiake*, *gagozkian*, *gagozkie*, *gagozkieke*, *gagozkien*, *gagozkik*, *gagozkin*, *gagozkinake*, *gagozkinan*, *gagozkio*, *gagozkioke*, *gagozkion*, *gagozkizu*, *gagozkizue*, *gagozkizueke*, *gagozkizuen*, *gagozkizuke*, *gagozkizun*, *gainean*, *gainera*, *gaitez*, *gaitezke*, *gaituk*, *gaitun*, *gakizkiake*, *gakizkian*, *gakizkieke*, *gakizkien*, *gakizkioke*, *gakizkion*, *gakizkizueke*, *gakizkizuen*, *gakizkizuke*, *gakizkizun*, *gara*, *garateke*, *gatzaizkiake*, *gatzaizkie*, *gatzaizkieke*, *gatzaizkik*, *gatzaizkin*, *gatzaizkinake*, *gatzaizkio*, *gatzaizkioke*, *gatzaizkizu*, *gatzaizkizue*, *gatzaizkizueke*, *gatzaizkizuke*, *gaude*, *gaudeke*, *gauden*, *gehiago*, *gehiegi*, *gehien*, *gengozkiake*, *gengozkian*, *gengozkieke*, *gengozkiekeen*, *gengozkien*, *gengozkinake*, *gengozkinan*, *gengozkioke*, *gengozkiokeen*, *gengozkion*, *gengozkizueke*, *gengozkizuekeen*, *gengozkizuen*, *gengozkizuke*, *gengozkizukeen*, *gengozkizun*, *genituzke*, *genituzkeen*, *genizue*, *genkizkiake*, *genkizkian*, *genkizkieke*, *genkizkiekeen*, *genkizkien*, *genkizkioke*, *genkizkiokeen*, *genkizkion*, *genkizkizueke*, *genkizkizuekeen*, *genkizkizuen*, *genkizkizuke*, *genkizkizukeen*, *genkizkizun*, *genuke*, *genukeen*, *gero*, *gerok*, *geroztik*, *geu*, *geundeke*, *geundekeen*, *geunden*, *geure*, *ginateke*, *ginatekeen*, *ginduan*, *gindunan*, *ginen*, *gintezen*, *gintezke*, *gintezkeen*, *gintzaizkiake*, *gintzaizkian*, *gintzaizkieke*, *gintzaizkiekeen*, *gintzaizkien*, *gintzaizkinake*, *gintzaizkinan*, *gintzaizkioke*, *gintzaizkiokeen*, *gintzaizkion*, *gintzaizkizueke*, *gintzaizkizuekeen*, *gintzaizkizuen*, *gintzaizkizuke*, *gintzaizkizukeen*, *gintzaizkizun*, *gisa*, *guhaur*, *gurea*, *gurekin*, *gutxi*, *gutxiago*, *gutxienez*, *guztiak*, *guztion*, *guztiontzat*, *hadin*, *hago*, *hagoke*, *hagokidake*, *hagokie*, *hagokieke*, *hagokigu*, *hagokiguke*, *hagokio*, *hagokioke*, *hagokit*, *hainbeste*, *haiteke*, *haiz*, *haizateke*, *hakidake*, *hakidan*, *hakieke*, *hakien*, *hakiguke*, *hakigun*, *hakioke*, *hakion*, *haraino*, *harantz*, *haratago*, *hargatik*, *hartara*, *hatzaidake*, *hatzaie*, *hatzaieke*, *hatzaigu*, *hatzaiguke*, *hatzaio*, *hatzaioke*, *hatzait*, *hauekin*, *hauen*, *hendin*, *hengoen*, *hengoke*, *hengokeen*, *hengokidake*, *hengokidakeen*, *hengokidan*, *hengokieke*, *hengokiekeen*, *hengokien*, *hengokiguke*, *hengokigukeen*, *hengokigun*, *hengokioke*, *hengokiokeen*, *hengokion*, *henkidake*, *henkidakeen*, *henkidan*, *henkieke*, *henkiekeen*, *henkien*, *henkiguke*, *henkigukeen*, *henkigun*, *henkioke*, *henkiokeen*, *henkion*, *herori*, *heu*, *heure*, *hihaur*, *hinteke*, *hintekeen*, *hintzaidake*, *hintzaidakeen*, *hintzaidan*, *hintzaieke*, *hintzaiekeen*, *hintzaien*, *hintzaiguke*, *hintzaigukeen*, *hintzaigun*, *hintzaioke*, *hintzaiokeen*, *hintzaion*, *hintzateke*, *hintzatekeen*, *hintzen*, *hituzke*, *hituzkeen*, *hola*, *honaino*, *honantz*, *honen*, *horraino*, *horrantz*, *horrek*, *horren*, *hortxe*, *huke*, *hukeen*, *hurrengo*, *ia*, *inguru*, *inguruan*, *inolako*, *inoren*, *izan*, *izanen*, *izango*, *izate*, *izatea*, *izaten*, *kanpo*, *kanpoan*, *kontra*, *landa*, *laster*, *legoke*, *legokiake*, *legokidake*, *legokieke*, *legokiguke*, *legokinake*, *legokioke*, *legokizueke*, *legokizuke*, *legozkiake*, *legozkidake*, *legozkieke*, *legozkiguke*, *legozkinake*, *legozkioke*, *legozkizueke*, *legozkizuke*, *lehenago*, *lekiake*, *lekidake*, *lekieke*, *lekiguke*, *lekioke*, *lekizkiake*, *lekizkidake*, *lekizkieke*, *lekizkiguke*, *lekizkioke*, *lekizkizueke*, *lekizkizuke*, *lekizueke*, *lekizuke*, *leudeke*, *lirateke*, *liteke*, *litezke*, *lituzke*, *lituzkeen*, *litzaiake*, *litzaidake*, *litzaieke*, *litzaiguke*, *litzainake*, *litzaioke*, *litzaizkiake*, *litzaizkidake*, *litzaizkieke*, *litzaizkiguke*, *litzaizkinake*, *litzaizkioke*, *litzaizkizueke*, *litzaizkizuke*, *litzaizueke*, *litzaizuke*, *litzateke*, *lortu*, *lortzen*, *luke*, *lukeen*, *maiatza*, *maiz*, *mundu*, *nadin*, *nago*, *nagoen*, *nagoke*, *nagokiake*, *nagokian*, *nagokie*, *nagokieke*, *nagokien*, *nagokik*, *nagokin*, *nagokinake*, *nagokinan*, *nagokio*, *nagokioke*, *nagokion*, *nagokizu*, *nagokizue*, *nagokizueke*, *nagokizuen*, *nagokizuke*, *nagokizun*, *nahiko*, *nahikoa*, *nahiz*, *naiteke*, *naiz*, *naizateke*, *nake*, *nakiake*, *nakian*, *nakieke*, *nakien*, *nakioke*, *nakion*, *nakizueke*, *nakizuen*, *nakizuke*, *nakizun*, *nan*, *natzaiake*, *natzaie*, *natzaieke*, *natzaik*, *natzain*, *natzainake*, *natzaio*, *natzaioke*, *natzaizu*, *natzaizue*, *natzaizueke*, *natzaizuke*, *nau*, *nauk*, *naun*, *nekatu*, *nekez*, *nendin*, *nengoen*, *nengoke*, *nengokeen*, *nengokiake*, *nengokian*, *nengokieke*, *nengokiekeen*, *nengokien*, *nengokinake*, *nengokinan*, *nengokioke*, *nengokiokeen*, *nengokion*, *nengokizueke*, *nengokizuekeen*, *nengokizuen*, *nengokizuke*, *nengokizukeen*, *nengokizun*, *nenkiake*, *nenkian*, *nenkieke*, *nenkiekeen*, *nenkien*, *nenkioke*, *nenkiokeen*, *nenkion*, *nenkizueke*, *nenkizuekeen*, *nenkizuen*, *nenkizuke*, *nenkizukeen*, *nenkizun*, *nerau*, *neu*, *neure*, *nihaur*, *nik*, *ninduan*, *nindunan*, *ninteke*, *nintekeen*, *nintzaiake*, *nintzaian*, *nintzaieke*, *nintzaiekeen*, *nintzaien*, *nintzainake*, *nintzainan*, *nintzaioke*, *nintzaiokeen*, *nintzaion*, *nintzaizueke*, *nintzaizuekeen*, *nintzaizuen*, *nintzaizuke*, *nintzaizukeen*, *nintzaizun*, *nintzateke*, *nintzatekeen*, *nintzen*, *nituzke*, *nituzkeen*, *norakoak*, *norbaiten*, *norbera*, *norberaren*, *nori*, *nuke*, *nukeen*, *omen*, *omeone*, *ondo*, *ondoan*, *ondoren*, *ondorioz*, *oraindik*, *ordea*, *oso*, *osoa*, *ostera*, *ote*, *sartu*, *segidan*, *sekula*, *truk*, *truke*, *utzi*, *utziz*, *zagozkidake*, *zagozkidakete*, *zagozkidate*, *zagozkie*, *zagozkieke*, *zagozkiekete*, *zagozkiete*, *zagozkigu*, *zagozkiguke*, *zagozkigukete*, *zagozkigute*, *zagozkio*, *zagozkioke*, *zagozkiokete*, *zagozkiote*, *zagozkit*, *zaiake*, *zaidake*, *zaie*, *zaieke*, *zaigu*, *zaiguke*, *zaik*, *zain*, *zainake*, *zaio*, *zaioke*, *zait*, *zaitez*, *zaitezke*, *zaitezkete*, *zaitezte*, *zaizkiake*, *zaizkidake*, *zaizkie*, *zaizkieke*, *zaizkigu*, *zaizkiguke*, *zaizkik*, *zaizkin*, *zaizkinake*, *zaizkio*, *zaizkioke*, *zaizkit*, *zaizkizu*, *zaizkizue*, *zaizkizueke*, *zaizkizuke*, *zaizu*, *zaizue*, *zaizueke*, *zaizuke*, *zakizkidake*, *zakizkidakete*, *zakizkidan*, *zakizkidaten*, *zakizkieke*, *zakizkiekete*, *zakizkien*, *zakizkieten*, *zakizkiguke*, *zakizkigukete*, *zakizkigun*, *zakizkiguten*, *zakizkioke*, *zakizkiokete*, *zakizkion*, *zakizkioten*, *zara*, *zarateke*, *zarete*, *zaretekete*, *zatekeen*, *zatia*, *zatzaizkidake*, *zatzaizkidakete*, *zatzaizkidate*, *zatzaizkie*, *zatzaizkieke*, *zatzaizkiekete*, *zatzaizkiete*, *zatzaizkigu*, *zatzaizkiguke*, *zatzaizkigukete*, *zatzaizkigute*, *zatzaizkio*, *zatzaizkioke*, *zatzaizkiokete*, *zatzaizkiote*, *zatzaizkit*, *zaude*, *zaudeke*, *zaudekete*, *zaudete*, *zazu*, *ze*, *zedin*, *zegoen*, *zegokeen*, *zegokian*, *zegokidakeen*, *zegokidan*, *zegokiekeen*, *zegokien*, *ze-*

gokigukeen, *zegokigun*, *zegokinan*, *zegokiokeen*, *zegokion*, *zegokizuekeen*, *zegokizuen*, *zegokizukeen*, *zegokizun*, *zegozkian*, *zegozkidakeen*, *zegozkidan*, *zegozkiekeen*, *zegozkien*, *zegozkigukeen*, *zegozkigun*, *zegozkinan*, *zegozkiokeen*, *zegozkion*, *zegozkizuekeen*, *zegozkizuen*, *zegozkizukeen*, *zegozkizun*, *zehar*, *zekian*, *zekidakeen*, *zekidan*, *zekiekeen*, *zekien*, *zekigukeen*, *zekigun*, *zekiokeen*, *zekion*, *zekizkian*, *zekizkidakeen*, *zekizkidan*, *zekizkiekeen*, *zekizkien*, *zekizkigukeen*, *zekizkigun*, *zekizkiokeen*, *zekizkion*, *zekizkizuekeen*, *zekizkizuen*, *zekizkizukeen*, *zekizkizun*, *zekizuekeen*, *zekizuen*, *zekizukeen*, *zekizun*, *zen*, *zengozkidake*, *zengozkidakeen*, *zengozkidakete*, *zengozkidaketen*, *zengozkidan*, *zengozkidaten*, *zengozkieke*, *zengozkiekeen*, *zengozkiekete*, *zengozkieketen*, *zengozkien*, *zengozkieten*, *zengozkiguke*, *zengozkigukeen*, *zengozkigukete*, *zengozkiguketen*, *zengozkigun*, *zengozkiguten*, *zengozkioke*, *zengozkiokeen*, *zengozkiokete*, *zengozkioketen*, *zengozkion*, *zengozkioten*, *zenidate*, *zenigute*, *zenituzke*, *zenituzkeen*, *zenituzkete*, *zenituzketen*, *zenkizkidake*, *zenkizkidakeen*, *zenkizkidakete*, *zenkizkidaketen*, *zenkizkidan*, *zenkizkidaten*, *zenkizkieke*, *zenkizkiekeen*, *zenkizkiekete*, *zenkizkieketen*, *zenkizkien*, *zenkizkieten*, *zenkizkiguke*, *zenkizkigukeen*, *zenkizkigukete*, *zenkizkiguketen*, *zenkizkigun*, *zenkizkiguten*, *zenkizkioke*, *zenkizkiokeen*, *zenkizkiokete*, *zenkizkioketen*, *zenkizkion*, *zenkizkioten*, *zenuke*, *zenukeen*, *zenukete*, *zenuketen*, *zeren*, *zergatik*, *zerok*, *zerori*, *zertaz*, *zeu*, *zeudekeen*, *zeuden*, *zeuek*, *zeundeke*, *zeundekeen*, *zeundeketen*, *zeunden*, *zeundeten*, *zeure*, *zinateke*, *zinatekeen*, *zinatekete*, *zinateketen*, *zinen*, *zineten*, *zintezen*, *zintezke*, *zintezkeen*, *zintezketen*, *zintezten*, *zintzaizkidake*, *zintzaizkidakeen*, *zintzaizkidakete*, *zintzaizkidaketen*, *zintzaizkidan*, *zintzaizkidaten*, *zintzaizkieke*, *zintzaizkiekeen*, *zintzaizkiekete*, *zintzaizkieketen*, *zintzaizkien*, *zintzaizkieten*, *zintzaizkiguke*, *zintzaizkigukeen*, *zintzaizkigukete*, *zintzaizkiguketen*, *zintzaizkigun*, *zintzaizkiguten*, *zintzaizkioke*, *zintzaizkiokeen*, *zintzaizkiokete*, *zintzaizkioketen*, *zintzaizkion*, *zintzaizkioten*, *ziratekeen*, *ziren*, *zitekeen*, *zitezen*, *zitezkeen*, *zituan*, *zitunan*, *zitzaian*, *zitzaidakeen*, *zitzaidan*, *zitzaiekeen*, *zitzaien*, *zitzaigukeen*, *zitzaigun*, *zitzainan*, *zitzaiokeen*, *zitzaion*, *zitzaizkian*, *zitzaizkidakeen*, *zitzaizkidan*, *zitzaizkiekeen*, *zitzaizkien*, *zitzaizkigukeen*, *zitzaizkigun*, *zitzaizkinan*, *zitzaizkiokeen*, *zitzaizkion*, *zitzaizkizuekeen*, *zitzaizkizuen*, *zitzaizkizukeen*, *zitzaizkizun*, *zitzaizuekeen*, *zitzaizuen*, *zitzaizukeen*, *zitzaizun*, *zuan*, *zuhaur*, *zuihauk*, *zunan*, *zurea*, *zutaz*.

Define the pattern **PronBasque** as any one from the following list:

bera, *beraiek*, *beraien*, *berak*, *beraren*, *berau*, *berauek*, *bere*, *beren*, *berori*, *beroriek*, *eurak*, *ezer*, *genian*, *genien*, *geninan*, *genion*, *genizuen*, *genizun*, *gu*, *gure*, *haiek*, *haien*, *hala*, *han*, *handik*, *hango*, *hara*, *haren*, *hau*, *hauek*, *hemen*, *hemendik*, *hemengo*, *hi*, *hidan*, *hien*, *higun*, *hion*, *hire*, *hona*, *honela*, *hor*, *hori*, *horiek*, *horko*, *horra*, *horrela*, *hortik*, *hura*, *inoiz*, *inola*, *inon*, *inor*, *ni*, *nian*, *nien*, *ninan*, *nion*, *nire*, *nizuen*, *nizun*, *noiz*, *noizbait*, *nola*, *nolabait*, *non*, *nonbait*, *nondik*, *nongo*, *nor*, *nora*, *norbait*, *orain*, *orduan*, *zein*, *zenidan*, *zenidaten*, *zenien*, *zenieten*, *zenigun*, *zeniguten*, *zenion*, *zenioten*, *zer*, *zerbait*, *zian*, *ziaten*, *zidan*, *zidaten*, *zien*, *zieten*, *zigun*, *ziguten*, *zinaten*, *zion*, *zioten*, *zizuen*, *zizueten*, *zizun*, *zizuten*, *zu*, *zuek*, *zuen*, *zure*.

Define **tenerBasque** as the following pattern:

$(\varnothing|ba|e)(\varnothing|hen|le|nen)(\varnothing|be|da|ga|ge|ha|na|za|ze)(\varnothing|n)uka\mathbf{X}$.

If a word exactly matches any one of the following:

(**BasqueStopRoot**)**X**,
BasqueStopString,
(∅|*edo*)**PronBasque**(∅|*bait*|*nahi*|*txe*|*xe*)**X**,
tenerBasque,

then we consider it a Basque stop word. All the Basque stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

9.1.1 Effective spelling and essential root

Definition 9.2 (Basque protected range)**.** Set $\mathbf{V} = (a|e|i|o|u)$, $\mathbf{C} = \overline{\mathbf{V}}$. Search for the pattern $\mathbf{V}_m\mathbf{C}_m\mathbf{V}\sim$ or

$$(\varnothing|ba|bi|ko)(\varnothing|hen|hin|le|li|nen)(\varnothing|be|da|ga|ge|gi|ha|na|ni|za|ze|zi)\mathbf{C}_m\mathbf{V}_m\mathbf{C}\sim$$

in the string $\hat{\sigma}$. If such a pattern cannot be found, or $\hat{\sigma}$ does not contain **V**, then define $\rho_* = 0$; otherwise, define ρ_* as the last position occupied by the aforementioned pattern. If $\rho_* > 0$, $\ell(\hat{\sigma}) > \rho_*$, $\hat{\sigma}^{\{\rho_*\}} = r$ (*i.e.* the ρ_*-th character is the letter r) and $\hat{\sigma}^{\{\rho_*+1\}} = \mathbf{C}$ (*i.e.* the $(\rho_* + 1)$-st character is not a vowel), then $\mathrm{ProtRg}(\hat{\sigma}) = 1 + \rho_*$; otherwise, set $\mathrm{ProtRg}(\hat{\sigma}) = \rho_*$.

Algorithm 9.3 (Basque effective spelling)**.** *Set* $\mathbf{V} = (a|e|i|o|u)$, $\mathbf{C} = \overline{\mathbf{V}}$. *For a Basque word* $\hat{\sigma}$*, its effective spelling* $\mathrm{EffSpell}(\hat{\sigma})$ *is constructed in the following steps:*

(1) *Convert to lowercase and do* $\hat{\chi}_m gabe(\varnothing|a)\sim \rightarrow xx\hat{\chi}_m$.

(2) *Do tx→č.*

(3) *Do (izeb|izeko)~→ αunτ,* **C***y→* **C***iy, anai~→neb, arreb~→ahizp, lehengus(in|u)***X***~→coωσiῖν.*

(4) *Replace*

aban~	*abe~*	*agir~*	*aldarri~*	*andereñ~*	*ater~*	*bakar~*	*bar~*	*begira~*	*begiru~*	*begitart~*
βa	*ab*	*aγir*	*aldrri*	*μiσ*	*atr*	*σiῖγr*	*λaφ*	*λuκa*	*ρeσπτu*	*faσt*
bil$_{\hat{\chi}\notin}$(a)~	*bila~*	*burg~*	*buru~*	*carolin~*	*charlotte~*	*dantz~*	*denbor~*	*des$_{\hat{\chi}\notin}$(i)~*	*desio~*	
βiλ$\hat{\chi}$	*βiaλ*	*buρg*	*bru*	*κρaln*	*καrλotte*	*δac*	*τmeπ*	*δds$\hat{\chi}$*	*dsir*	
desi~	*desk~*	*egona~*	*eliza***X***~*	*era$_{\hat{\chi}\notin}$(m|t)~*	*err~*	*familia~*	*fitz~*	*gela~*	*gorrot~*	*hertf~*
dsi	*dsk*	*σtaya*	*εelζ*	*exq$\hat{\chi}$*	*εrr*	*φμaλi*	*φτζ*	*gelα*	*ηat*	*ηrτeφ*
*hobe***X***~*	*hunsf~*	*ibil(ald|er)***X***~*	*ideia~*	*igaro~*	*ikar~*	*indi~*	*irr~*	*izen~*	*jane***X***~*	*kant~*
βeστ	*hnsf*	*ibili*	*ιδea*	*σpendo*	*ckar*	*ndi*	*ιrr*	*ῖaμ*	*jaῖ*	*canτ*
karta~	*kol~*	*liburuteg~*	*liydia~*	*londres~*	*lots~*	*maria~*	*mariy~*	*meriy~*	*mingarri~*	*musik~*
καρτa	*kl*	*λβiρ*	*ληδiα*	*lndres*	*σhaμ*	*μaριa*	*maρη*	*μry*	*παiῖi*	*muσκ*
negozio~	*negu~*	*rst*	*senarra~*	*senti~*	*tentel~*	*ugar~*	*zakar~*	*amak*	*de*	*long*	*on(∅|a|ak|ik)*
ngoz	*ῖegω*	*ρστ*	*huσβa*	*φeeλi*	*φωol*	*ugr*	*ρuδr*	*ama*	*δeε*	*lῖoγ*	*βeστ*

(5) *Replace*

*(∅|ba|bi|zi)(∅|li)(d|g|h|j|n|z)(∅|in)(∅|d)oa(∅|k|l|n|t|z)***X***~*
βgoβ
*(∅|ba|bi|zi)(∅|li)(d|g|h|j|n|z)(∅|in)(∅|d)(∅|a|e)ki(∅|k|l|n|t|z)***X***~*
βknowβ
*(∅|ba|e)(∅|hen|le|nen)(∅|be|da|ga|ge|ha|na|za|ze)(∅|n)ka(r|z)***X***~*
ekarr
*(∅|ba|e)(∅|hen|le|nen)(∅|be|da|ga|ge|ha|na|za|ze)(∅|n)rama***X***~*
βτaκβ

ahazt~	*begi***X***~*	*ego~*	*eliz$_{\hat{\chi}\notin}$(a)~*	*elizab~*	*harro~*	*kont(a|u)***X***~*	*bru(∅|bid***X***|ko)*
foργt	*eye*	*eg*	*εliζ$\hat{\chi}$*	*εlizaβ*	*πiρδo*	*κoῖτ*	*bruak*

(6) *Call the result so far as $\hat{\sigma}^{\dagger}$, and break it down into $\hat{\sigma}^{\dagger} = \hat{\sigma}_1^{\dagger}\hat{\sigma}_2^{\dagger}$, where $\ell(\hat{\sigma}_1^{\dagger}) = \mathrm{ProtRg}(\hat{\sigma}^{\dagger})$. Work on $\hat{\sigma}_2^{\dagger}$ as follows:*

(6.1) *Do č→tx.*

(6.2) *Delete any pattern in the following list:*

ago, aldatu, aldi, aldia, an, arazi, ari, atu, atze, bera, bide, bidea, dako, du, dura, ean, era, errez, erreza, eta, etan, etari, ez, eza, ezin, ezina, gai, gaia, gailu, gailua, gaitz, gaitza, gale, galea, go, gune, gunea, gura, ide, idea, ka, kaitz, kaitza, kan, kari, karia, karri, karria, kera, keta, ki, kide, kidea, kin, kina, kizun, kizuna, kor, korra, kuna, kunde, kundea, kune, kunea, kuntza, kura, la, lari, le, men, mena, or, orra, pen, pena, pera, pide, pidea, rean, rekin, taile, tailea, taldi, taldia, tarazi, tari, tatu, tezin, tezina, tio, tu, tun, tuna, tura, tzaga, tzaile, tzailea, tzaka, tzake, tzat, tze, tzeke, tzez

and the letters (if any) thereafter.

(6.3) *Delete any pattern in the following list:*

ada, ail, aizun, ak, alde, aldea, aldi, aldia, anda, anga, antza, ar, ara, ari, aria, aro, aroa, arte, artea, asi, asia, asun, asuna, aurre, aurrea, behar, bera, bizia, burua, dar, dara, degi, degia, denda, di, du, dua, dun, duna, duri, duria, duru, durua, egi, egia, ek, eko, eme, emea, ena, enea, eria, ero, eroa, eroz, eroza, estu, estua, eta, etako, etan, etara, etxe, etxea, ez, eza, ezia, ga, gabe, gabea, gai, gaia, garna, garren, garrena, ge, gei, geia, gela, gerren, gerrena, gibel, gibela, gile, gilea, gintza, gintzo, gintzu, giro, go, goi, gune, gunea, handi, handia, i, ka, kabe, kabea, kada, kail, kaila, kalde, kaldea, kan, kana, kari, karia, kera, keria, ket, keta, ki, kia, kide, kin, kina, kintza, kirri, kirria, ko, koa, koi, koia, koitz, koitza, kondo, kondoa, kor, korra, kote, kotea, kume, kumea, kuntza, lari, laria, larri, larria, leku, lekua, liar, liara, mendi, mendia, mendu, mendua, mentu, mentua, min, mina, n, na, nahi, nahia, ne, nea, ngo, ngoa, no, noa, o, oa, ohi, ohia, oi, oia, ola, ondo, ondoa, ontzi, ontzia, orde, ordea, ordu, ordua, oro, oroa, os, osa, oso, osoa,

oste, ostea, pe, pea, pera, ra, ro, sa, ska, skila, sko, sta, ta, tako, takoa, talde, taldea, taldi, taldia, tan, tar, tara, tari, taria, tarik, tariko, taro, taroa, tasun, tasuna, te, tea, tegi, tegia, teria, ti, tia, tiar, tiara, tila, to, toa, toki, tokia, tra, tsu, tsua, tto, ttoa, tu, tua, tuko, txo, txoa, txu, txua, tz, tzain, tzaina, tzale, tzalea, tzar, tzara, tzarra, tzo, tzoa, tzu, tzua, una, une, unea, urren, urrena, xka, z, za, zain, zaina, zale, zalea, zaro, zaroa, zino, zinoa, zio, zioa, zione, zionea, zka, zko, zkoa, zp, zto, ztoa, zu, zua

and the letters (if any) thereafter.

(6.4) *Delete any pattern in the following list:*

dade, date, era, ero, gi, go, ik, keria, ki, la, lanik, larik, rik, ro, tade, tate, to, ztik

and the letters (if any) thereafter.

(6.5) *Call the result so far as* $\hat{\sigma}_2^{\ddagger}$. *Concatenate* $\hat{\sigma}_1^{\dagger}$ *and* $\hat{\sigma}_2^{\ddagger}$.

(7) *Replace*

$$(\varnothing|ba|e)(\varnothing|hen|le|nen)(\varnothing|be|da|ga|ge|ha|na|za|ze)(\varnothing|n)to(r|z)\mathbf{X}\sim$$
$$etorr$$

$$(\varnothing|ba|i)(\varnothing|hen|le|nen)(\varnothing|be|da|ga|ge|ha|na|za|ze)(\varnothing|n)bil\mathbf{X}\sim \qquad (b|d|gen|h|n|z|zen)io(d|e|g|k|n|s|t|z)\mathbf{X}\sim \qquad \underline{one(\varnothing|g)}$$
$$\beta walk\beta \qquad esa \qquad \beta e\sigma\tau$$

Algorithm 9.4 (Basque essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Basque word, then its corresponding essential root* $\mathrm{EssRoot}(\hat{\sigma})$ *is constructed as* $\hat{\sigma}^{[\mathrm{ProtRg}(\hat{\sigma})]}$.

9.1.2 Approximate clustering

There will not be any mutation rules in our algorithm for Basque, so SimpHrdTest = HrdTest.

Algorithm 9.5 (Simple heredity test)**.** *Suppose that* $\hat{\alpha}$ *and* $\hat{\beta}$ *are both lowercase strings. The Boolean-valued function* $\mathrm{SimpHrdTest}(\hat{\alpha},\hat{\beta})$ *returns* **TRUE** *if* $\hat{\alpha}$ *contains at least one instance of* **V***,* **AND** *at least one of the following two conditions holds:*[107]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\alpha}' = \hat{\beta}'$, *where the strings with prime result from doing* $\sim^{\mathbf{X}\in}(\mathbf{V}\hat{\chi})(\varnothing|b)(a|e) \to \mathbf{X}$ *on their counterparts without prime.*

Algorithm 9.6 (Approximate clustering of Basque words)**.** *The approximate clustering of a list of Basque words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1, \mathrm{EffSpell}(\hat{\alpha}_1)), \ldots, (\hat{\alpha}_N, \mathrm{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the last component. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1)})), \ldots, (\hat{\alpha}_{(N)}, \mathrm{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfy* $\mathrm{HrdTest}(\mathrm{EffSpell}(\hat{\alpha}_{(n)}), \mathrm{EffSpell}(\hat{\alpha}_{(n+1)})) = \mathbf{FALSE}$*, where* $n \in \mathbb{Z} \cap [1, N)$*, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)}, *)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, *), \ldots, (\hat{\alpha}_{(M,n_M)}, *)\}\}$*, where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, *), \ldots, (\hat{\alpha}_{(m,n_m)}, *)\}$ *where* $m \in \mathbb{Z} \cap [1, M]$*, we augment it into a tagged entry* $G_m = (g_m, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)}), \mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with higher priority), and* $\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lower priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)})\}$ *satisfy*

$$\mathrm{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \mathbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \mathbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$*, then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots) from each* Γ_k*, where* $k \in \mathbb{Z} \cap [1, K]$.

Example 9.6.1. Our clustering algorithm correctly groups the following families of Basque words:

[107] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

ahizpa, *ahizpak*, *ahizparekin*, *ahizparen*, *ahizparengan*, *ahizparengana*, *ahizparenganaino*, *ahizparenganako*, *ahizparenganantz*, *ahizparengandik*, *ahizparengatik*, *ahizparentzat*, *ahizpari*, *ahizparik*, *ahizpatzat*, *ahizpaz*, *ahizpei*, *ahizpek*, *ahizpekin*, *ahizpen*, *ahizpengan*, *ahizpengana*, *ahizpenganaino*, *ahizpenganako*, *ahizpenganantz*, *ahizpengandik*, *ahizpengatik*, *ahizpentzat*, *ahizpez*, *arreba*, *arrebak*, *arrebarekin*, *arrebaren*, *arrebarengan*, *arrebarengana*, *arrebarenganaino*, *arrebarenganako*, *arrebarenganantz*, *arrebarengandik*, *arrebarengatik*, *arrebarentzat*, *arrebari*, *arrebarik*, *arrebatzat*, *arrebaz*, *arrebei*, *arrebek*, *arrebekin*, *arreben*, *arrebengan*, *arrebengana*, *arrebenganaino*, *arrebenganako*, *arrebenganantz*, *arrebengandik*, *arrebengatik*, *arrebentzat*, *arrebez* — "sister";
anaia, *anaiak*, *anaiarekin*, *anaiaren*, *anaiarengan*, *anaiarengana*, *anaiarenganaino*, *anaiarenganako*, *anaiarenganantz*, *anaiarengandik*, *anaiarengatik*, *anaiarentzat*, *anaiari*, *anaiarik*, *anaiatzat*, *anaiaz*, *anaiei*, *anaiek*, *anaiekin*, *anaien*, *anaiengan*, *anaiengana*, *anaienganaino*, *anaienganako*, *anaienganantz*, *anaiengandik*, *anaiengatik*, *anaientzat*, *anaiez*, *neba*, *nebak*, *nebarekin*, *nebaren*, *nebarengan*, *nebarengana*, *nebarenganaino*, *nebarenganako*, *nebarenganantz*, *nebarengandik*, *nebarengatik*, *nebarentzat*, *nebari*, *nebarik*, *nebatzat*, *nebaz*, *nebei*, *nebek*, *nebekin*, *neben*, *nebengan*, *nebengana*, *nebenganaino*, *nebenganako*, *nebenganantz*, *nebengandik*, *nebengatik*, *nebentzat*, *nebez* — "brother";
izeba, *izeba*, *izebak*, *izebak*, *izebak*, *izebarekin*, *izebarekin*, *izebaren*, *izebaren*, *izebarengan*, *izebarengan*, *izebarengana*, *izebarengana*, *izebarenganaino*, *izebarenganaino*, *izebarenganako*, *izebarenganako*, *izebarenganantz*, *izebarenganantz*, *izebarengandik*, *izebarengandik*, *izebarengatik*, *izebarengatik*, *izebarentzat*, *izebarentzat*, *izebari*, *izebari*, *izebarik*, *izebatzat*, *izebaz*, *izebaz*, *izebei*, *izebek*, *izebekin*, *izeben*, *izebengan*, *izebengana*, *izebenganaino*, *izebenganako*, *izebenganantz*, *izebengandik*, *izebengatik*, *izebentzat*, *izebez*, *izeko*, *izekoa*, *izekoak*, *izekoak*, *izekoarekin*, *izekoaren*, *izekoarengan*, *izekoarengana*, *izekoarenganaino*, *izekoarenganako*, *izekoarenganantz*, *izekoarengandik*, *izekoarengatik*, *izekoarentzat*, *izekoari*, *izekoaz*, *izekoei*, *izekoek*, *izekoekin*, *izekoen*, *izekoengan*, *izekoengana*, *izekoenganaino*, *izekoenganako*, *izekoenganantz*, *izekoengandik*, *izekoengatik*, *izekoentzat*, *izekoez*, *izekok*, *izekorekin*, *izekoren*, *izekorengan*, *izekorengana*, *izekorenganaino*, *izekorenganako*, *izekorenganantz*, *izekorengandik*, *izekorengatik*, *izekorentzat*, *izekori*, *izekorik*, *izekotzat*, *izekoz* — "aunt".

It also correctly cluster each of the following verbs:

bagenbiltza, *bagenbilzkie*, *bagenbilzkik*, *bagenbilzkin*, *bagenbilzkio*, *bagenbilzkizu*, *bagenbilzkizue*, *bahenbilkie*, *bahenbilkigu*, *bahenbilkio*, *bahenbilkit*, *balebilkie*, *balebilkigu*, *balebilkik*, *balebilkin*, *balebilkio*, *balebilkit*, *balebilkizu*, *balebilkizue*, *balebiltza*, *balebilzkie*, *balebilzkigu*, *balebilzkik*, *balebilzkin*, *balebilzkio*, *balebilzkit*, *balebilzkizu*, *balebilzkizue*, *banenbilkie*, *banenbilkik*, *banenbilkin*, *banenbilkio*, *banenbilkizu*, *banenbilkizue*, *bazenbiltza*, *bazenbiltzate*, *bazenbilzkidate*, *bazenbilzkie*, *bazenbilzkiete*, *bazenbilzkigu*, *bazenbilzkigute*, *bazenbilzkio*, *bazenbilzkiote*, *bazenbilzkit*, *bebiltza*, *bebilzkie*, *bebilzkigu*, *bebilzkik*, *bebilzkin*, *bebilzkio*, *bebilzkit*, *bebilzkizu*, *bebilzkizue*, *benbilkie*, *benbilkigu*, *benbilkik*, *benbilkin*, *benbilkio*, *benbilkit*, *benbilkizu*, *benbilkizue*, *dabilela*, *dabilke*, *dabilkiake*, *dabilkidake*, *dabilkie*, *dabilkieke*, *dabilkigu*, *dabilkiguke*, *dabilkik*, *dabilkin*, *dabilkinake*, *dabilkio*, *dabilkioke*, *dabilkit*, *dabilkizu*, *dabilkizue*, *dabilkizueke*, *dabilkizuke*, *dabiltza*, *dabiltzala*, *dabilzke*, *dabilzkiake*, *dabilzkidake*, *dabilzkie*, *dabilzkieke*, *dabilzkigu*, *dabilzkiguke*, *dabilzkik*, *dabilzkin*, *dabilzkinake*, *dabilzkio*, *dabilzkioke*, *dabilzkit*, *dabilzkizu*, *dabilzkizue*, *dabilzkizueke*, *dabilzkizuke*, *gabiltza*, *gabiltzan*, *gabilzke*, *gabilzkiake*, *gabilzkian*, *gabilzkie*, *gabilzkieke*, *gabilzkien*, *gabilzkik*, *gabilzkin*, *gabilzkinake*, *gabilzkinan*, *gabilzkio*, *gabilzkioke*, *gabilzkion*, *gabilzkizu*, *gabilzkizue*, *gabilzkizueke*, *gabilzkizuen*, *gabilzkizuke*, *gabilzkizun*, *genbiltzan*, *genbilzke*, *genbilzkeen*, *genbilzkiake*, *genbilzkian*, *genbilzkieke*, *genbilzkiekeen*, *genbilzkien*, *genbilzkinake*, *genbilzkinan*, *genbilzkioke*, *genbilzkiokeen*, *genbilzkion*, *genbilzkizueke*, *genbilzkizuekeen*, *genbilzkizuen*, *genbilzkizuke*, *genbilzkizukeen*, *genbilzkizun*, *habilke*, *habilkidake*, *habilkie*, *habilkieke*, *habilkigu*, *habilkiguke*, *habilkio*, *habilkioke*, *habilkit*, *henbilen*, *henbilke*, *henbilkeen*, *henbilkidake*, *henbilkidakeen*, *henbilkidan*, *henbilkieke*, *henbilkiekeen*, *henbilkien*, *henbilkiguke*, *henbilkigukeen*, *henbilkigun*, *henbilkioke*, *henbilkiokeen*, *henbilkion*, *ibili*, *ibiliko*, *ibiltze*, *ibiltzen*, *lebilke*, *lebilkiake*, *lebilkidake*, *lebilkieke*, *lebilkiguke*, *lebilkinake*, *lebilkioke*, *lebilkizueke*, *lebilkizuke*, *lebilzke*, *lebilzkiake*, *lebilzkidake*, *lebilzkieke*, *lebilzkiguke*, *lebilzkinake*, *lebilzkioke*, *lebilzkizueke*, *lebilzkizuke*, *nabilen*, *nabilke*, *nabilkiake*, *nabilkian*, *nabilkie*, *nabilkieke*, *nabilkien*, *nabilkik*, *nabilkin*, *nabilkinake*, *nabilkinan*, *nabilkio*, *nabilkioke*, *nabilkion*, *nabilkizu*, *nabilkizue*, *nabilkizueke*, *nabilkizuen*, *nabilkizuke*, *nabilkizun*, *nenbilen*, *nenbilke*, *nenbilkeen*, *nenbilkiake*, *nenbilkian*, *nenbilkieke*, *nenbilkiekeen*, *nenbilkien*, *nenbilkinake*, *nenbilkinan*, *nenbilkioke*, *nenbilkiokeen*, *nenbilkion*, *nenbilkizueke*, *nenbilkizuekeen*, *nenbilkizuen*, *nenbilkizuke*, *nenbilkizukeen*, *nenbilkizun*, *zabiltza*, *zabiltzate*, *zabilzke*, *zabilzkete*, *zabilzkidake*, *zabilzkidakete*, *zabilzkidate*, *zabilzkie*, *zabilzkieke*, *zabilzkiekete*, *zabilzkiete*, *zabilzkigu*, *zabilzkiguke*, *zabilzkigukete*, *zabilzkigute*, *zabilzkio*, *zabilzkioke*, *zabilzkiokete*, *zabilzkiote*, *zabilzkit*, *zebilen*, *zebilkeen*, *zebilkian*, *zebilkidakeen*, *zebilkidan*, *zebilkiekeen*, *zebilkien*, *zebilkigukeen*, *zebilkigun*, *zebilkinan*, *zebilkiokeen*, *zebilkion*, *zebilkizuekeen*, *zebilkizuen*, *zebilkizukeen*, *zebilkizun*, *zebiltzan*, *zebilzkeen*, *zebilzkian*, *zebilzkidakeen*, *zebilzkidan*, *zebilzkiekeen*, *zebilzkien*, *zebilzkigukeen*, *zebilzkigun*, *zebilzkinan*, *zebilzkiokeen*, *zebilzkion*, *zebilzkizuekeen*, *zebilzkizuen*, *zebilzkizukeen*, *zebilzkizun*, *zenbiltzan*, *zenbiltzaten*, *zenbilzke*, *zenbilzkeen*, *zenbilzkete*, *zenbilzketen*, *zenbilzkidake*, *zenbilzkida-*

keen, *zenbilzkidakete*, *zenbilzkidaketen*, *zenbilzkidan*, *zenbilzkidaten*, *zenbilzkieke*, *zenbilzkiekeen*, *zenbilzkiekete*, *zenbilzkieketen*, *zenbilzkien*, *zenbilzkieten*, *zenbilzkiguke*, *zenbilzkigukeen*, *zenbilzkigukete*, *zenbilzkiguketen*, *zenbilzkigun*, *zenbilzkiguten*, *zenbilzkioke*, *zenbilzkiokeen*, *zenbilzkiokete*, *zenbilzkioketen*, *zenbilzkion*, *zenbilzkioten* — “walk”;

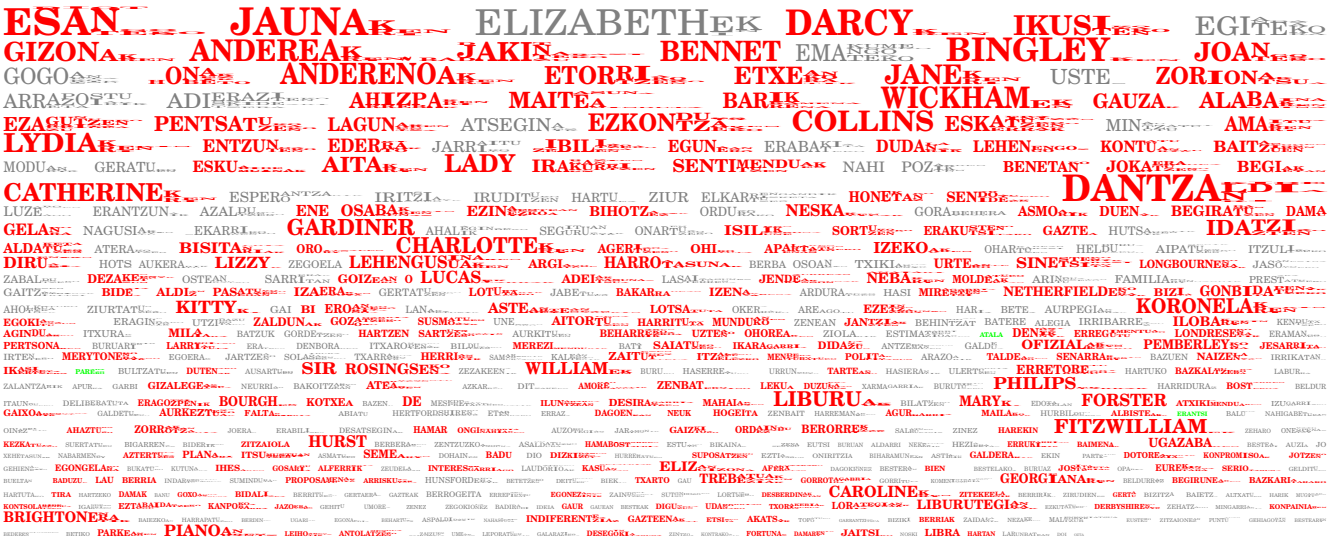
diodaz, *diogu*, *dioguz*, *diok*, *diosagu*, *diosat*, *diosate*, *dioska*, *diosknа*, *diosku*, *dioskuk*, *dioskute*, *dioskutez*, *dioskuz*, *dioskuzak*, *dioskuzu*, *dioskuzue*, *dioskuzuez*, *dioskuzuz*, *diosnagu*, *diosnat*, *diosnate*, *diost*, *diostak*, *diostate*, *diostatez*, *diostaz*, *diostazak*, *diostazu*, *diostazue*, *diostazuez*, *diostazuz*, *diot*, *diote*, *diotez*, *diotse*, *diotsedaz*, *diotsegu*, *diotseguz*, *diotsek*, *diotset*, *diotsete*, *diotsetez*, *diotsez*, *diotsezak*, *diotsezu*, *diotsezue*, *diotsezuez*, *diotsezuz*, *diotso*, *diotsodaz*, *diotsogu*, *diotsoguz*, *diotsok*, *diotsot*, *diotsote*, *diotsotez*, *diotsoz*, *diotsozak*, *diotsozu*, *diotsozue*, *diotsozuez*, *diotsozuz*, *diotsu*, *diotsudaz*, *diotsue*, *diotsuedaz*, *diotsuegu*, *diotsueguz*, *diotsuet*, *diotsuete*, *diotsuetez*, *diotsuez*, *diotsugu*, *diotsuguz*, *diotsut*, *diotsute*, *diotsutez*, *diotsuz*, *dioz*, *diozaa*, *diozaagu*, *diozaat*, *diozaate*, *diozak*, *diozana*, *diozanagu*, *diozanat*, *diozanate*, *diozu*, *diozue*, *diozuez*, *diozuz*, *esan*, *esango*, *esate*, *esaten*, *genioen*, *genioezen*, *geniosan*, *geniosnan*, *geniotsen*, *geniotsezen*, *geniotson*, *geniotsozen*, *geniotsuen*, *geniotsuezen*, *geniotsun*, *geniotsuzen*, *geniozaa*, *geniozana*, *hioen*, *hioezen*, *hioskun*, *hioskuzen*, *hiostan*, *hiostazen*, *hiotsen*, *hiotsezen*, *hiotson*, *hiotsozen*, *nioen*, *nioezen*, *niosan*, *niosnan*, *niostsuen*, *niostsuezen*, *niotsen*, *niotsezen*, *niotson*, *niotsozen*, *niotsun*, *niotsuzen*, *niozaan*, *niozanan*, *zenioen*, *zenioezen*, *zenioskun*, *zenioskuten*, *zenioskutezen*, *zenioskuzen*, *zeniostan*, *zeniostaten*, *zeniostatezen*, *zeniostazen*, *zenioten*, *zeniotezen*, *zeniotsen*, *zeniotseten*, *zeniotsetezen*, *zeniotsezen*, *zeniotson*, *zeniotsoten*, *zeniotsotezen*, *zeniotsozen*, *zioen*, *zioezen*, *ziosan*, *ziosaten*, *zioskun*, *zioskuten*, *zioskutezen*, *zioskuzen*, *ziosnan*, *ziosnaten*, *ziostan*, *ziostaten*, *ziostatezen*, *ziostazen*, *zioten*, *ziotezen*, *ziotsen*, *ziotseten*, *ziotsetezen*, *ziotsezen*, *ziotson*, *ziotsoten*, *ziotsotezen*, *ziotsozen*, *ziotsuen*, *ziotsueten*, *ziotsuetezen*, *ziotsuezen*, *ziotsun*, *ziotsuten*, *ziotsutezen*, *ziotsuzen*, *ziozaa*, *ziozaate*, *ziozana*, *ziozanate* — “say”;

bagentoz, *bagentozkie*, *bagentozkik*, *bagentozkin*, *bagentozkio*, *bagentozkizu*, *bagentozkizue*, *bahentor*, *bahentorkie*, *bahentorkigu*, *bahentorkio*, *bahentorkit*, *baletor*, *baletorkie*, *baletorkigu*, *baletorkik*, *baletorkin*, *baletorkio*, *baletorkit*, *baletorkizu*, *baletorkizue*, *baletoz*, *baletozkie*, *baletozkigu*, *baletozkik*, *baletozkin*, *baletozkio*, *baletozkit*, *baletozkizu*, *baletozkizue*, *banentor*, *banentorkie*, *banentorkik*, *banentorkin*, *banentorkio*, *banentorkizu*, *banentorkizue*, *bazentoz*, *bazentozkidate*, *bazentozkie*, *bazentozkiete*, *bazentozkigu*, *bazentozkigute*, *bazentozkio*, *bazentozkiote*, *bazentozkit*, *bazentozte*, *bentorkie*, *bentorkigu*, *bentorkik*, *bentorkin*, *bentorkio*, *bentorkit*, *bentorkizu*, *bentorkizue*, *betor*, *betoz*, *betozkie*, *betozkigu*, *betozkik*, *betozkin*, *betozkio*, *betozkit*, *betozkizu*, *betozkizue*, *dator*, *datorke*, *datorkiake*, *datorkidake*, *datorkie*, *datorkieke*, *datorkigu*, *datorkiguke*, *datorkik*, *datorkin*, *datorkinake*, *datorkio*, *datorkioke*, *datorkit*, *datorkizu*, *datorkizue*, *datorkizueke*, *datorkizuke*, *datorrela*, *datoz*, *datozela*, *datozke*, *datozkiake*, *datozkidake*, *datozkie*, *datozkieke*, *datozkigu*, *datozkiguke*, *datozkik*, *datozkin*, *datozkinake*, *datozkio*, *datozkioke*, *datozkit*, *datozkizu*, *datozkizue*, *datozkizueke*, *datozkizuke*, *etor*, *etorri*, *etorriko*, *etortze*, *etortzen*, *gatoz*, *gatozen*, *gatozke*, *gatozkiake*, *gatozkian*, *gatozkie*, *gatozkieke*, *gatozkien*, *gatozkik*, *gatozkin*, *gatozkinake*, *gatozkinan*, *gatozkio*, *gatozkioke*, *gatozkion*, *gatozkizu*, *gatozkizue*, *gatozkizueke*, *gatozkizuen*, *gatozkizuke*, *gatozkizun*, *gentozen*, *gentozke*, *gentozkeen*, *gentozkiake*, *gentozkian*, *gentozkieke*, *gentozkiekeen*, *gentozkien*, *gentozkinake*, *gentozkinan*, *gentozkioke*, *gentozkiokeen*, *gentozkion*, *gentozkizueke*, *gentozkizuekeen*, *gentozkizuen*, *gentozkizuke*, *gentozkizukeen*, *gentozkizun*, *hator*, *hatorke*, *hatorkidake*, *hatorkie*, *hatorkieke*, *hatorkigu*, *hatorkiguke*, *hatorkio*, *hatorkioke*, *hatorkit*, *hentorke*, *hentorkeen*, *hentorkidake*, *hentorkidakeen*, *hentorkidan*, *hentorkieke*, *hentorkiekeen*, *hentorkien*, *hentorkiguke*, *hentorkigukeen*, *hentorkigun*, *hentorkioke*, *hentorkiokeen*, *hentorkion*, *hentorren*, *letorke*, *letorkiake*, *letorkidake*, *letorkieke*, *letorkiguke*, *letorkinake*, *letorkioke*, *letorkizueke*, *letorkizuke*, *letozke*, *letozkiake*, *letozkidake*, *letozkieke*, *letozkiguke*, *letozkinake*, *letozkioke*, *letozkizueke*, *letozkizuke*, *nator*, *natorke*, *natorkiake*, *natorkian*, *natorkie*, *natorkieke*, *natorkien*, *natorkik*, *natorkin*, *natorkinake*, *natorkinan*, *natorkio*, *natorkioke*, *natorkion*, *natorkizu*, *natorkizue*, *natorkizueke*, *natorkizuen*, *natorkizuke*, *natorkizun*, *natorren*, *nentorke*, *nentorkeen*, *nentorkiake*, *nentorkian*, *nentorkieke*, *nentorkiekeen*, *nentorkien*, *nentorkinake*, *nentorkinan*, *nentorkioke*, *nentorkiokeen*, *nentorkion*, *nentorkizueke*, *nentorkizuekeen*, *nentorkizuen*, *nentorkizuke*, *nentorkizukeen*, *nentorkizun*, *nentorren*, *zatoz*, *zatozke*, *zatozkete*, *zatozkidake*, *zatozkidakete*, *zatozkidate*, *zatozkie*, *zatozkieke*, *zatozkiekete*, *zatozkiete*, *zatozkigu*, *zatozkiguke*, *zatozkigukete*, *zatozkigute*, *zatozkio*, *zatozkioke*, *zatozkiokete*, *zatozkiote*, *zatozkit*, *zatozte*, *zentozen*, *zentozke*, *zentozkeen*, *zentozketen*, *zentozkidake*, *zentozkidakeen*, *zentozkidakete*, *zentozkidaketen*, *zentozkidan*, *zentozkidaten*, *zentozkieke*, *zentozkiekeen*, *zentozkiekete*, *zentozkieketen*, *zentozkien*, *zentozkieten*, *zentozkiguke*, *zentozkigukeen*, *zentozkigukete*, *zentozkiguketen*, *zentozkigun*, *zentozkiguten*, *zentozkioke*, *zentozkiokeen*, *zentozkiokete*, *zentozkioketen*, *zentozkion*, *zentozkioten*, *zentozten*, *zetorkeen*, *zetorkian*, *zetorkidakeen*, *zetorkidan*, *zetorkiekeen*, *zetorkien*, *zetorkigukeen*, *zetorkigun*, *zetorkinan*, *zetorkiokeen*, *zetorkion*, *zetorkizuekeen*, *zetorkizuen*, *zetorkizukeen*, *zetorkizun*, *zetorren*, *zetozen*, *zetozkeen*, *zetozkian*, *zetozkidakeen*, *zetozkidan*, *zetozkiekeen*, *zetozkien*, *zetozkigukeen*, *zetozkigun*, *zetozkinan*, *zetozkiokeen*, *zetozkion*, *zetozkizuekeen*, *zetozkizuen*, *zetozkizukeen*, *zetozkizun* — “come”;

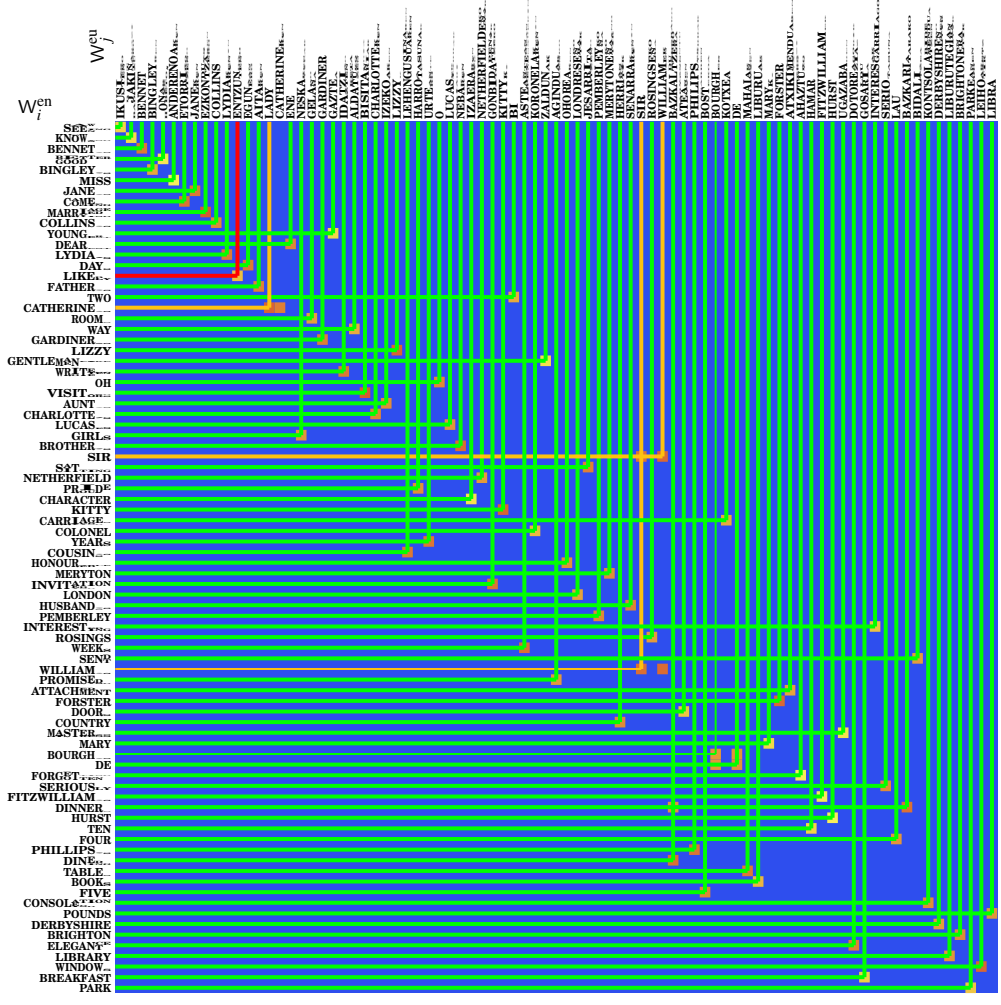
bagindoaz, *bagindoazkie*, *bagindoazkik*, *bagindoazkin*, *bagindoazkio*, *bagindoazkizu*, *bagindoazkizue*, *bahindoa*, *bahindoakie*, *bahindoakigu*, *bahindoakio*, *bahindoakit*, *balihoa*, *balihoakie*, *balihoakigu*, *balihoakik*, *balihoakin*,

balihoakio, *balihoakit*, *balihoakizu*, *balihoakizue*, *balihoaz*, *balihoazkie*, *balihoazkigu*, *balihoazkik*, *balihoazkin*, *balihoazkio*, *balihoazkit*, *balihoazkizu*, *balihoazkizue*, *banindoa*, *banindoakie*, *banindoakik*, *banindoakin*, *banindoakio*, *banindoakizu*, *banindoakizue*, *bazindoaz*, *bazindoazkidate*, *bazindoazkie*, *bazindoazkiete*, *bazindoazkigu*, *bazindoazkigute*, *bazindoazkio*, *bazindoazkiote*, *bazindoazkit*, *bazindoazte*, *bihoa*, *bihoaz*, *bihoazkie*, *bihoazkigu*, *bihoazkik*, *bihoazkin*, *bihoazkio*, *bihoazkit*, *bihoazkizu*, *bihoazkizue*, *bindoakie*, *bindoakigu*, *bindoakik*, *bindoakin*, *bindoakio*, *bindoakit*, *bindoakizu*, *bindoakizue*, *doa*, *doake*, *doakiake*, *doakidake*, *doakie*, *doakieke*, *doakigu*, *doakiguke*, *doakik*, *doakin*, *doakinake*, *doakio*, *doakioke*, *doakit*, *doakizu*, *doakizue*, *doakizueke*, *doakizuke*, *doala*, *doaz*, *doazela*, *doazke*, *doazkiake*, *doazkidake*, *doazkie*, *doazkieke*, *doazkigu*, *doazkiguke*, *doazkik*, *doazkin*, *doazkinake*, *doazkio*, *doazkioke*, *doazkit*, *doazkizu*, *doazkizue*, *doazkizueke*, *doazkizuke*, *gindoazen*, *gindoazke*, *gindoazkeen*, *gindoazkiake*, *gindoazkian*, *gindoazkieke*, *gindoazkiekeen*, *gindoazkien*, *gindoazkinake*, *gindoazkinan*, *gindoazkioke*, *gindoazkiokeen*, *gindoazkion*, *gindoazkizueke*, *gindoazkizuekeen*, *gindoazkizuen*, *gindoazkizuke*, *gindoazkizukeen*, *gindoazkizun*, *goaz*, *goazen*, *goazke*, *goazkiake*, *goazkian*, *goazkie*, *goazkieke*, *goazkien*, *goazkik*, *goazkin*, *goazkinake*, *goazkinan*, *goazkio*, *goazkioke*, *goazkion*, *goazkizu*, *goazkizue*, *goazkizueke*, *goazkizuen*, *goazkizuke*, *goazkizun*, *hindoake*, *hindoakeen*, *hindoakidake*, *hindoakidakeen*, *hindoakidan*, *hindoakieke*, *hindoakiekeen*, *hindoakien*, *hindoakiguke*, *hindoakigukeen*, *hindoakigun*, *hindoakioke*, *hindoakiokeen*, *hindoakion*, *hindoan*, *hoa*, *hoake*, *hoakidake*, *hoakie*, *hoakieke*, *hoakigu*, *hoakiguke*, *hoakio*, *hoakioke*, *hoakit*, *joan*, *joango*, *joate*, *joaten*, *lihoake*, *lihoakiake*, *lihoakidake*, *lihoakieke*, *lihoakiguke*, *lihoakinake*, *lihoakioke*, *lihoakizueke*, *lihoakizuke*, *lihoazke*, *lihoazkiake*, *lihoazkidake*, *lihoazkieke*, *lihoazkiguke*, *lihoazkinake*, *lihoazkioke*, *lihoazkizueke*, *lihoazkizuke*, *nindoake*, *nindoakeen*, *nindoakiake*, *nindoakian*, *nindoakieke*, *nindoakiekeen*, *nindoakien*, *nindoakinake*, *nindoakinan*, *nindoakioke*, *nindoakiokeen*, *nindoakion*, *nindoakizueke*, *nindoakizuekeen*, *nindoakizuen*, *nindoakizuke*, *nindoakizukeen*, *nindoakizun*, *nindoan*, *noa*, *noake*, *noakiake*, *noakian*, *noakie*, *noakieke*, *noakien*, *noakik*, *noakin*, *noakinake*, *noakinan*, *noakio*, *noakioke*, *noakion*, *noakizu*, *noakizue*, *noakizueke*, *noakizuen*, *noakizuke*, *noakizun*, *noan*, *zihoakeen*, *zihoakian*, *zihoakidakeen*, *zihoakidan*, *zihoakiekeen*, *zihoakien*, *zihoakigukeen*, *zihoakigun*, *zihoakinan*, *zihoakiokeen*, *zihoakion*, *zihoakizuekeen*, *zihoakizuen*, *zihoakizukeen*, *zihoakizun*, *zihoan*, *zihoazen*, *zihoazkeen*, *zihoazkian*, *zihoazkidakeen*, *zihoazkidan*, *zihoazkiekeen*, *zihoazkien*, *zihoazkigukeen*, *zihoazkigun*, *zihoazkinan*, *zihoazkiokeen*, *zihoazkion*, *zihoazkizuekeen*, *zihoazkizuen*, *zihoazkizukeen*, *zihoazkizun*, *zindoazen*, *zindoazke*, *zindoazkeen*, *zindoazketen*, *zindoazkidake*, *zindoazkidakeen*, *zindoazkidakete*, *zindoazkidaketen*, *zindoazkidan*, *zindoazkidaten*, *zindoazkieke*, *zindoazkiekeen*, *zindoazkiekete*, *zindoazkieketen*, *zindoazkien*, *zindoazkieten*, *zindoazkiguke*, *zindoazkigukeen*, *zindoazkigukete*, *zindoazkiguketen*, *zindoazkigun*, *zindoazkiguten*, *zindoazkioke*, *zindoazkiokeen*, *zindoazkiokete*, *zindoazkioketen*, *zindoazkion*, *zindoazkioten*, *zindoazten*, *zoaz*, *zoazke*, *zoazkete*, *zoazkidake*, *zoazkidakete*, *zoazkidate*, *zoazkie*, *zoazkieke*, *zoazkiekete*, *zoazkiete*, *zoazkigu*, *zoazkiguke*, *zoazkigukete*, *zoazkigute*, *zoazkio*, *zoazkioke*, *zoazkiokete*, *zoazkiote*, *zoazkit*, *zoazte* — “go”.

Example 9.6.2. In Fig. S14, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).



(a)



(b)

Fig. S14. Text mining in Basque. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Basque version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{eu})$ between selected topics in English and Basque versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.

9.2 Modified Porter stemming algorithm for Korean

We will be interested in analyzing Modern Korean literary works, written according to the South Korean standard after 1971. This means that texts are spelt out in the native Korean alphabet (known as *Hangul*), with spaces between words.[108] Each *Hangul* block represents exactly one syllable (see Algorithm 9.8 below for romanization of *Hangul*), hence we can use the "uniform reading speed" approximation (§**??**).[109] We are not going to deal with an archaic mixed script style (where Sino-Korean vocabulary is spelt out with *Hanja*, the Chinese characters used in the Korean language), which is mostly restricted to academic and legal settings nowadays.[110]

In a Modern Korean translation (see Table S1 for text source) of *Pride and Prejudice*, the opening line reads

재산이 많은 미혼 남성이라면 반드시 아내를 필요로 한다는 말은 널리 인정되는 진리이다.

Its structure can be analyzed as follows:

Hangul:	재산이	많은	미혼	남성이라면	반드시	아내를
Mixed:	財産이	많은	未婚	男性이라면	반드시	아내를
Romanization:	jaesan-i	manh-eun	mihon	namseong-ilamyeon	bandeusi	anae-leul
Gloss:	fortune.NOM	much.PRS.DET	unmarried	male.be.COND	certainly	wife.ACC

Hangul:	필요로	한다는	말은	널리	인정되는	진리이다
Mixed:	必要로	한다는	말은	널리	認定되는	眞理이다
Romanization:	pilyo-lo	handa-neun	mal-eun	neolli	injeong-doeneun	jinli-ida
Gloss:	necessity.INS	do.IND.NPST.DET	word.TOP	widely	affirm.PAS.PRS.DET	truth.be.IND.NPST

where word stems are shown in red, and various functions of the suffixes include ACC = accusative, COND = conditional, DET = determiner, INS = instrumental, NOM = nominative, NPST = non-past, PAS = passive, PRS = present, TOP = topic. In principle, we want to automatically remove all kinds of grammatical suffixes from a Korean word, no matter the remaining stem is Sino-Korean (*i.e.* of ultimately Chinese origin) or not. In practice, we must admit to certain limitations in our current technology:

- Native Korean stems are sometimes subject to consonant changes. For example, 널리 neolli "widely", 넓다 neolbda "to be wide" and 너비 neobi "width" are supposed to share the same stem, but their *Hangul* spellings in the first syllable have subtle differences. We will try our best to accommodate to such irregular Korean inflections, at the cost of misclassifying certain native Korean words.
- While Sino-Korean stems are immutable in inflections, they suffer from another serious problem: polysemy. Many words with distinguished pronunciations in Chinese become (nearly) homophonic[111] when adapted into Korean, so the disuse of Chinese characters in Korean writing causes ambiguities. Such pairs of *Hangul* homographs [and also (near) homophones] include 인정(認定) injeong "acknowledgement" vs. 인정(人情) injeong "humanity"; 화장(化粧) hwajang "cosmetics" vs. 화장(火葬) hwajang "cremation" etc. We can do nothing to solve such polysemy puzzles in our current work.
- Some native Korean words are spelt in the same way as Sino-Korean words, despite having very different meanings.[112] For example, we have 거리 geoli "road, street" vs. 거리(距離) geoli "distance"; 시장 sijang "hunger" vs. 시장(市場) sijang "market" and 시장(市長) sijang "mayor" etc. Again, we can do nothing right now, as we face Korean documents free from Chinese characters.
- Depending on context (which determines delicate shades of meaning), a given English word might be translated into either a native Korean word or a Sino-Korean word, with drastically different spellings. For example, both 마음 maeum and 심(心) sim mean "heart" in Korean, but there is no simple way to build such links algorithmically.[113] Traditionally, Koreans would learn each *Hanja* by juxtaposing its native Korean gloss with its Sino-Korean pronunciation, in a formula like 心 마음 심. For technical reasons, we have chosen not to feed our computer a long list of *Hanja* readings for rote learning.

[108] This distinguishes Korean texts from Chinese and Japanese ones. In the latter two cases, word segmentation is a non-trivial task.

[109] Vowel length is not marked in Korean spelling. Its effects on reading speed are ignored in our analysis.

[110] When Korean is written in the mixed script, rules for word segmentation are also somewhat different from the pure *Hangul* style. The current North Korean standard does not employ Chinese characters, but its spelling/spacing rules for Sino-Korean vocabulary are different from the practice in South Korea.

[111] For near homophones, we allow differences in vowel lengths that are not marked in *Hangul* spelling.

[112] There is a similar problem for words transliterated from English. For example, 다시 dasi could be a *Hangul* version of "Darcy" or a native Korean word meaning "again"; 제인 jein could be a *Hangul* version of "Jane" or a Sino-Korean word 제인(諸人) "all people". In the present study, we draw on neighboring words [cf. Algorithm 3.25(4)] to decide whether 다시 means "again" (which is to be removed as a stop word), and regard all words in the form of 제인**X** as derivatives of "Jane".

[113] There is a similar situation in Japanese, where the Chinese character 心 can be pronounced either こころ kokoro (native Japanese) or しん shin (Sino-Japanese). In Japanese, "heart" is always written out using the *kanji* 心, irrespective of pronunciation. Thus in text processing, it is trivial to identify native Japanese and Sino-Japanese versions of the same concept. However, even when Korean is written in the mixed script style, only the Sino-Korean vocabulary can be spelt out with *Hanja*, which typographically obscures the connection between native Korean and Sino-Korean counterparts.

Since the phonetic writing system of Modern Korean is a compromise between accuracy and efficiency, so will be our approximate clustering algorithm based on [45].

9.2.1 Stop words and transliterations

Korean stop words can carry various agglutinative suffixes. Instead of enumerating all of them in a single list, we will define several string patterns that match the criterion for stop words.

Definition 9.7 (Korean stop words)**.** Define the pattern **beKorean** as any one from the following list:

> 계세요, 계셔, 계셔서, 계셔야, 계셔요, 계셨기, 계셨느냐, 계셨다, 계셨습니까, 계셨습니다, 계셨어, 계셨어요, 계셨음, 계시겠다, 계시겠습니다, 계시겠어, 계시겠어요, 계시고, 계시기, 계시느냐, 계시는, 계시는데, 계시니, 계시니까, 계시다, 계시더니, 계시려고, 계시면, 계시지만, 계신, 계신다, 계실, 계심, 계십니까, 계십니다, 계십시오, 아녀, 아녀서, 아녀야, 아녀요, 아녔기, 아녔냐, 아녔다, 아녔습니까, 아녔습니다, 아녔어, 아녔어요, 아녔음, 아니겠다, 아니겠습니다, 아니겠어, 아니겠어요, 아니고, 아니기, 아니냐, 아니니, 아니니까, 아니다, 아니다, 아니더니, 아니면, 아니세요, 아니셔, 아니셔서, 아니셔야, 아니셔요, 아니셨기, 아니셨냐, 아니셨다, 아니셨습니까, 아니셨습니다, 아니셨어, 아니셨어요, 아니셨음, 아니시겠다, 아니시겠습니다, 아니시겠어, 아니시겠어요, 아니시고, 아니시기, 아니시냐, 아니시니, 아니시니까, 아니시다, 아니시더니, 아니시면, 아니시지만, 아니신, 아니신데, 아니실, 아니심, 아니십니까, 아니십니다, 아니어서, 아니어야, 아니어요, 아니었기, 아니었음, 아니지만, 아닌, 아닌데, 아닐, 아님, 아닙니까, 아닙니다, 였기, 였느냐, 였다, 였습니까, 였습니다, 였어, 였어요, 였음, 예요, 이겠다, 이겠습니다, 이겠어, 이겠어요, 이고, 이기, 이냐, 이니, 이니까, 이다, 이다, 이더니, 이면, 이세요, 이셔, 이셔서, 이셔야, 이셔요, 이셨기, 이셨냐, 이셨다, 이셨습니까, 이셨습니다, 이셨어, 이셨어요, 이셨음, 이시겠다, 이시겠습니다, 이시겠어, 이시겠어요, 이시고, 이시기, 이시냐, 이시니, 이시니까, 이시다, 이시더니, 이시면, 이시지만, 이신, 이신데, 이실, 이심, 이십니까, 이십니다, 이어서, 이어야, 이었기, 이었느냐, 이었다, 이었습니까, 이었습니다, 이었어, 이었어요, 이었음, 이에, 이에요, 이지만, 인데, 입니까, 입니다, 있겠다, 있겠습니다, 있겠어, 있겠어요, 있고, 있기, 있느냐, 있는, 있는데, 있다, 있다, 있더니, 있습니까, 있습니다, 있어, 있어서, 있어야, 있어요, 있었기, 있었느냐, 있었다, 있었습니까, 있었습니다, 있었어, 있었어요, 있었음, 있으니, 있으니까, 있으면, 있으세요, 있으셔, 있으셔서, 있으셔야, 있으셔요, 있으셨기, 있으셨냐, 있으셨다, 있으셨습니까, 있으셨습니다, 있으셨어, 있으셨어요, 있으셨음, 있으시겠다, 있으시겠습니다, 있으시겠어, 있으시겠어요, 있으시고, 있으시기, 있으시냐, 있으시니, 있으시니까, 있으시다, 있으시더니, 있으시면, 있으시지만, 있으신, 있으신데, 있으실, 있으심, 있으십니까, 있으십니다, 있을, 있음, 있지만.

Define the pattern **notbeKorean** as any one from the following list:

> 없겠다, 없겠습니다, 없겠어, 없겠어요, 없고, 없기, 없느냐, 없는, 없는데, 없다, 없다, 없더니, 없습니까, 없습니다, 없어, 없어서, 없어야, 없어요, 없었기, 없었느냐, 없었다, 없었습니까, 없었습니다, 없었어, 없었어요, 없었음, 없으니, 없으니까, 없으면, 없으세요, 없으셔, 없으셔서, 없으셔야, 없으셔요, 없으셨기, 없으셨냐, 없으셨다, 없으셨습니까, 없으셨습니다, 없으셨어, 없으셨어요, 없으셨음, 없으시겠다, 없으시겠습니다, 없으시겠어, 없으시겠어요, 없으시고, 없으시기, 없으시냐, 없으시니, 없으시니까, 없으시다, 없으시더니, 없으시면, 없으시지만, 없으신, 없으신데, 없으실, 없으심, 없으십니까, 없으십니다, 없을, 없음, 없지만.

Define the pattern **becomeKorean** as any one from the following list:

> 돼, 돼라, 돼서, 돼야, 돼요, 됐기, 됐느냐, 됐다, 됐습니까, 됐습니다, 됐어, 됐어요, 됐음, 되, 되겠다, 되겠습니다, 되겠어, 되겠어요, 되고, 되기, 되느냐, 되는, 되는데, 되니, 되니, 되니까, 되다, 되더니, 되려고, 되면, 되세요, 되셔, 되셔서, 되셔야, 되셔요, 되셨기, 되셨느냐, 되셨다, 되셨습니까, 되셨습니다, 되셨어, 되셨어요, 되셨음, 되시겠다, 되시겠습니다, 되시겠어, 되시겠어요, 되시고, 되시기, 되시느냐, 되시는, 되시는데, 되시니, 되시니까, 되시더니, 되시라, 되시려고, 되시면, 되시지만, 되신, 되신다, 되실, 되심, 되십니까, 되십니다, 되십시오, 되어, 되어라, 되어서, 되어야, 되어요, 되었기, 되었음, 되자, 되지만, 된, 된다, 될, 됨, 됩니까, 됩니다, 됩시다, 됩시오.

Define the pattern **haveKorean** as any one from the following list:

> 가져라, 가져서, 가져야, 가져요, 가졌기, 가졌느냐, 가졌다, 가졌습니까, 가졌습니다, 가졌어, 가졌어요, 가졌음, 가지겠다, 가지겠습니다, 가지겠어, 가지겠어요, 가지고, 가지기, 가지느냐,

가지는, 가지는데, 가지니, 가지니까, 가지다, 가지더니, 가지려고, 가지면, 가지세요, 가지셔, 가지셔서, 가지셔야, 가지셔요, 가지셨기, 가지셨느냐, 가지셨다, 가지셨습니까, 가지셨습니다, 가지셨어, 가지셨어요, 가지셨음, 가지시겠다, 가지시겠습니다, 가지시겠어, 가지시겠어요, 가지시고, 가지시기, 가지시느냐, 가지시는, 가지시는데, 가지시니, 가지시니까, 가지시더니, 가지시라, 가지시려고, 가지시면, 가지시지만, 가지신, 가지신다, 가지실, 가지심, 가지십니까, 가지십니다, 가지십시오, 가지어라, 가지어서, 가지어야, 가지어요, 가지었기, 가지었음, 가지자, 가지지만, 가진, 가진다, 가질, 가짐, 가집니까, 가집니다, 가집시다, 가집시오, 갖겠다, 갖겠습니다, 갖겠어, 갖겠어요, 갖고, 갖기, 갖느냐, 갖는, 갖는다, 갖는데, 갖더니, 갖습니까, 갖습니다, 갖아, 갖아라, 갖아서, 갖아야, 갖아요, 갖았기, 갖았느냐, 갖았다, 갖았습니까, 갖았습니다, 갖았어, 갖았어요, 갖았음, 갖으니, 갖으니까, 갖으려고, 갖으면, 갖으세요, 갖으셔, 갖으셔서, 갖으셔야, 갖으셔요, 갖으셨기, 갖으셨느냐, 갖으셨다, 갖으셨습니까, 갖으셨습니다, 갖으셨어, 갖으셨어요, 갖으셨음, 갖으시겠다, 갖으시겠습니다, 갖으시겠어, 갖으시겠어요, 갖으시고, 갖으시기, 갖으시느냐, 갖으시는, 갖으시는데, 갖으시니, 갖으시니까, 갖으시더니, 갖으시라, 갖으시려고, 갖으시면, 갖으시지만, 갖으신, 갖으신다, 갖으실, 갖으심, 갖으십니까, 갖으십니다, 갖으십시오, 갖은, 갖을, 갖음, 갖읍시다, 갖읍시오, 갖자, 갖지만.

Define the pattern **KoreanStopRoot** as any one from the following list:

가운데, 가져, 걔, 거기, 거니, 거라, 거야, 거였, 거예, 거의, 건너편, 겁니, 곳, 귀하, 그러나, 꼬나겠, 꼬나느, 꼬나는, 꼬나니, 꼬나더, 꼬나려, 꼬나세, 꼬나셔, 꼬나셨, 꼬나시, 꼬나신, 꼬나십, 꼬나자, 꼬나지, 꼬납니, 꼬납시, 꼬났, 꼬보, 꼬본, 꼬볼, 꼬봄, 꼬봅, 꼬봐, 꼬봤, 나네, 나도, 나라면, 나로, 나를, 나만, 나보다, 나중, 나처, 나한테, 난다, 남게, 남는, 남들, 남에, 남은, 남을, 남의, 남한테, 내게, 내겐, 너도, 너를, 너만, 너만큼, 너무, 너에, 너와, 너한, 너희, 누구, 누군가, 다음, 당신, 대개, 대신, 대하여, 덕분, 도중, 되겠, 되나, 되던, 되려, 되리, 되었, 되지, 뒤에, 들끼리, 때까, 때는, 때도, 때때로, 때로, 때마다, 때만, 때만큼, 때면, 때문, 때보다, 때부터, 때에, 때조차, 라고, 마찬가지, 만큼, 많, 몇, 모두, 모든, 못하, 못했, 무렵, 무엇, 무척, 뭐, 뭔, 뭘, 뭣, 바에, 별로, 본인, 불과, 뿐이, 뿐입, 속으로, 수시로, 스스로, 심지어, 십상이, 아니, 아니요, 아무, 아직, 안에, 안으로, 안은, 안을, 않, 얘들, 얘야, 어느, 어디, 어딘, 어딜, 어때, 어땠, 어떠, 어떤, 어떨, 어떰, 어떻, 어쩌, 어쩔, 어쩜, 언제, 얼마, 없게, 없겠, 없었, 없으, 없이, 없지, 여러, 여부, 여전, 였, 예가, 오로지, 왜냐, 외치곤, 우리, 위까, 위로, 위를, 위부터, 위에, 위의, 이라고, 이래, 이랬, 이러, 이런, 이럴, 이럼, 이렇, 이로, 이리, 이만, 이분, 이와, 이외, 이전, 이제, 이쯤, 이후, 일도, 일쑤, 일은, 일을, 일이, 일인, 일하, 있, 자기, 자네, 자신, 자체, 저기, 저까, 저로, 저만큼, 저처럼, 저한테, 저희, 전과, 전까, 전보다, 전부터, 전에, 전으로, 전쯤, 전혀, 절대, 정도, 제가, 제게, 중에, 중이, 지금, 지난, 지마느, 지마세, 지마셨, 지마시, 지마십, 지말겠, 지말더, 지말려, 지말았, 지말지, 지맙니, 지맙시, 차라리, 최소, 텐데, 하거, 하겠, 하고, 하곤, 하기, 하긴, 하나, 하네, 하는, 하니, 하다, 하더, 하던, 하든, 하려, 하며, 하면, 하면서, 하세, 하셨, 하시, 하여, 하지, 한다, 한데, 할지, 함께, 해서, 해야, 해왔, 해주, 해준, 해줄, 했, 현재, 혹시, 혹은, 혼자, 후에, 훨씬.

Define the pattern **KoreanStopString** as any one from the following list:

가끔, 가장, 간에, 갖게, 갖은, 같이, 걔, 거, 거고, 거나, 거는, 거닌, 거다, 거면, 거야, 거에요, 거예요, 거요, 건, 걸, 걸로, 겁니다, 것이다, 것입니다, 게, 게다가, 결코, 겸, 고로, 곧, 곧잘, 그, 그대, 그래서, 그러나, 그러면, 그리, 그리고, 그리도, 그만, 그만큼, 금방, 기껏해야, 꼬나, 꼬나고, 꼬나기, 꼬나는, 꼬나니, 꼬나라, 꼬나면, 꼬나서, 꼬나셔, 꼬나신, 꼬나실, 꼬나심, 꼬나야, 꼬나요, 꼬나자, 꼬난, 꼬난다, 꼬날, 꼬남, 나, 나는, 나에, 나와, 나의, 난, 남, 내, 내가, 내게, 내내, 내로, 너, 너는, 너무, 너의, 널, 네, 네가, 누가, 늘, 다, 다들, 다소, 대로, 대체로, 댁, 댁은, 더, 더구나, 더욱, 덜, 데, 데가, 데서, 도로, 동안, 되어, 된, 둥, 듯, 듯도, 따라, 따라서, 따름, 딴, 때, 땼씿, 또, 라도, 리, 리가, 마, 마나, 마니, 만, 매우, 맨, 먼저, 모든, 못, 무릇, 무슨, 미리, 바람에, 밖에, 번째, 벌써, 뿐, 사이, 사이에, 서로, 서로서로, 수, 수가, 수는, 아니면, 아마, 아마도, 아주, 안, 앞으로, 얘, 어느, 어따, 어떤, 어떻게, 어차피, 언제나, 얼마나, 없, 여기, 옆에, 예, 오히려, 온, 온갖, 왜, 웬, 이, 이는, 이들, 이래, 이로, 이를, 이리, 이미, 이번, 이상, 이상는, 이상에, 이상으로, 이상은, 이상을, 이상의, 이상이, 일, 자, 자넨, 자넬, 자주, 잘, 쟤, 저, 저기, 저나, 저는, 저도, 저를, 저의, 적, 적이, 전, 전에, 전의, 제, 조금, 좀, 좀처럼, 좀체, 줄, 즉, 지, 지가, 지는, 지마는, 지마니, 지마셔, 지마신, 지마실, 지마심, 지만, 지말, 지말고, 지말기, 지말면, 지말아, 지말자, 지맒, 짝이, 쪽, 채, 채로, 퍽, 하고, 하여금, 하지만, 한, 한번, 한번도, 해, 해도, 해라, 해요, 했어, 후, 후로, 흔히.

Define **doKorean** as the following pattern:

(하|하거|하게|하겠다|하겠습니다|하겠어|하겠어요|하고|하곤|하기|하긴|하느냐|하는|하는데|하니|하니까|하다|하더니|하던|하든|하려|하려고|하면|하세요|하셔|하셔서|하셔야|하셔요|하셨기|하셨느냐|하셨다|하셨습니까|하셨습니다|하셨어|하셨어요|하셨음|하시|하시겠다|하시겠습니다|하시겠어|하시겠어요|하시고|하시기|하시느냐|하시는|하시는데|하시니|하시니까|하시더니|하시라|하시려고|하시면|하시지만|하신|하신다|하실|하심|하십니까|하십니다|하십시오|하여라|하여서|하여야|하였기|하였음|하자|하지만|한|한다|할|함|합니까|합니다|합시다|합시오|해|해라|해서|해야|해요|했|했기|했느냐|했다|했습니까|했습니다|했어|했어요|했음)(∅|까|요)$_m$.

Define **DemPronKorean** as the following pattern:

(그(려|렸|리|림|립|만|쳐|쳤|치|친|칠|침|칩)|(∅|이|저)(것|곳|보다|쪽)|(이|저)(거|건|걸|게|들|때|번))**X**.

If a word exactly matches any one of the following:

(**beKorean**|**becomeKorean**|**haveKorean**|**KoreanStopRoot**)**X**,
KoreanStopString,
(∅|관|못|인)(**doKorean**)**X**,
DemPronKorean,
(전|중)**beKorean**,

then we consider it a Korean stop word. All the Korean stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

To accommodate to consonant changes in Korean inflections, we need to convert between *Hangul* and its romanized form.

Algorithm 9.8 (Romanization of *Hangul*)**.** *We transliterate each individual Hangul syllable as follows:*[114]

(1) *Replace the initial Hangul letter by*

ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
g	*kk*	*n*	*d*	*tt*	*l*	*m*	*b*	*pp*	*s*	*ss*	∅	*j*	*jj*	*ch*	*k*	*t*	*p*	*h*

(2) *Replace the medial Hangul vowel letter by*

ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ
a	*ae*	*ya*	*yae*	*eo*	*e*	*yeo*	*ye*	*o*	*wa*	*wae*	*oe*	*yo*	*u*	*wo*	*we*	*wi*	*yu*	*eu*	*ui*	*i*

(3) *Replace the final Hangul consonant letter (if present) by*

ㄱ	ㄲ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅆ	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ	ㄾ	ㅄ	ㄺ	ㄵ	ㄻ	ㄼ	ㄽ	ㅀ
g	*kk*	*n*	*d*	*l*	*m*	*b*	*s*	*ss*	*ng*	*j*	*ch*	*k*	*t*	*p*	*h*	*lh*	*bs*	*gs*	*nj*	*nh*	*lg*	*lm*	*lb*	*ls*	*lt*	*lp*

For a string $\hat{\sigma}$ written in the Hangul alphabet, we construct HgRom($\hat{\sigma}$) *in three steps:*

(1′) *Transliterate each syllable according to the aforementioned rules.*

(2′) *Insert letter Q as the syllable separator.*

(3′) *Convert double letters into their capitalized equivalents.*

The reverse operation RomHg *is defined so as to ensure that* $\hat{\sigma}$ = RomHg(HgRom($\hat{\sigma}$)).

Example 9.8.1. We have HgRom(밖에) = *baKQe* and HgRom(한글) = *hanQgeul*.

[114] As mentioned before, we concern ourselves with Modern Korean spelling, which eliminates some consonant and vowel forms in Middle Korean: such as 끼 kki "meal" (from earlier 뻐 pski), 따름 ttaleum "only, just" (from earlier 뚜룸 stolom), 땅 ttang "soil, land" (from earlier 짜 sta), 때 ttae "time" (from earlier 뻬 pstae).

9.2.2 Effective spelling and essential root

Like Japanese, the Korean language has two separate numeral systems, one native and one Chinese, which appear in different contexts. The following algorithm handles this special feature of Korean.

Algorithm 9.9 (Clustering of Korean numerals)**.** *Define* **CntKorean** *as the following pattern:*

(가지|개|대|마리|명|번|분|사람|잔|장|채)(∅|가|과|와|의|이).

For a string $\hat{\sigma}$*, we define* KorNum($\hat{\sigma}$) *by the following procedures:*

(1) *Check whether* $\hat{\sigma}$ *matches exactly the pattern* (∅|제)(**D**|구|넷|다섯|두|둘|삼|세|셋|십|아홉|어섯|여덟|열|육|이|일|일곱|칠|팔|하나|한)$_m$(∅|**CntKorean**|시|째)*, where* **D** = (0|1|2|3|4|5|6|7|8|9)*. If no, define* KorNum($\hat{\sigma}$) = $\hat{\sigma}$ *and quit; if yes, go to the next step.*

(2) *Work on* $\hat{\sigma}$ *as follows:*

(2.1) *Do* ~(시|째|**CntKorean**) → ∅, 제~ → ∅.

(2.2) *Do* ~(구|아홉) → ’9, ~넷 → ’4, ~다섯 → ’5, ~(두|둘|이) → ’2, ~(삼|세|셋) → ’3, ~(십|열) → ’0, ~(어섯|육) → ’6, ~(여덟|팔) → ’8, ~(하나|한) → ’1, ~(일곱|칠) → ’7.

(2.3) *Repeat the last step.*

(2.4) *Do* ’ → ∅.

(2.5) *Call the result so far as* $\hat{\sigma}'$*, and define* KorNum($\hat{\sigma}$) =쑸$\hat{\sigma}'$.

Algorithm 9.10 (Regularization of Korean verbs)**.** *For a Korean word* $\hat{\sigma}$*, its “regularized verb form”* RegVb($\hat{\sigma}$) *is defined through the following procedures:*

(1) *If* $\hat{\sigma}$ *contains* 없 *and does not start with* 없*, then replace it by* 렜$\hat{\sigma}$*; otherwise, leave it intact.*

(2) *Replace*

(∅|새|시)(파|퍼)(라|란|랗|래|랬|러|런|렇|레|렜)**X**~			(거|까|꺼|꺼)(마|만|맣|매|맸|머|먼|멓|메|멨)**X**~	
파랗			까맣	
(검|검정|깜장)~	(노|누)(라|란|랑|랗|래|랬|러|런|렁|렇|레|렜)**X**~		(발|벌|빨|뻘)(가|간|갛|개|갰|거|건|겋|게|겠)~	
까맣	노랗		발갛	
(보|뽀)(야|얀|얗|애|얬)**X**~	(부|뿌)(여|연|옇|예|옜)**X**~		(하|허)(야|얀|양|얗|애|얬|여|연|영|옇|예|옜)**X**~	
뽀얗	뽀얗		하얗	
X$^{\in}$(곱|뽑|수줍|씹|입|잡|접|좁)~	**X**$^{\in}$(낳|넣|닿|찧)~		**X**$^{\in}$(닫|돋|쏟|얻)~	**X**$^{\in}$(빗|빼앗|솟|씻)~
X뽔	**X**흻		**X**뜥	**X**씆
그(쳐|쳤|치|친|칠|침|칩)~	그만~		노래(**doKorean**|가|는|로|를|에|와|의)**X**~	
그칮	그맗		노럒	
돌아(보|본|볼|봄|봅|봐|봤)**X**~			만(나|난|날|남|납|났)**X**~	붉~
렛뽔다			뮕	발갛
여(느|는|는데|니|세|셔|셨|시|실|심|십)~			열(겠|고|기|더니|면|어|었|지만)~	
엽			엽	
일(은|을)~	허영심~	(연다|엶)	건(∅|다)	이름(∅|은|을)
일하다	허용씸	엽	걸다	성함

(3) *Do* 마(셔|셨|시|신|실|심|십)~ → 마씎, $\hat{\chi}$어(보|본|볼|봄|봅|봐|봤)**X**~ → $\hat{\chi}'$다*, where* $\hat{\chi}'$ *derives from* $\hat{\chi}$*, after replacing the latter’s final* ㄹ *(if present) with* ㄷ.[115]

Algorithm 9.11 (Korean normalization)**.** *For a Korean word* $\hat{\sigma}$*, its normalized form* KN($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do* **X**$^{\in}$(가게|가구|거실|구실|소리|시시|시집|일시|포도|포함)~ → **X**핦 *on* RegVb($\hat{\sigma}$).[116]

[115] In practice, such a replacement needs both HgRom and RomHg to take effect.

[116] The second syllable in many Sino-Korean stems may resemble a Korean suffix. To avoid misidentification of suffixes, we need to protect such stems by a blocker syllable 핦.

(2) *Replace*

(∅|맏)(아드님|아들)~ 얷쏊			(가르|가른|가를|가름|가릅|갈라|갈랐)~ 컽톹			
(다|아|어)(보(이)|본|볼|봄|봅|봐|봤(자))**X**~ 틡뜅				(매일|하루)~ 날	(묻|물었|물으)~ 몯뜰	
(벼서|벼야|벼요|볐기|볐냐|볐다|볐습|볐어|볐음|빈데|빕니)**X**~ 옒틽					(봬|뵀|뵈|뵌|뵐|뵘|뵙)~ 보	
(이르|이른|이를|이름|이릅|일러|일렀)**X**~ 릀읉				가(려(고)|렸|리|린|릴|림|립)~ 픸윝		갈~ 가다
결혼식~ 결혼썪	경(**doKorean**)**X**~ 경쪙	그리워~ ∅		늘어(놓|서|선|설|섬|섭)~ 뇛늟		다(르|른|를|름|릅)**X**~ 띒폝
떨어(져|졌|지|진|질|짐|집)**X**~ 뽍뙾	떨어뜨**X**~ 뙾	메리턴~ 메릝턙	목사관~ 목사괁	물려받~ 잃혩	배우자~ 배웍짳	부인(**doKorean**)**X**~ 뿗잁
비(겠|고|기|냐|니|다|더|면|세|셔|셨|시|신|실|심|십|어|었|우|운|울|움|웁|워|웠|지)~ 옒틽					쓸데~ 씙뗉	쓸쓸(**doKorean**)**X**~ 푍뢵
어(렵)(려(우|운|울|움|워|웠))~ 어럞		어리석~ 어리썪		일으(켜|켰|키|킨|킬|킴|킵)**X**~ 쾾	일이**X**~ 일하다	조지아나~ 조찢앉
친(**doKorean**)**X**~ 친구	침~ 챔	하녀장~ 하녀쨍	햄~ 햶	(벼|비|빈|빌|빔) 옒틽		달(라고|라셔|래) 주다

(3) *If the string so far begins with any one of the following list:*

> 가문, 가치, 강요, 거리, 거만, 게임, 겨우, 겨울, 견해, 결과, 결심, 결함, 경고, 경련, 경쟁, 고개, 고기, 고려, 고문, 고민, 고의, 고집, 곰곰, 과거, 과시, 관대, 관련, 관리, 관심, 구름, 구입, 군대, 권리, 기대, 기만, 기분, 기여, 기운, 기울, 기질, 기혼, 끝내, 나름, 나무, 나이, 날아, 노골, 노출, 단언, 달리, 당당, 대가, 대기, 대기, 대단, 대담, 대면, 대지, 도구, 도도, 도리, 도서, 동기, 동네, 동요, 동의, 두려, 둘러, 드러, 떠올, 런던, 리지, 마디, 마련, 마을, 마음, 마치, 맞이, 머리, 메리, 모습, 목숨, 몹시, 무례, 무시, 무심, 물건, 물러, 물려, 물론, 물리, 미래, 바람, 바보, 반대, 방도, 방심, 방어, 방임, 방자, 방치, 방해, 배려, 배신, 변심, 부대, 부러, 부분, 부여, 부자, 분개, 분들, 불리, 비중, 비참, 사건, 사고, 사과, 사라, 사랑, 사실, 사이, 사치, 상대, 서재, 섬세, 성함, 세대, 세련, 세심, 소개, 소리, 소심, 소중, 속도, 손님, 수고, 수입, 수치, 시골, 시기, 시도, 시선, 식구, 신랑, 신분, 신중, 아래, 아첨, 안심, 암시, 압니, 압시, 양도, 양심, 얼굴, 여름, 여보, 여성, 여인, 여자, 역할, 연구, 연기, 연대, 연습, 연인, 연주, 연출, 열심, 열중, 염려, 예의, 오래, 오만, 오해, 외출, 요구, 요리, 용서, 우려, 우아, 우울, 원래, 원인, 위대, 유감, 유리, 유일, 유지, 의도, 의무, 의문, 의미, 의심, 의아, 의자, 의지, 이유, 이의, 이자, 이치, 이해, 일반, 일어, 일치, 임대, 입구, 자랑, 자리, 자립, 자만, 자세, 자연, 자질, 잠시, 장교, 장군, 장담, 장면, 재산, 적당, 전이, 절대, 점심, 점잖, 정당, 정리, 정보, 정신, 정오, 정중, 제인, 조건, 조심, 조지, 조치, 존재, 존중, 주로, 주요, 주의, 주인, 중대, 중심, 중요, 중재, 지나, 지시, 지인, 지참, 지체, 지출, 직접, 진실, 진심, 진지, 질녀, 집중, 차이, 차지, 책임, 처리, 처신, 처치, 천재, 초과, 초대, 초래, 추구, 축하, 충고, 친구, 태도, 편안, 편지, 피로, 하녀, 하인, 한가, 한심, 합의, 합치, 호감, 호기, 화랑, 화려, 화해, 확고, 확보, 확신, 확실, 확인,

then insert 핝 *as the third syllable.*

(4) *Replace*

(∅|고)남성~ 남쎵	(∅|고)남아~ 뾔갌	(∅|고)남자~ 뭹먅	(∅|고)남편~ 남펹	(∅|맏)(따님|딸)~ 딿톹	(가|나)싶~ 쌩깎
(계수|동서|새언니|시누이|올케|제수|처제|처형|형수)~ 켴쯽				(고모|숙모|이모|작은어머니|큰어머니)**X**~ 쯌쁚	
(금일|오늘)~ 금잁	(길겠|길고|길기|길다|길더|길면|길어|길었|길지|깁니)~ 롱룔				

(남동생|브라더|아우|오라버님|오빠|형님|형제)**X**~ (내|명)일~ (낟|단|란|잔)말(씀|이)~
쁠떢 뫍끉 얛믡

(누나|누이|시스터|언니|여동생|자매)**X**~ (도움|돕)**X**~
찠턲 혪혪

(동서|매부|시동생|시매부|시숙|시아주버니|제부|처남|형부)~ (드|올)(려|렸|리|린|릴|림|립)~
먫쀺 주

(마누라|와이프|집사람)~ (먼저|우선|처음|첫)~ (모르|모른|모를|모름|모릅|몰라|몰랐)~ (미덥|믿)~
아내 풿쓽 앑놄 믨폝

(바로|바르|바른|바를|바름|바릅|발라|발랐)**X**~ (봅|봐|봤)~ (부르|부른|부를|부름|부릅|불러|불렀)~
쓅렱 보 콽곬

(부친|아버지|아빠|어비)~ (숙부|아저씨|외삼촌|친삼촌)**X**~ (써|썼|쓴|쓸|씁)**X**~ (아이|애들|자식)~
빭쯝 꾺쓲 쓰다 첉뜵

(안다|알)~ (얘|이야)기~ (어머니|어미|엄마)~ (이어|이었|이은|이을|이음|잇)~
얉놄 톭킉 모친 잇렧

(자르|자른|자를|자름|자릅|잘라|잘랐)~ (조카딸|질녀)**X**~
짂캍 짂뻮

(추우|추운|추울|추움|추워|추웠|춥겠|춥고|춥기|춥다|춥더|춥습|춥지)~
캝콡

X∈(∅|대)답~ 가르(쳐|쳤|치|친|칠|침|칩)**X**~ 갈(겠|고|더|려|면|아|았|지만)~ 갊~ 감𝒳~
X땊 팇톹 첁 첁 꺫𝒳

강(**doKorean**)**X**~ 걸렸**X**~ 검~ 겁(니|시)**X**~ 경우**X**~
강챵 걸리다 검컴 행거걸다 끵욲

고말(**doKorean**)**X**~ 구(**doKorean**)**X**~ 굳~ 그~ 금~
말하다 큨큐 꾍 끥 김금

기(냐|니|세|셔|셨|시|신|실|심|심|십)~ 길이~ 긺~ 까(보|본|볼|봄|봅|봐|봤)~ 나(가|간|갈|감|갑|갔)**X**~
롱룔 롱룔 롱룔 쌩깎 나가니

나(아|았|으|은|을|음|읍)**X**~ 나(온|올|옴|옵|와|왔)**X**~ 너비**X**~ 놀(라|란|랄|람|랍|랐)**X**~ 눌~
냿 곎왩 넓다 쎮럧 픒렋

느(껴|꼈|끼|낀|낄|낌|낍)~ 다가**X**~ 다시**X**~ 달(**doKorean**)**X**~
느끼 앒뢫 탙찎 릦맃

달(도|만|에|은|이|째|쯤)~ 달(라|랐|리)**X**~ 달라(요)**X**~ 닮~ 담~ 대(**doKorean**)**X**~ 데(려|리)~
월 띒폝 얛츸 똹 땲 팫퍬 데렱

돌아(가|간|갈|감|갑|갔)**X**~ 돌아(오|온|올|옴|옵|와|왔)**X**~ 드(느|는|니|세|셔|셨|시|신|실|심|십)**X**~
렋깎다 렋왂라 뜰뜰다

드(세|셔|셨|시|신|실|심|십)**X**~ 드러(나|난|날|남|납|났)**X**~ 들(러|렀|르|른|를|름|릅)**X**~
먹다 얦픭 뙾뺚

들(렸|리|린|릴|림|립)**X**~ 들려(오|온|올|옴|옵|와|왔)**X**~ 들어(가|간|갈|감|갑|갔)**X**~
듣다 듣다 엓깎다

들어(도|주|준|줄|줌|줍|줘)**X**~ 들어(라|서|야|요)~
듣다 듣다

들어(서𝒳|선|설|섬|섭)**X**~ 들어(오|온|올|옴|옵|와|왔)**X**~ 들었**X**~
엓서다 엓왂라 듣다

듭(니|시)**X**~ 떠(나|난|날|났)~ 뜻~ 리디아~ 마리아~
뜰뜰다 릪뚎 믦닝 리띸얕 맱랉

만(**doKorean**)**X**~ 말(려|렸|리|린|릴|림|립)~ 맘~ 멈~
먅쵺 마 마핥 멏

모(아|았|여|였|으|은|을|음|읍|이|인|일|임|입)~ 몸~
꺭톑 뽍띆

못~	무도회~	미움**X**~	바(라|란|랄|랍|랐)~	받~	받아들~	밤~	법~	벗~
뫴	무또횎	밉	홾	빹팥	빹해들	냻흝	럮껉	뼱

보잘~	비(**doKorean**)**X**~	상(**doKorean**)**X**~	샘~	생각**X**~	세(우|운|울|움|웁|워|웠)**X**~
뿢짨	쾲퍭삆	쌍햫	썚	생깎	쏉

손(실|해)~	숨~	싫~	심(**doKorean**)**X**~	쓸(데|모)**X**~
룄롰	쑱	싫긽	쏊삂	쓖쓖

아(느|는|니|세|셨|시|신|실|심|십)**X**~	앓~	어(려|렸|리|린|릴|림|립)~	어(째|쨌|쩌|쩐|쩔|쩜|쩝)~	업~
얉놄	빩	어럒	어챿	업업

엿~	오(가|간|갈|감|갑|갔)**X**~	옮~	옳~	옷~	웃~	의(∅|거)(**doKorean**)**X**~
얥	콞꾂	뭎픞	랱랱	읭쑋	랚쁲	뺖앆

이(루|룬|룰|룸|룹|뤄)~	일(**doKorean**)**X**~	일라이자~	잃~	입(에|은|을|이)**X**~
얯칲	일이라는	엘리자베스	룼쯗	뫪믐

자(라|란|랄|람|랍|랐)~	잡수**X**~	적(혀|혔|히|힌|힐|힘|힙)**X**~	전(**doKorean**)**X**~	젊~
끍롥	먹다	쓰다	췑띝	쮍

정(**doKorean**)**X**~	좋아(요)**X**~	주기**X**~	주시(**doKorean**)**X**~	준다**X**~	줘~	즐거**X**~	집~
딍딩	럒	주뀌	쫒씎	주	주	즐겁다	찞

차(려|렸|리|린|릴|림|립)~	참~	청(**doKorean**)**X**~	춤~	크리스마스~	탓~
챷릝	챮	첉	떉	앣맺	턗

피(**doKorean**)**X**~	합~	해리엇~	화가$\hat{\chi}$~	힐~	간	(긴|긴데|길)
엎푇	햪	해렐엩	화깎$\hat{\chi}$	횙	가다	롱룔

(∅|매)주(∅|까지|나|만큼|부터|쯤)((에|였|일)**X**)	(거|걸|겋)	(둔|둠)	(든다|들|듦)	(본|본다|볼|봄)
윜옪	행거걸다	킢릞	뜰뜰다	보다

(안|앒)	달	드	들려(∅|야|요)	들어	쓸	아	야	오(게|지)	입	준
얉놄	월	릍뜵	든다	든다	쓰다	항항	썋썋	오면	뫪믐	주

and do (두겠\|두고\|두기\|두느\|두는\|두니\|두더\|두려\|두면\|두세\|두셔\|두셨\|두시\|두신\|두실\|두심\|두십\|두어\|두었\|두자\|두지\|둔다\|둡니\|둡시)~→킢릞, 나오(∅\|겠다\|겠습니다\|겠어\|겠어요\|고\|기\|너라\|느냐\|는\|는데\|니\|니까\|더니\|려고\|면\|세요\|셔\|셔서\|셔야\|셔요\|셨기\|셨느냐\|셨다\|셨습니까\|셨습니다\|셨어\|셨어요\|셨음\|시겠다\|시겠습니다\|시겠어\|시겠어요\|시고\|시기\|시느냐\|시는\|시는데\|시니\|시니까\|시더니\|시라\|시려고\|시면\|시지만\|신\|신다\|실\|심\|십니까\|십니다\|십시\|자\|지만)**X**~→꼚욅, 매$^{\mathbf{X}_1\in}$(년\|달\|번\|월\|일)(∅\|까지\|나\|만큼\|부터\|을\|이\|쯤)((에\|일)**X**)→$\mathbf{X}_1'$, *where* $\mathbf{X}_1'$ *results from doing* 년→연, 달→월 *on* $\mathbf{X}_1$.

(5) *Do* $^{\mathbf{X}\in}$(들어\|부당\|부인\|부추\|상당\|상세\|상실\|상심\|상의\|상인)~→**X**핞.

(6) *Do* (가거라\|가려고\|가면\|가서\|가세요\|가셔\|가셔서\|가셔야\|가셔요\|가시느냐\|가시려고\|가시면\|가신\|가신다\|가실\|가심\|간다)**X**~→가다, (거냐\|거느\|거는\|거니\|거세\|거셔\|거셨\|거시\|거신\|거심\|거십\|건데\|걸겠\|걸고\|걸기\|걸다\|걸더\|걸려\|걸면\|걸자\|걸지\|겁니\|겁시)**X**~→행거걸다, (낫\|좋\|최선)~→냇, (커\|컸\|크\|큰\|클\|큼\|큽)~→쁚, 고말(∅\|야)~→고먩, 꺫(추\|춘\|출\|춤\|춥\|춰\|췄)~→햹횥, 남~→랆랅, 달(려\|렸\|리\|린\|릴\|림\|립)~→햻릻, 들(겠\|려고\|면\|어)**X**~→뜰뜰다, 들(어\|으\|은\|을\|음)**X**~→듣다, 말~→말먍, 오(겠\|고\|기\|너라\|느냐\|는\|니\|더니\|도록\|라\|려고\|면\|세요\|셔\|셔\|셨\|시\|신\|신다\|실\|심\|십니까\|십니다\|십시\|자\|지\|지만)**X**~→와라, 이(어\|었)~→잇, 입~→읪, 점~→쪔점, 주(게\|까)**X**~→주, (나가\|나간)→꼚욅, (온\|온다\|올\|옴\|옵니까\|옵니다\|옵시다\|옵시오\|와\|와서)→와라.

Algorithm 9.12 (Korean effective spelling)**.** *Set* **V** = (*a*\|*e*\|*i*\|*o*\|*u*\|*y*)*. For a Korean word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do* $\hat{\chi}$한테**X**→$\hat{\chi}$ *before applying Korean normalization.*

(2) *Do* 리핞→릮핞, 하핞→햫핞, 한핞→햱핞, 할핞→햺핞, 함핞→햶핞, 해핞→햸핞.

(3) *Do* $\hat{\chi}$(**beKorean**|**becomeKorean**|(∅\|당)**doKorean**)**X**→$\hat{\chi}$.

(4) *Define* **KoreanNounSuff** *as the pattern* (∅\|건\|걸\|것\|게\|쪽)(∅\|들\|뿐\|뿐만)(∅\|가\|과\|랑\|로\|로부터\|로서\|로써\|를\|보다\|아\|에\|에게\|에게서\|에다\|에다가\|에서\|와\|으로\|으로부터\|으로서\|으로써\|을\|의\|이\|이랑\|하고)(∅\|같이\|까지\|까지\|나\|는\|대로\|도\|따라\|마다\|마저\|만\|만큼\|보다\|보다도\|부터\|씩\|야\|야말로\|은\|이나\|조차\|조차도\|쯤\|처럼). *Do* $\hat{\chi}\hat{\chi}$적**X**→$\hat{\chi}\hat{\chi}$ *and* ~$\hat{\chi}\hat{\chi}$(∅\|가\|감\|객\|경\|게\|과\|관\|권\|금\|기\|력\|료\|방\|범\|별\|복\|부\|비\|사\|상\|생\|서\|석\|성\|세\|소\|수\|식\|실\|아\|어\|용\|원\|인\|자\|장\|지\|직\|초\|탕\|판\|편\|품\|학\|형\|화\|회)**KoreanNounSuff**(∅\|커녕)→$\hat{\chi}\hat{\chi}$.

(5) *Do* ~$\hat{\chi}$(∅|감|개|게|껏|꾸러기|꾼|님|둥이|맹이|보|어치|오|우|음|이|쟁이|질|짜리|째|투성이|히)**KoreanNounSuff**(∅|다는데|듯이|라는데|라서|려무나|련마는|렴|망정|수록|야말고|자마|지언정|커녕|테니|테니까|텐데) → $\hat{\chi}$.

(6) *Apply* KorNum.

(7) *Do the following in a sequel, searching for the longest match in each step:*

(7.1) *Remove one of these word-final patterns if it follows a character:*

길래, 끼리, 네요, 다며, 다시피, 담, 되며, 라도, 라며, 라면서, 만, 만은, 며, 면서, 어요, 에서, 에선, 요, 이란, 이며, 이요, 인, 하며.

(7.2) *Remove one of these word-final patterns if it follows a character:*

내, 는, 대, 라면, 래, 를, 은, 을, 이군, 이라뇨, 이라니, 재

(7.3) *Remove one of these word-final patterns if it follows a character:*

가, 까, 끔, 늘, 니와, 더러, 도, 든, 라, 뿐, 아, 아라, 야, 어라, 에, 오, 자

and do ~$\hat{\chi}$(∅|냔|단|란|잔)다 → $\hat{\chi}$, ~$\hat{\chi}$(∅|냤|댔|랬|쟀)(∅|어) → $\hat{\chi}$ *in the same sweep.*

(7.4) *Define* **KoreanVerbSuff** *as any one of the following patterns:*

∅, 가, 같이, 거, 겠, 고, 고말고, 구나, 구만, 구먼, 군요, 기, 기만, 기엔, 까지, 께, 는, 다, 다니, 더니, 데, 도록, 든지, 들, 따라, 려고, 로, 로부터, 로서, 로써, 를, 마다, 마저, 만큼, 면, 보다, 부터, 서, 서부터, 셔, 씩, 야, 어, 에, 에게, 에게서, 에겐, 에다, 에다가, 에서, 에서부터, 으로, 으로부터, 으로서, 으로써, 은, 은데, 을, 음, 의, 이, 자마자, 져서, 조차, 중, 지만, 쯤, 처럼, 처럼, 하곤, 해선, (∅|는|은)커녕, (∅|이)나, (과|랑|와|이랑|하고)(∅|는|만|은), (냐|냡|다|답|라|랍|자|잡)(고|니).

Do $\hat{\chi}$(**KoreanVerbSuff**)$_m$ → $\hat{\chi}$.

(7.5) *Remove one of these word-final patterns if it follows a character:*

게, 네, 는, 니, 더냐, 더라, 던, 데, 디, 세, 시, 어라, 자.

(7.6) *Do* ~$\hat{\chi}$(∅|으)(∅|셨|시|신|실|심|십)(∅|겠|았|았겠|었|었겠)(∅|습|읍) → $\hat{\chi}$.

(7.7) *Remove one of these word-final patterns if it follows a character:*

겨, 겼, 구, 군, 굴, 굼, 굽, 궈, 궜, 기, 긴, 길, 김, 깁, 려, 련, 렸, 리, 릴, 림, 립, 받, 여, 였, 우, 운, 울, 움, 웁, 워, 웠, 이, 인, 일, 임, 입, 져, 졌, 지, 진, 질, 짐, 집, 추, 춘, 출, 춤, 춥, 춰, 췄, 혀, 혔, 히, 힌, 힐, 힘, 힙.

(7.8) *Do* ~$\hat{\chi}$(∅|대로|(겹|답|롭|맞|스러|스럽)(∅|겨 우|겨 운|겨 울|겨 움|겨 웁|겨 워|겨 웠|다우|다운|다울|다움|다웁|다워|다웠|로우|로운|로울|로움|로웁|로워|로웠|우|운|울|움|웁|워|웠)) → $\hat{\chi}$.

(8) *Do* ~$\hat{\chi}$(러|러서|셔서) → $\hat{\chi}$, $\hat{\chi}$(뻐|뻤|쁘|쁜|쁠|쁨|쁩)**X** → $\hat{\chi}$쁘, $\hat{\chi}$(려|리|린|림)**X** → $\hat{\chi}$리.

(9) *Apply* HgRom.

(10) *Do nolQl***X**~ → *nolQ, Q*(∅|*g*|*l*|*s*)**V**$_m$*S***X** → ∅, *QsiQk***X** → ∅, ~(∅|*b*)*Q*(∅|*eub*|*seub*)(∅|*Q*)(∅|*de*|*deo*|*di*|*ge*|*go*|*ja*|*n*|*ne*|*neu*|*ni*|*se*|*si*)(∅|*l*|*m*|*n*) → ∅.

(11) *Apply* RomHg *and do* ~$\hat{\chi}$(∅|냐|시|으시) → $\hat{\chi}$.[117]

(12) *Do* $\hat{\chi}$(느|으)**X** → $\hat{\chi}$.

Algorithm 9.13 (Korean essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Korean word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do* ~$\hat{\chi}$(리|아|아서|어|어서|음|이) → $\hat{\chi}$.

(2) *Apply* HgRom *and do* ~(*b*|*h*|*m*|*s*|*S*) → ∅, ~(*d*|*Qleu*(∅|*b*|*h*|*m*|*n*|*Qleo*|*s*|*S*)) → *l*.

(3) *Apply* RomHg *and do* ~와 → ∅.

[117] To save execution time, one could have manipulated only the romanized strings, without converting back and forth between *Hangul* and its transliteration. However, we choose to state our algorithm using multiple applications of HgRom and RomHg, to enhance the readability of our substitution rules.

9.2.3 Approximate clustering

There will not be any mutation rules in our algorithm for Korean, so SimpHrdTest = HrdTest.

Algorithm 9.14 (Simple heredity test)**.** *Suppose that $\hat{\alpha}$ and $\hat{\beta}$ are both Hangul strings. The Boolean-valued function* SimpHrdTest($\hat{\alpha}$,$\hat{\beta}$) *returns* **TRUE** *if at least one of the following five conditions holds:*[118]

(i) $\hat{\alpha} = \hat{\beta}$;

(ii) $\hat{\beta} = \hat{\alpha}$네;

(iii) HgRom($\hat{\alpha}$)l = HgRom($\hat{\beta}$);

(iv) $\min\{\ell(\hat{\alpha}), \ell(\hat{\beta})\} \geq 2$, **AND** HgRom($\hat{\alpha}$)′ = HgRom($\hat{\beta}$)′, *where the prime denotes an operation* $\sim\mathbf{V}_m(\varnothing|n) \rightarrow \varnothing$;

(v) $\min\{\ell(\hat{\alpha}), \ell(\hat{\beta})\} \geq 2$, **AND** HgRom($\hat{\alpha}$)* = HgRom($\hat{\beta}$)*, *where the asterisk denotes an operation* Q(*jeog*|*man*|*Pun*|*seong*)$\mathbf{X} \rightarrow \varnothing$.

Algorithm 9.15 (Approximate clustering of Korean words)**.** *The approximate clustering of a list of Korean words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1, \text{HgRom}(\text{EffSpell}(\hat{\alpha}_1))), \ldots, (\hat{\alpha}_N, \text{HgRom}(\text{EffSpell}(\hat{\alpha}_N)))\}$ *alphabetically according to the last component. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \text{HgRom}(\text{EffSpell}(\hat{\alpha}_1))), \ldots, (\hat{\alpha}_{(N)}, \text{HgRom}(\text{EffSpell}(\hat{\alpha}_N)))\}$ *satisfy* HrdTest(EffSpell($\hat{\alpha}_{(n)}$),EffSpell($\hat{\alpha}_{(n+1)}$)) = **FALSE**, *where* $n \in \mathbb{Z} \cap [1, N)$, *then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \text{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)}, *)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, *), \ldots, (\hat{\alpha}_{(M,n_M)}, *)\}\}$, *where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$.

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, *), \ldots, (\hat{\alpha}_{(m,n_m)}, *)\}$ *where* $m \in \mathbb{Z} \cap [1, M]$, *we augment it into a tagged entry* $G_m = (g_m, \text{HgRom}(\text{EffSpell}(\hat{\alpha}_{(m,1)})), \text{HgRom}(\text{EssRoot}(\text{EffSpell}(\hat{\alpha}_{(m,1)}))))$. *The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* HgRom(EssRoot(EffSpell($\hat{\alpha}_{(m,1)}$))) *(with higher priority), and* HgRom(EffSpell($\hat{\alpha}_{(m,1)}$)) *(with lower priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)})\}$ *satisfy*

$$\text{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\text{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$, *then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$. *Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots) from each* Γ_k, *where* $k \in \mathbb{Z} \cap [1, K]$.

Example 9.15.1. Our clustering algorithm correctly groups the following families of Korean words:

걷겠다, 걷겠습니다, 걷겠어, 걷겠어요, 걷고, 걷기, 걷느냐, 걷는, 걷는다, 걷는데, 걷더니, 걷습니까, 걷습니다, 걷자, 걷지만, 걸어, 걸어라, 걸어서, 걸어야, 걸어요, 걸었기, 걸었느냐, 걸었다, 걸었습니까, 걸었습니다, 걸었어, 걸었어요, 걸었음, 걸으니, 걸으니까, 걸으려고, 걸으면, 걸으세요, 걸으셔, 걸으셔서, 걸으셔야, 걸으셔요, 걸으셨기, 걸으셨느냐, 걸으셨다, 걸으셨습니까, 걸으셨습니다, 걸으셨어, 걸으셨어요, 걸으셨음, 걸으시겠다, 걸으시겠습니다, 걸으시겠어, 걸으시겠어요, 걸으시고, 걸으시기, 걸으시느냐, 걸으시는, 걸으시는데, 걸으시니, 걸으시니까, 걸으시더니, 걸으시라, 걸으시려고, 걸으시면, 걸으시지만, 걸으신, 걸으신다, 걸으실, 걸으심, 걸으십니까, 걸으십니다, 걸으십시오, 걸은, 걸을, 걸음, 걸읍시다, 걸읍시오 — "walk";

넓겠다, 넓겠습니다, 넓겠어, 넓겠어요, 넓고, 넓기, 넓다, 넓더니, 넓습니까, 넓습니다, 넓어, 넓어서, 넓어야, 넓어요, 넓었기, 넓었냐, 넓었다, 넓었습니까, 넓었습니다, 넓었어, 넓었어요, 넓었음, 넓으냐, 넓으니, 넓으니까, 넓으면, 넓으세요, 넓으셔, 넓으셔서, 넓으셔야, 넓으셔요, 넓으셨기, 넓으셨냐, 넓으셨다, 넓으셨습니까, 넓으셨습니다, 넓으셨어, 넓으셨어요, 넓으셨음, 넓으시겠다, 넓으시겠습니다, 넓으시겠어, 넓으시겠어요, 넓으시고, 넓으시기, 넓으시냐, 넓으시니, 넓으시니까, 넓으시다, 넓으시더니, 넓으시면, 넓으시지만, 넓으신, 넓으신데, 넓으실, 넓으심, 넓으십니까, 넓으십니다, 넓은, 넓은데, 넓을, 넓음, 넓지만 — "wide";

[118] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

오겠다, 오겠습니다, 오겠어, 오겠어요, 오고, 오기, 오너라, 오느냐, 오는, 오는데, 오니, 오니까, 오더니, 오려고, 오면, 오세요, 오셔, 오셔서, 오셔야, 오셔요, 오셨기, 오셨느냐, 오셨다, 오셨습니까, 오셨습니다, 오셨어, 오셨어요, 오셨음, 오시겠다, 오시겠습니다, 오시겠어, 오시겠어요, 오시고, 오시기, 오시느냐, 오시는, 오시는데, 오시니, 오시니까, 오시더니, 오시라, 오시려고, 오시면, 오시지만, 오신, 오신다, 오실, 오심, 오십니까, 오십니다, 오십시오, 오자, 오지만, 온, 온다, 올, 옴, 옵니까, 옵니다, 옵시다, 옵시오, 와, 와라, 와서, 와야, 와요, 왔기, 왔느냐, 왔다, 왔습니까, 왔습니다, 왔어, 왔어요, 왔음 — “come”;

높겠다, 높겠습니다, 높겠어, 높겠어요, 높고, 높기, 높다, 높더니, 높습니까, 높습니다, 높아, 높아서, 높아야, 높아요, 높았기, 높았냐, 높았다, 높았습니까, 높았습니다, 높았어, 높았어요, 높았음, 높으냐, 높으니, 높으니까, 높으면, 높으세요, 높으셔, 높으셔서, 높으셔야, 높으셔요, 높으셨기, 높으셨냐, 높으셨다, 높으셨습니까, 높으셨습니다, 높으셨어, 높으셨어요, 높으셨음, 높으시겠다, 높으시겠습니다, 높으시겠어, 높으시겠어요, 높으시고, 높으시기, 높으시냐, 높으시니, 높으시니까, 높으시다, 높으시더니, 높으시면, 높으시지만, 높으신, 높으신데, 높으실, 높으심, 높으십니까, 높으십니다, 높은, 높은데, 높을, 높음, 높지만 — “tall”;

가거라, 가겠다, 가겠습니다, 가겠어, 가겠어요, 가고, 가기, 가느냐, 가는, 가는데, 가니, 가니까, 가더니, 가라, 가려고, 가면, 가서, 가세요, 가셔, 가셔서, 가셔야, 가셔요, 가셨기, 가셨느냐, 가셨다, 가셨습니까, 가셨습니다, 가셨어, 가셨어요, 가셨음, 가시겠다, 가시겠습니다, 가시겠어, 가시겠어요, 가시고, 가시기, 가시느냐, 가시는, 가시는데, 가시니, 가시니까, 가시더니, 가시라, 가시려고, 가시면, 가시지만, 가신, 가신다, 가실, 가심, 가십니까, 가십니다, 가십시오, 가야, 가요, 가자, 가지만, 간다, 갑니까, 갑니다, 갑시다, 갑시오, 갔기, 갔느냐, 갔다, 갔습니까, 갔습니다, 갔어, 갔어요, 갔음 — “go”;

빗겠다, 빗겠습니다, 빗겠어, 빗겠어요, 빗고, 빗기, 빗느냐, 빗는, 빗는다, 빗는데, 빗더니, 빗습니까, 빗습니다, 빗어, 빗어라, 빗어서, 빗어야, 빗어요, 빗었기, 빗었느냐, 빗었다, 빗었습니까, 빗었습니다, 빗었어, 빗었어요, 빗었음, 빗으니, 빗으니까, 빗으려고, 빗으면, 빗으세요, 빗으셔, 빗으셔서, 빗으셔야, 빗으셔요, 빗으셨기, 빗으셨느냐, 빗으셨다, 빗으셨습니까, 빗으셨습니다, 빗으셨어, 빗으셨어요, 빗으셨음, 빗으시겠다, 빗으시겠습니다, 빗으시겠어, 빗으시겠어요, 빗으시고, 빗으시기, 빗으시느냐, 빗으시는, 빗으시는데, 빗으시니, 빗으시니까, 빗으시더니, 빗으시라, 빗으시려고, 빗으시면, 빗으시지만, 빗으신, 빗으신다, 빗으실, 빗으심, 빗으십니까, 빗으십니다, 빗으십시오, 빗은, 빗을, 빗음, 빗읍시다, 빗읍시오, 빗자, 빗지만 — “comb”;

나아, 나아라, 나아서, 나아야, 나아요, 나았기, 나았느냐, 나았다, 나았습니까, 나았습니다, 나았어, 나았어요, 나았음, 나으니, 나으니까, 나으려고, 나으면, 나으세요, 나으셔, 나으셔서, 나으셔야, 나으셔요, 나으셨기, 나으셨느냐, 나으셨다, 나으셨습니까, 나으셨습니다, 나으셨어, 나으셨어요, 나으셨음, 나으시겠다, 나으시겠습니다, 나으시겠어, 나으시겠어요, 나으시고, 나으시기, 나으시느냐, 나으시는, 나으시는데, 나으시니, 나으시니까, 나으시더니, 나으시라, 나으시려고, 나으시면, 나으시지만, 나으신, 나으신다, 나으실, 나으심, 나으십니까, 나으십니다, 나으십시오, 나은, 나을, 나음, 나읍시다, 나읍시오, 낫겠다, 낫겠습니다, 낫겠어, 낫겠어요, 낫고, 낫기, 낫느냐, 낫는, 낫는다, 낫는데, 낫더니, 낫습니까, 낫습니다, 낫자, 낫지만 — “recover”;

흐르겠다, 흐르겠습니다, 흐르겠어, 흐르겠어요, 흐르고, 흐르기, 흐르느냐, 흐르는, 흐르는데, 흐르니, 흐르니까, 흐르더니, 흐르려고, 흐르면, 흐르세요, 흐르셔, 흐르셔서, 흐르셔야, 흐르셔요, 흐르셨기, 흐르셨느냐, 흐르셨다, 흐르셨습니까, 흐르셨습니다, 흐르셨어, 흐르셨어요, 흐르셨음, 흐르시겠다, 흐르시겠습니다, 흐르시겠어, 흐르시겠어요, 흐르시고, 흐르시기, 흐르시느냐, 흐르시는, 흐르시는데, 흐르시니, 흐르시니까, 흐르시더니, 흐르시라, 흐르시려고, 흐르시면, 흐르시지만, 흐르신, 흐르신다, 흐르실, 흐르심, 흐르십니까, 흐르십니다, 흐르십시오, 흐르자, 흐르지만, 흐른, 흐른다, 흐를, 흐름, 흐릅니까, 흐릅니다, 흐릅시다, 흐릅시오, 흘러, 흘러라, 흘러서, 흘러야, 흘러요, 흘렀기, 흘렀느냐, 흘렀다, 흘렀습니까, 흘렀습니다, 흘렀어, 흘렀어요, 흘렀음 — “flow”.

Moreover, the words 널리 “widely” and 너비 “width” also cluster correctly with the inflected forms of “wide”.

Example 9.15.2. In Fig. S15, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).

Following the standard practice in Korean dictionaries, we have inserted etymological information for loanwords from Chinese[119] and English, in the parentheses. This was done by using a manually built look-up table. The etyma in Chinese

[119]Due to partial conflations of native Korean and Sino-Korean vocabulary (as discussed earlier), a *Hanja* form may not apply to all the words in a certain cluster. Furthermore, since we have merged 숙모(叔母) “paternal aunt” and 이모(姨母) “maternal aunt” into the same cluster, but have not built a stacking mechanism for Chinese characters (for legibility reasons), the *Hanja* forms of such word stems are not displayed.

or English were not directly inferrable from the Korean text under investigation, nor were they an integral part of our word clustering algorithm: we track the etymologies of the word stems so that we can assess the reliability of our experimental methods.

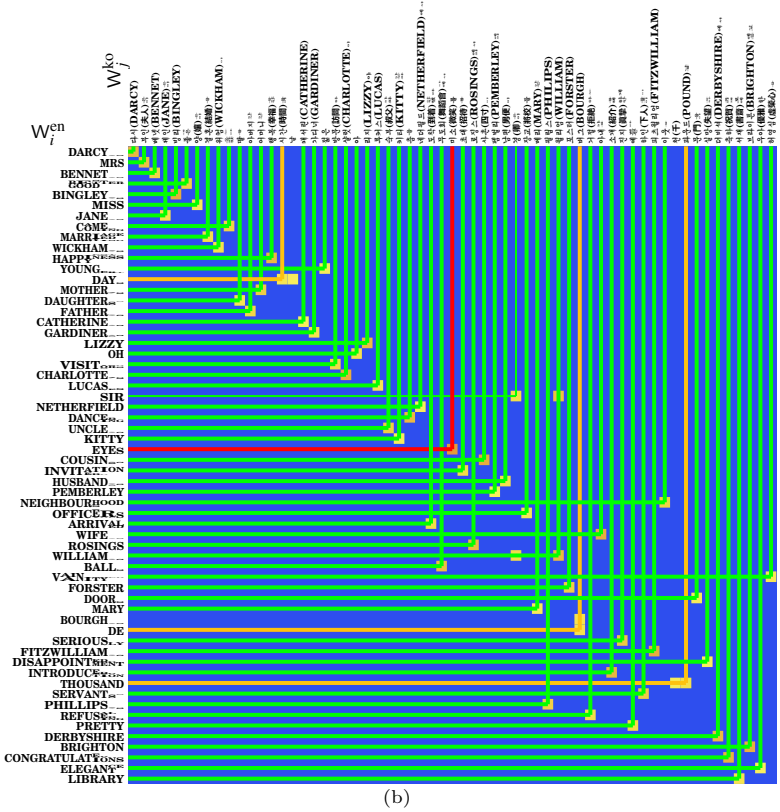
The Korean case shown here offers a telling example that meanings of words are determined by their numerical fingerprints, rather than their etymological sources. Korean words of native, Chinese and English origins all behave in a consistent way in the displayed results.

We caution our readers, however, that the performance of our numerical algorithm is indeed affected by mixed etymological sources of a given concept in Korean. From the control experiments in Fig. S15c, we see that certain correct matching pairs have similarity scores below the 0.7 threshold applicable to Fig. S15b. Part of this can be explained by the coexistence of native Korean and Sino-Korean stems, the latter usually reserved for abstract, formal and polite settings. See the (partial) list below:

English gloss	native Korean stem	Sino-Korean stem
“amiable”	상냥 sangnyang	친절(親切) chinjeol
“feel”	느끼 neukki	감정(感情) gamjeong, 기분(氣分) kibun
“forget”	잊 ij	망각(忘却) manggag
“love”	사랑 salang	애정(愛情) aejeong
“pride”	자랑 jalang	오만(傲慢) oman, 자부심(自負心) jabusim, 거만(倨慢) geoman
“read”	읽 ilg	독서(讀書) dogseo
“walk”	걷 geod	산책(散策) sanchaeg
“wish”	싶 sip	원(願) won

Had we forcibly merged these native Korean stems with their Sino-Korean counterparts, before performing the numerical experiments in Fig. S15b, we would have achieved a slightly better yield for correct matchings with similarity scores above 0.7.

(a)



(b)

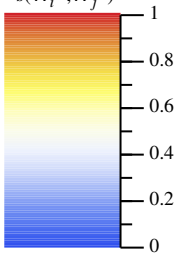
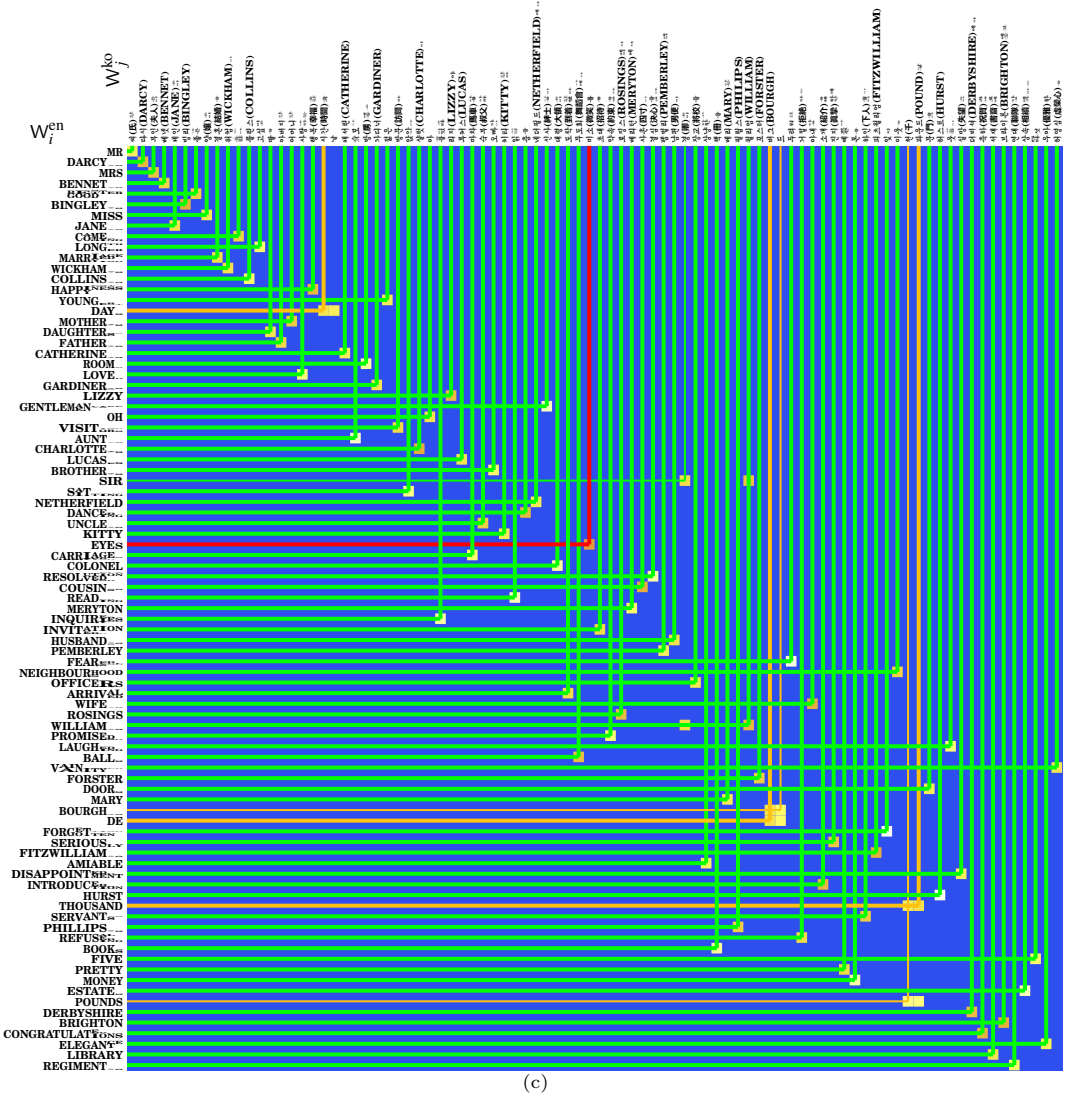


Fig. S15. Text mining in Korean. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Korean version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(W_i^{en}, W_j^{ko})$ between selected topics in English and Korean versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *red*) cross-hair indicates a correct (resp. incorrect) match. *Amber* cross-hair marks a link between distinct concepts that share the same hypernyms.



(c)

Fig. S15. Text mining in Korean. (*Continued*) (c) A control experiment in the spirit of Fig. S8c. The censorship $s(\mathbf{v}_i^{en}, \mathbf{v}_j^{ko}) \geq 0.7$ is not imposed.

9.3 Modified Porter stemming algorithm for Turkish

Our Turkish stemming algorithm is adapted from the method of Gülşen Eryiğit and Eşref Adalı [51], which was previously implemented by Evren (Kapusuz) Çilden (`http://snowball.tartarus.org/algorithms/turkish/stemmer.html`).

Our algorithm works on Modern Turkish (after the Atatürk Reform of 1928), which uses (a modified version of) the Latin alphabet, and phases out most grammatical forms of Arabic and Persian origins. Since Atatürk's Language Reform, a significant amount of Arabic and Persian loanwords have become obsolete, even if they are apparently more concise (using fewer syllables) than their counterparts in Modern Turkish, such as

English gloss	via Proto-Turkic	via Arabic	via Persian
"friend"	*arkadaş*	———	*dost*
"war"	*savaş*	*harp*	*cenk*

While the native Turkish forms *arkadaş* and *savaş* contain more syllables than their Arabic/Persian equivalents, they are free from awkward consonant clusters.[120] Some Arabic/Persian loanwords still survive in Modern Turkish, as they are phonologically compatible with native Turkish words, such as *saat* "hour" (from Arabic) and *zaman* "time, epoch" (from either Arabic or Persian). It is also possible that a native Turkish word coexists with an Arabic/Persian loan for the same concept, with slightly different connotations, such as *ak* "white" (native Turkish) vs. *beyaz* "white" (via Arabic); *kara* "black" (native Turkish) vs. *siyah* "black" (via Persian). The last situation is similar to the coexistence of native Korean and Sino-Korean vocabularies in Modern Korean (see §9.2).

While we design our stemming algorithm for Modern Turkish, we take into account its phonological restrictions on syllable structure. Like what we did for Korean, we are not going to merge native and non-native etyma into the same word cluster.

Turkish stop words can carry various agglutinative suffixes. Instead of enumerating all of them in a single list, we will define several string patterns that match the criterion for stop words.

Definition 9.16 (Turkish stop words)**.** Define the pattern **TurkishStopRoot** as any one from the following list:

> *aksi*, *altında*, *ama*, *anca*, *ancak*, *andan*, *aralarında*, *arasında*, *artık*, *aşağı*, *aşağıda*, *aynı*, *ayrı*, *ayrıca*, *ayrıyeten*, *ayriyeten*, *az*, *bana*, *başına*, *başka*, *bazen*, *bazı*, *bazıları*, *bazılarıda*, *bazılarına*, *bazılarından*, *bazılarını*, *bazılarının*, *belgeyle*, *belki*, *ben*, *bende*, *benden*, *beni*, *benim*, *beri*, *bile*, *bir*, *biraz*, *biri*, *birileri*, *birisi*, *birkaç*, *birlikte*, *biz*, *bizde*, *bizden*, *bize*, *bizi*, *bizim*, *bizler*, *boylece*, *böyle*, *böylece*, *bu*, *buna*, *bunda*, *bundan*, *bunlar*, *bunlara*, *bunlarda*, *bunlardan*, *bunları*, *bunların*, *bunu*, *bunun*, *bununla*, *bura*, *burada*, *buradaki*, *buradan*, *buralar*, *buralara*, *buralarda*, *buralardan*, *buraları*, *buraların*, *buranın*, *buraya*, *burayı*, *bütün*, *civarda*, *çoğu*, *çoğunlukla*, *çok*, *çünkü*, *da*, *daha*, *dahası*, *dair*, *de*, *dek*, *dendenle*, *dışarı*, *dışında*, *diğer*, *diğerleri*, *diye*, *dolayı*, *dolayısıyla*, *durumda*, *e*, *eden*, *eder*, *ederim*, *ederiz*, *ederler*, *edersin*, *edersiniz*, *edildi*, *ediyor*, *ediyordu*, *ediyorduk*, *ediyordum*, *ediyordun*, *ediyordunuz*, *ediyorlar*, *ediyorlardı*, *ediyorsun*, *ediyorsunuz*, *ediyorum*, *ediyoruz*, *eğer*, *en*, *etmedi*, *etmedik*, *etmediler*, *etmedim*, *etmedin*, *etmediniz*, *etmem*, *etmeyiz*, *etmez*, *etmezler*, *etmezsin*, *etmezsiniz*, *etmiyor*, *etmiyordu*, *etmiyorduk*, *etmiyordum*, *etmiyordun*, *etmiyordunuz*, *etmiyorlar*, *etmiyorlardı*, *etmiyorsun*, *etmiyorsunuz*, *etmiyorum*, *etmiyoruz*, *etrafında*, *etsen*, *etti*, *ettik*, *ettiler*, *ettim*, *ettin*, *ettiniz*, *ettirdi*, *ettirdik*, *ettirdiler*, *ettirdim*, *ettirdin*, *ettirdiniz*, *ettireceğim*, *ettireceğiz*, *ettirecek*, *ettirecekler*, *ettireceksin*, *ettireceksiniz*, *ettirir*, *ettirirım*, *ettiririz*, *ettirirler*, *ettirirsin*, *ettirirsiniz*, *ettiriyor*, *ettiriyordu*, *ettiriyorduk*, *ettiriyordum*, *ettiriyordun*, *ettiriyordunuz*, *ettiriyorlar*, *ettiriyorlardı*, *ettiriyorsun*, *ettiriyorsunuz*, *ettiriyorum*, *ettiriyoruz*, *ettirmedi*, *ettirmedik*, *ettirmediler*, *ettirmedim*, *ettirmedin*, *ettirmediniz*, *ettirmem*, *ettirmeyeceğim*, *ettirmeyeceğiz*, *ettirmeyecek*, *ettirmeyecekler*, *ettirmeyeceksin*, *ettirmeyeceksiniz*, *ettirmeyiz*, *ettirmez*, *ettirmezler*, *ettirmezsin*, *ettirmezsiniz*, *ettirmiyor*, *ettirmiyordu*, *ettirmiyorduk*, *ettirmiyordular*, *ettirmiyordum*, *ettirmiyordun*, *ettirmiyordunuz*, *ettirmiyorlar*, *ettirmiyorsun*, *ettirmiyorsunuz*, *ettirmiyorum*, *ettirmiyoruz*, *evet*, *evvel*, *fakat*, *gayet*, *gelen*, *gelişme*, *gereçler*, *gerek*, *gerektiği*, *geride*, *gibi*, *göre*, *günü*, *hadi*, *hakkında*, *hala*, *hâlâ*, *halbuki*, *halde*, *haline*, *hangi*, *hangileri*, *hangilerine*, *hangilerini*, *hangilerinin*, *hangisi*, *hangisinde*, *hangisinden*, *hangisine*, *hangisini*, *hangisinin*, *hangsilerinde*, *hangsilerinden*, *hani*, *hariç*, *hatta*, *hayır*, *hem*, *hemen*, *henüz*, *hepsi*, *hepsinde*, *hepsinden*, *hepsine*, *hepsini*, *hepsinin*, *her*, *herhangi*, *herkes*, *herkesin*, *hiç*, *hiçbir*, *icar*, *içeride*, *için*, *içinde*, *içinden*, *içine*, *ile*, *ilgili*, *irade*, *ise*, *ister*, *işte*, *iyi*, *izin*, *kadar*, *kâh*, *kapalı*, *karşı*, *karşın*, *karşısında*, *kazanılmış*, *kenara*, *keşke*, *kez*, *ki*, *kim*, *kimde*, *kimden*, *kime*, *kimi*, *kimin*, *kimisi*, *kimler*, *kimlerde*, *kimlerden*, *kimlere*, *kimleri*, *kimlerin*, *kimse*, *kişinin*, *kişiye*, *konum*, *lâkin*, *madem*, *mademki*, *meğer*, *meğerki*, *meğerse*, *meselâ*, *mı*, *mısın*, *mısınız*, *mıydı*, *mıyım*, *mıyız*, *mi*, *misin*, *misiniz*, *miyim*, *miyiz*, *mu*, *musun*, *musunuz*, *muydu*, *muyduk*, *muydum*, *muydun*, *muydunuz*, *muyum*, *muyuz*, *mü*, *nasıl*, *ne*, *nede*, *neden*, *nedenle*, *neler*, *nelerde*, *nelerden*, *nelere*, *neleri*, *nelerin*, *nere*, *nerede*, *nereden*, *neredeyse*, *nereler*, *nerelerde*, *nerelerden*, *nerelere*, *nereleri*, *nerelerin*, *nereye*, *nereyi*, *nereyin*, *neye*, *neyi*, *neyin*, *neyse*, *niçin*, *o*, *ol*, *olacağım*, *olacağız*, *olacak*, *olacaklar*, *olacaksın*, *olacaksınız*, *olan*, *olanları*, *olarak*, *oldu*, *olduğu*, *olduk*, *oldukça*, *oldular*, *oldum*, *oldun*, *oldunuz*, *olma*, *olmadan*, *olmadı*, *olmadık*, *olmadıkça*, *olmadılar*,

[120] A typical syllable in a native Turkish word takes the form $(C_1)V(C_2)$, containing a single vowel flanked by (optionally present) single consonants at both ends.

olmadım, *olmadın*, *olmadınız*, *olmak*, *olmali*, *olmam*, *olmamalı*, *olmamalıdır*, *olmanın*, *olmayacağım*, *olmayacağız*, *olmayacak*, *olmayacaklar*, *olmayacaksın*, *olmayacaksınız*, *olmayız*, *olmaz*, *olmazlar*, *olmazsın*, *olmazsınız*, *olmuştur*, *olmuyor*, *olmuyordu*, *olmuyorduk*, *olmuyordular*, *olmuyordum*, *olmuyordun*, *olmuyordunuz*, *olmuyorlar*, *olmuyorsun*, *olmuyorsunuz*, *olmuyorum*, *olmuyoruz*, *olsa*, *olsun*, *olup*, *olur*, *olurken*, *olurlar*, *olursa*, *olursun*, *olursunuz*, *olurum*, *oluruz*, *oluyor*, *oluyordu*, *oluyorduk*, *oluyordum*, *oluyordun*, *oluyordunuz*, *oluyorlar*, *oluyorlardı*, *oluyorsun*, *oluyorsunuz*, *oluyorum*, *oluyoruz*, *ona*, *onda*, *ondan*, *onlar*, *onlara*, *onlarda*, *onlardan*, *onları*, *onların*, *onu*, *onun*, *ora*, *orada*, *oradan*, *oralar*, *oralara*, *oralarda*, *oralardan*, *oraları*, *oraların*, *oranın*, *oraya*, *orayı*, *ordan*, *ortasında*, *oysa*, *ön*, *önce*, *önde*, *örneğin*, *ötesinde*, *ötürü*, *pek*, *rağmen*, *sadece*, *sahip*, *sahiptir*, *salt*, *sana*, *sayede*, *sebeple*, *sen*, *sende*, *senden*, *seni*, *senin*, *sık*, *sırasında*, *sırt*, *siz*, *sizde*, *sizden*, *size*, *sizi*, *sizin*, *son*, *sonra*, *sonradan*, *sonraki*, *sözgelimi*, *şayet*, *şekilde*, *şey*, *şimdi*, *şimdiye*, *şu*, *şuna*, *şunda*, *şundan*, *şunlar*, *şunlara*, *şunlarda*, *şunlardan*, *şunları*, *şunların*, *şunu*, *şunun*, *şura*, *şurada*, *şuradan*, *şuralar*, *şuralara*, *şuralarda*, *şuralardan*, *şuraları*, *şuraların*, *şuranın*, *şuraya*, *şurayı*, *tabii*, *takdirde*, *tam*, *tarafından*, *tek*, *tekrar*, *uzak*, *üstelik*, *üzere*, *üzerinde*, *üzerine*, *vakte*, *var*, *vardı*, *ve*, *verdim*, *vesile*, *veya*, *veyahut*, *vs*, *ya*, *yahut*, *yakında*, *yaklaşık*, *yanında*, *yani*, *yapacağım*, *yapacağız*, *yapacak*, *yapacaklar*, *yapacaksın*, *yapacaksınız*, *yapamadım*, *yapamaz*, *yapar*, *yaparım*, *yaparız*, *yaparlar*, *yaparsın*, *yaparsınız*, *yapıyor*, *yapıyordu*, *yapıyorduk*, *yapıyordum*, *yapıyordun*, *yapıyordunuz*, *yapıyorlar*, *yapıyorlardı*, *yapıyorsun*, *yapıyorsunuz*, *yapıyorum*, *yapıyoruz*, *yapma*, *yapmadı*, *yapmadık*, *yapmadılar*, *yapmadım*, *yapmadın*, *yapmadınız*, *yapmam*, *yapmayacağım*, *yapmayacağız*, *yapmayacak*, *yapmayacaklar*, *yapmayacaksın*, *yapmayacaksınız*, *yapmayız*, *yapmaz*, *yapmazlar*, *yapmazsın*, *yapmazsınız*, *yapmıyor*, *yapmıyordu*, *yapmıyorduk*, *yapmıyordular*, *yapmıyordum*, *yapmıyordun*, *yapmıyordunuz*, *yapmıyorlar*, *yapmıyorsun*, *yapmıyorsunuz*, *yapmıyorum*, *yapmıyoruz*, *yaptı*, *yaptık*, *yaptılar*, *yaptım*, *yaptın*, *yaptınız*, *yerine*, *yeterince*, *yine*, *yoksa*, *yoluyla*, *yukarı*, *yukarıdaki*, *yüzden*, *zaman*, *zamanda*, *zaten*, *zira*.

If a word exactly matches the string pattern **TurkishStopRoot**(∅|*ki*) or any one of the following:[121]

(*b*(*en*|*iz*)*im*|*on*(*ları*|*u*)*n*|*s*(*en*|*iz*)*in*)(∅|*ki*(∅|*n*))(∅|*de*(∅|*n*)|*e*|*i*(∅|*n*)),
(*değil*|*edebil*|*öyle*|*tüm*|*yok*)**X**,
edece(*ğ*|*k*)**X**,
etm(*e*|*iş*)**X**,
kendi(*leri*|(*m*|*n*)(∅|*iz*)|*si*)(∅|*n*)(∅|*de*(∅|*n*)|*e*|*i*(∅|*n*)),
ol(*abil*|*ma*|*u*(*n*|*r*|*y*))**X**,
ol(*aca*|*du*)(*ğ*|*k*)**X**,

then we consider it a Turkish stop word. All the Turkish stop words that appear in a particular document need to be ignored before we perform word clustering on the rest of the vocabulary list. □

9.3.1 Effective spelling and essential root

Definition 9.17 (Turkish protected range)**.** Set **V** = (*a*|*â*|*e*|*ı*|*i*|*î*|*o*|*ö*|*u*|*û*|*ü*). If $\hat{\sigma}$ does not contain **V**, define ProtRg($\hat{\sigma}$) = 0; otherwise, define ProtRg($\hat{\sigma}$) as the first position occupied by a letter in **V**.

Algorithm 9.18 (Turkish effective spelling)**.** *Set* **V** = (*a*|*â*|*e*|*ı*|*i*|*î*|*o*|*ö*|*u*|*û*|*ü*), **C** = $\overline{\mathbf{V}}$. *For a Turkish word* $\hat{\sigma}$*, its effective spelling* EffSpell($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Do* $I \to ı$, $\dot{I} \to i$, *before converting to lowercase.*

(2) *Look for the pattern* **CV**(*m*|*p*|*r*|*s*)**CV**~*. If the two occurrences of* **CV** *represent the same consonant-vowel combination, reduce the aforementioned pattern to just* **CV***; otherwise, leave as is.*[122]

(3) *Do* $ad\sim \to naμ$, $apacı\sim \to acı$, $emel\sim \to eμeλ$, $emin\mathbf{X}\sim \to σuρε$, $ıpıslak\sim \to ıslak$, $uzun\mathbf{X}\sim \to λoῦγ$, $\underline{mr} \to ηερρ$, $\underline{mrs} \to φραu$.

(4) *Call the result so far as* $\hat{\sigma}'$*, and break it down into* $\hat{\sigma}' = \hat{\sigma}_1\hat{\sigma}_2$*, where* $\ell(\hat{\sigma}_1)$ = ProtRg($\hat{\sigma}'$)*. Work on* $\hat{\sigma}_2$ *as follows:*

(4.1) *Do* $\sim(∅|u)ms(ı|i|u) \to ∅$, $\sim(∅|u)mt(ı|i)rak \to ∅$, $\sim(∅|y)(ı|i|u|ü)p \to ∅$, $\sim(c|ç)(a|e) \to ∅$, $\sim(c|ç)(ı|i)k \to ∅$*;*

(4.2) *Call the result so far as* $\hat{\sigma}'_2$*. If* $\hat{\sigma}'_2$ *contains* $s(ı|i|u|ü)z$*, then define* $\hat{\sigma}''_2 = \tilde{v}\hat{\sigma}^*_2$*, where* $\hat{\sigma}^*_2$ *results from doing* $s(ı|i|u|ü)z\mathbf{X} \to ∅$ *on* $\hat{\sigma}'_2$*; otherwise define* $\hat{\sigma}''_2 = \hat{\sigma}'_2$.

Concatenate $\hat{\sigma}_1$ *and* $\hat{\sigma}''_2$.

[121] The Turkish letter *ş* (Latin Small Letter S With Cedilla) should not be confused with the Romanian letter *ș* (Latin Small Letter S With Comma Below), as they occupy different positions in Unicode.

[122] This step treats the intensified form of certain Turkish adjectives, such as *bembeyaz* "totally white" (*cf. beyaz* "white") and *bomboş* "totally empty" (*cf. boş* "empty".)

(5) *Do* (*amca*|*dayı*)**X**~→ *ωṽκλe,* (*bibi*|*eme*|*hala*|*teyze*)**X**~→ *αuṽτ.*[123]

(6) *Replace*

*anla***X**~	*der*(∅|*iz*)~	*di*(∅|*me*)*y*(*ece*|*or*)**X**~	*düşün***X**~	*eleştir***X**~	*getir***X**~	*gid*(*e*|*i*)~	*götür***X**~	*gül*~
κομπρ	*de*	*de*	*θiṽκ*	*κριτr*	*βριγ*	*gitmek*	*καρρ*	*γülλ*
hanım~	*hatır***X**~	*his*~	*ileri***X**~	*itiraf***X**~	*itiraz***X**~	*izl***X**~	*kısmet***X**~	*kızar*~	*naμam*~	*oda*~	*sözü*~
λaδ	*μiṽδ*	*hisset*	*φροṽτ*	*κοṽφσ*	*ωβjεκτ*	*φολω*	*φατε*	*kızar*	*αdaμ*	*ρuμa*	*söz*
su(*lar*|*yu*)**X**~	*ulaş***X**~	*yemin***X**~	*yer*(∅|*iz*)~	*yi*(∅|*me*)*y*(*ece*|*or*)**X**~	*anda*	*su*(∅|*da*|*dan*|*dur*|*sun*|*sunuz*|*uz*|*ya*)				
hho	*ρeαch*	*νoω*	*ye*	*ye*	*noω*	*hho*				

(7) *Call the result so far as* $\hat{\sigma}^{\dagger}$*, and break it down into* $\hat{\sigma}^{\dagger} = \hat{\sigma}_1^{\dagger}\hat{\sigma}_2^{\dagger}$*, where* $\ell(\hat{\sigma}_1^{\dagger}) = \mathrm{ProtRg}(\hat{\sigma}^{\dagger})$*. Work on* $\hat{\sigma}_2^{\dagger}$ *as follows:*

(7.1) *Replace*

♯**X**	(*abil*|*ebil*|*mak*|*mek*)**X**	(*aca*|*ece*)(*ğ*|*k*)**X**	*c*(*a*|*e*)*s*(*ı*|*i*)*n*(*a*|*e*)**X**	*l*(*a*|*e*)(*r*|*ş*)**X**
∅	∅	∅	∅	∅
m(*ı*|*i*|*u*|*ü*)*ş***X**	*r*~	*yor***X**	~(∅|*n*)(*d*|*t*)(*a*|*e*)(∅|*n*)	
∅	*ρ*	∅	*Ta*	
~(*d*|*t*)(*ı*|*i*|*u*|*ü*)(∅|*k*|*l*(*a*|*e*)*r*)|*m*|*n*(∅|(*ı*|*i*|*u*|*ü*)*z*)|*r***X**)	~*ki*(∅|*n*)(*de*|*den*|*e*|*i*|*in*)	~*s*(*ı*|*i*|*u*|*ü*)*n*(∅|(*ı*|*i*|*u*|*ü*)*z*)
Tu	∅	∅

(7.2) *Do* (*d*\|*t*)(*ı*\|*i*\|*u*\|*ü*)(*ğ*\|*k*)**X**→ ∅, **V***ym*(*ı*\|*i*\|*u*\|*ü*)*ş***X**→ **V**, *m*(*a*\|*e*\|*ı*\|*i*\|*u*\|*ü*)**X**→ ∅, ~(*ı*\|*i*\|*u*\|*ü*)(*m*\|*n*)(∅\|(*ı*\|*i*\|*u*\|*ü*)*z*) → ∅, ~**V***y*(∅\|*l*)(*a*\|*e*) → **V**, ~**V***y*(∅\|*Tu*) → **V**, ~**V***y*(*ı*\|*i*\|*u*\|*ü*)(∅\|*m*\|*z*) → **V**, ~**V***yken* → **V**, ~**V***ys*(*a*\|*e*) → **V**.

(7.3) *Do* ~(∅\|*ı*\|*i*\|*u*\|*ü*)(*m*\|*n*) → ∅.

(7.4) *Call the result so far as* $\hat{\sigma}_2^{\ddagger}$*. Concatenate* $\hat{\sigma}_1^{\dagger}$ *and* $\hat{\sigma}_2^{\ddagger}$*.*

(8) *Do T***X**→ ∅, ~*r*(∅\|(*ı*\|*i*\|*u*\|*ü*)*z*) → ∅.

Algorithm 9.19 (Turkish essential root)**.** *Let* $\hat{\sigma}$ *be the effective spelling of a Turkish word, then its corresponding essential root* EssRoot($\hat{\sigma}$) *is constructed in the following steps:*

(1) *Search for the pattern* $\mathbf{C}_{m_0}\mathbf{V}\mathbf{C}_{m_0}\mathbf{V}_{m_0}(\varnothing|\overline{(n|s)})$~*. If this search is successful, define* $\hat{\sigma}^{\flat}$ *as the result from doing* ~(*ı*\|*i*\|*u*\|*ü*) → ∅ *on such a pattern; otherwise, define* $\hat{\sigma}^{\flat}$ *as* $\hat{\sigma}$*.*

(2) *On* $\hat{\sigma}^{\flat}$*, do* ~*b*→*p,* ~*c*→*ç,* ~*d*→*t,* ~*ğ*→*k.*

9.3.2 Approximate clustering

There will not be any mutation rules in our algorithm for Turkish, so SimpHrdTest = HrdTest.

Algorithm 9.20 (Simple heredity test)**.** *Suppose that* $\hat{\alpha}$ *and* $\hat{\beta}$ *are both lowercase strings. The Boolean-valued function* SimpHrdTest($\hat{\alpha}$,$\hat{\beta}$) *returns* **TRUE** *if* $\hat{\alpha}$ *contains at least one instance of* **V***,* **AND** *at least one of the following two conditions holds:*[124]

(i) $\hat{\alpha} = \hat{\beta}$*;*

(ii) $\hat{\alpha}' = \hat{\beta}'$*, where the strings with prime result from doing* $\sim^{\mathbf{X}\in}(\mathbf{V}\hat{\chi})(\varnothing|l)\mathbf{V} \to \mathbf{X}$ *on their counterparts without prime.*

Algorithm 9.21 (Approximate clustering of Turkish words)**.** *The approximate clustering of a list of Turkish words* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_N\}$ *is completed in two stages:*

(1) *We sort the list* $\{(\hat{\alpha}_1, \mathrm{EffSpell}(\hat{\alpha}_1)), \ldots, (\hat{\alpha}_N, \mathrm{EffSpell}(\hat{\alpha}_N))\}$ *alphabetically according to the last component. If two consecutive neighbors in the alphabetized list* $\{(\hat{\alpha}_{(1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1)})), \ldots, (\hat{\alpha}_{(N)}, \mathrm{EffSpell}(\hat{\alpha}_{(N)}))\}$ *satisfy* $\mathrm{HrdTest}(\mathrm{EffSpell}(\hat{\alpha}_{(n)}), \mathrm{EffSpell}(\hat{\alpha}_{(n+1)})) = \mathbf{FALSE}$*, where* $n \in \mathbb{Z} \cap [1, N)$*, then add a demarcation line between these two entries. In this way, the alphabetized list is divided into separate groups of words tagged with their effective spellings:* $\{g_1 = \{(\hat{\alpha}_{(1,1)}, \mathrm{EffSpell}(\hat{\alpha}_{(1,1)})), \ldots, (\hat{\alpha}_{(1,n_1)}, *)\}, \ldots, g_M = \{(\hat{\alpha}_{(M,1)}, *), \ldots, (\hat{\alpha}_{(M,n_M)}, *)\}\}$*, where each sublist also preserves alphabetic order. In particular, we have* $\hat{\alpha}_{(1,1)} = \hat{\alpha}_{(1)}$ *and* $\hat{\alpha}_{(M,n_M)} = \hat{\alpha}_{(N)}$*.*

[123] In this step, we group certain kinship terms in Turkish so as to match their English counterparts. However, since the Turkish word *kardeş* means both "brother" and "sister", we cannot do anything to compensate for this.

[124] As a general rule in this document, the truth values of items labeled with Roman numerals are connected to each other with logical **OR**.

(2) *For each group of words* $g_m = \{(\hat{\alpha}_{(m,1)}, *), \ldots, (\hat{\alpha}_{(m,n_m)}, *)\}$ *where* $m \in \mathbb{Z} \cap [1, M]$*, we augment it into a tagged entry* $G_m = (g_m, \mathrm{EffSpell}(\hat{\alpha}_{(m,1)}), \mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})))$*. The list* $\{G_1, \ldots, G_M\}$ *is sorted alphabetically, with respect to* $\mathrm{EssRoot}(\mathrm{EffSpell}(\hat{\alpha}_{(m,1)}))$ *(with higher priority), and* $\mathrm{EffSpell}(\hat{\alpha}_{(m,1)})$ *(with lower priority). If two consecutive neighbors in the alphabetized list* $\{G_{(1)} = (g_{(1)}, \hat{\gamma}'_{(1)}, \hat{\gamma}''_{(1)}), \ldots, G_{(M)} = (g_{(M)}, \hat{\gamma}'_{(M)}, \hat{\gamma}''_{(M)})\}$ *satisfy*

$$\mathrm{HrdTest}(\hat{\gamma}''_{(m)}, \hat{\gamma}''_{(m+1)}) = \textbf{FALSE}$$

AND

$$\mathrm{HrdTest}(\hat{\gamma}'_{(m)}, \hat{\gamma}'_{(m+1)}) = \textbf{FALSE}$$

where $m \in \mathbb{Z} \cap [1, M)$*, then add a demarcation line between these two entries. In this manner, the alphabetized list* $\{G_{(1)}, \ldots, G_{(M)}\}$ *is divided into separate groups of tagged clusters* $\{\Gamma_1 = \{G_{(1,1)}, \ldots, G_{(1,m_1)}\}, \ldots, \Gamma_K = \{G_{(K,1)}, \ldots, G_{(1,m_K)}\}\}$*. Finally, the output list of word clusters* $\{\check{\Gamma}_1, \ldots, \check{\Gamma}_K\}$ *is constructed by discarding all the tags (effective spellings, essential roots) from each* Γ_k*, where* $k \in \mathbb{Z} \cap [1, K]$*.*

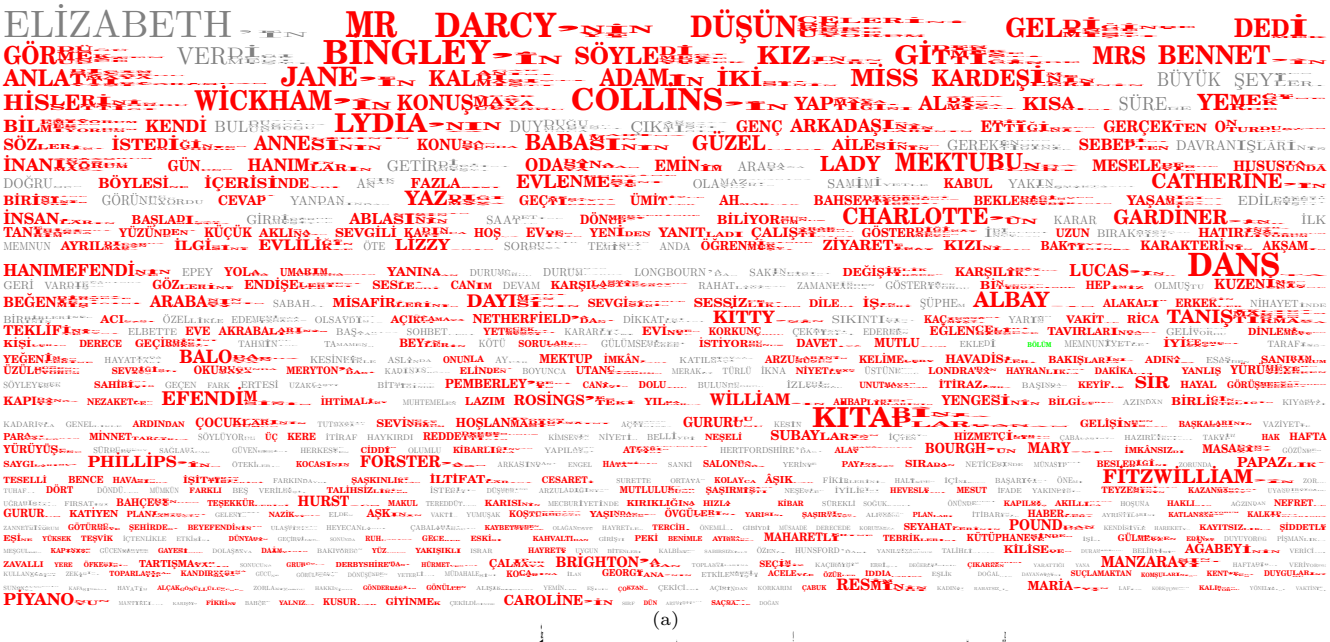
Example 9.21.1. Our clustering algorithm correctly groups the following families of Turkish words:

ev, *evde*, *evden*, *eve*, *evi*, *evin*, *evler*, *evlerde*, *evlerden*, *evlere*, *evleri*, *evlerin* — "house";
dedi, *dedik*, *dediler*, *dedim*, *dedin*, *dediniz*, *demedi*, *demedik*, *demediler*, *demedim*, *demedin*, *demediniz*, *demem*, *demeyeceğim*, *demeyeceğiz*, *demeyecek*, *demeyecekler*, *demeyeceksin*, *demeyeceksiniz*, *demeyiz*, *demez*, *demezler*, *demezsin*, *demezsiniz*, *demiyor*, *demiyordu*, *demiyorduk*, *demiyordular*, *demiyordum*, *demiyordun*, *demiyordunuz*, *demiyorlar*, *demiyorsun*, *demiyorsunuz*, *demiyorum*, *demiyoruz*, *der*, *derim*, *deriz*, *derler*, *dersin*, *dersiniz*, *dimeyecek*, *dimeyecekler*, *diyeceğim*, *diyeceğiz*, *diyecek*, *diyecekler*, *diyeceksin*, *diyeceksiniz*, *diyor*, *diyordu*, *diyorduk*, *diyordum*, *diyordun*, *diyordunuz*, *diyorlar*, *diyorlardı*, *diyorsun*, *diyorsunuz*, *diyorum*, *diyoruz* — "say";
al, *alacağım*, *alacağız*, *alacak*, *alacaklar*, *alacaksın*, *alacaksınız*, *aldı*, *aldık*, *aldılar*, *aldım*, *aldın*, *aldınız*, *alır*, *alırım*, *alırız*, *alırlar*, *alırsın*, *alırsınız*, *alıyor*, *alıyordu*, *alıyorduk*, *alıyordum*, *alıyordun*, *alıyordunuz*, *alıyorlar*, *alıyorlardı*, *alıyorsun*, *alıyorsunuz*, *alıyorum*, *alıyoruz*, *almadı*, *almadık*, *almadılar*, *almadım*, *almadın*, *almadınız*, *almam*, *almayacağım*, *almayacağız*, *almayacak*, *almayacaklar*, *almayacaksın*, *almayacaksınız*, *almayız*, *almaz*, *almazlar*, *almazsın*, *almazsınız*, *almıyor*, *almıyordu*, *almıyorduk*, *almıyordular*, *almıyordum*, *almıyordun*, *almıyordunuz*, *almıyorlar*, *almıyorsun*, *almıyorsunuz*, *almıyorum*, *almıyoruz* — "buy";
okudu, *okuduk*, *okudular*, *okudum*, *okudun*, *okudunuz*, *okumadı*, *okumadık*, *okumadılar*, *okumadım*, *okumadın*, *okumadınız*, *okumam*, *okumayacağım*, *okumayacağız*, *okumayacak*, *okumayacaklar*, *okumayacaksın*, *okumayacaksınız*, *okumayız*, *okumaz*, *okumazlar*, *okumazsın*, *okumazsınız*, *okumuyor*, *okumuyordu*, *okumuyorduk*, *okumuyordular*, *okumuyordum*, *okumuyordun*, *okumuyordunuz*, *okumuyorlar*, *okumuyorsun*, *okumuyorsunuz*, *okumuyorum*, *okumuyoruz*, *okur*, *okurlar*, *okursun*, *okursunuz*, *okurum*, *okuruz*, *okuyacağım*, *okuyacağız*, *okuyacak*, *okuyacaklar*, *okuyacaksın*, *okuyacaksınız*, *okuyor*, *okuyordu*, *okuyorduk*, *okuyordum*, *okuyordun*, *okuyordunuz*, *okuyorlar*, *okuyorlardı*, *okuyorsun*, *okuyorsunuz*, *okuyorum*, *okuyoruz* — "read";
kitaba, *kitabı*, *kitabın*, *kitap*, *kitaplar*, *kitaplara*, *kitaplarda*, *kitaplardan*, *kitapları*, *kitapların*, *kitapta*, *kitaptan* — "book".

Example 9.21.2. In Fig. S16, we further apply the aforementioned word clustering algorithm to topic extraction and machine translation (see Table S1 for text source).

Note that the Turkish translation for *disappointment* consists of two words *hayal kırıklığı* (literally "dream brokenness"), where *hayal* means "dream". This explains why there are two hot spots in the row for *disappointment* in Fig. S16b.

In Turkish, *de* "also, too" is a stop word.



(a)

(b)

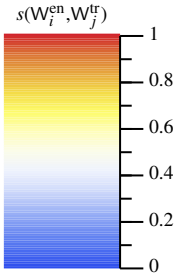


Fig. S16. Text mining in Turkish. (a) Statistically identified topics ($n_{ii} \geq 20$) in a Turkish version of *Pride and Prejudice*, with the same color encoding scheme as Fig. S3. (b) Semantic similarities $s(\mathrm{W}_i^{\mathrm{en}}, \mathrm{W}_j^{\mathrm{tr}})$ between selected topics in English and Turkish versions of *Pride and Prejudice*. Cross-hairs meet at optimal nodes that solve the bipartite matching problem. The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) rank of the similarity score for the optimal node. *Green* (resp. *amber*) cross-hair indicates an exact (resp. a close but non-exact) match.

[1] W. E and Y. Zhou, "A mathematical model for universal semantics," 2020, arXiv:1907.12293v6 [cs.CL].

[2] W. A. Gale, K. W. Church, and D. Yarowsky, "A method for disambiguating word senses in a large corpus," *Comput. Humanities*, vol. 26, no. 5-6, pp. 415–439, 1992.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artificial Intelligence Res.*, vol. 37, pp. 141–188, Feb. 2010.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*. La Jolla, CA: NIPS, 2013, pp. 3111–3119.

[6] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "A latent variable model approach to PMI-based word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 385–399, 2016.

[7] Y. Zhou and X. Zhuang, "Robust reconstruction of the rate constant distribution using the phase function method," *Biophys. J.*, vol. 91, no. 11, pp. 4045–4053, 2006.

[8] N. Haydn, Y. Lacroix, and S. Vaienti, "Hitting and return times in ergodic dynamical systems," *Ann. Probab.*, vol. 33, pp. 2043–2050, 2005.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley Interscience, 2006.

[10] M. Pollicott and M. Yuri, *Dynamical Systems and Ergodic Theory*, ser. London Mathematical Society Student Texts. Cambridge, UK: Cambridge University Press, 1998, vol. 40.

[11] M. Ružička, "Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen)," *Biológia (Bratislava)*, vol. 13, pp. 647–661, 1958.

[12] E. Deza and M.-M. Deza, *Dictionary of distances*. Amsterdam, The Netherlands: Elsevier, 2006.

[13] G. Gilbert, "Distance between sets," *Nature*, vol. 239, no. 5368, pp. 174–174, 1972.

[14] F. McSherry and M. Najork, "Computing information retrieval performance measures efficiently in the presence of tied scores," in *Advances in Information Retrieval. ECIR 2008*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, Eds. Berlin, Germany: Springer, 2008, pp. 414–421.

[15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Networks ISDN*, vol. 30, no. 1-7, pp. 107–117, 1998.

[16] M. Bressan and E. Peserico, "Choose the damping, choose the ranking?" *Journal of Discrete Algorithms*, vol. 8, no. 2, pp. 199–213, 2010.

[17] D. Lardiere, "Words and their parts," in *An Introduction to Language and Linguistics*, R. Fasold and J. Connor-Linton, Eds. Cambridge, UK: Cambridge University Press, 2006, ch. 2, pp. 55–96.

[18] A. D. Friederici, "The neurobiology of language comprehension," in *Language Comprehension: A Biological Perspective*, A. D. Friederici, Ed. Berlin, Germany: Springer, 1999, ch. 9, pp. 265–304.

[19] R. Jakobson, "The identification of phonemic entities," in *Recherches structurales 1949. Interventions dans le débat glossématique*, ser. Travaux du Cercle Linguistique de Copenhague. Copenhagen, Denmark: Nordisk Sprog- og Kulturforlag, 1949, vol. V, pp. 205–213.

[20] C. Everett, D. E. Blasí, and S. G. Roberts, "Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots," *Proc. Natl. Acad. Sci. USA*, vol. 112, pp. 1322–1327, 2015.

[21] C. R. Ember, "Commentary: Considering language as expressive culture," *Journal of Language Evolution*, vol. 1, no. 1, pp. 60–61, 2016.

[22] C. Everett, "Languages in drier climates use fewer vowels," *Frontiers in Psychology*, vol. 8, p. Article 1285, 2017.

[23] J. G. Frazer, *The Golden Bough: a study of magic and religion*, abridged ed. New York, NY: MacMillan, 1922.

[24] W. Labov, *Principles of Linguistic Change: Social Factors*. Malden, MA: Blackwell Publishers, 2001.

[25] ——, *Principles of Linguistic Change: Cognitive and Cultural Factors*. Malden, MA: Wiley-Blackwell, 2010.

[26] K. Allan and K. Burridge, *Forbidden Words: Taboo and the Censoring of Language*. Cambridge, UK: Cambridge University Press, 2006.

[27] A. D. Friederici, J. Bahlmann, S. Heim, R. I. Schubotz, and A. Anwander, "The brain differentiates human and non-human grammars: Functional localization and structural connectivity," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 7, pp. 2458–2463, 2006.

[28] S. M. Wilson, S. Galantucci, M. C. Tartaglia, K. Rising, D. K. Patterson, M. L. Henry, J. M. Ogar, J. DeLeon, B. L. Miller, and M. L. Gorno-Tempini, "Syntactic processing depends on dorsal language tracts," *Neuron*, vol. 72, no. 2, pp. 397–403, 2011.

[29] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based *n*-gram models of natural language," *Computat. Linguist.*, vol. 18, no. 4, pp. 467–479, 1992.

[30] S. Pinker, *The Stuff of Thought: Language as a Window into Human Nature*. New York: Viking, 2007.

[31] ——, *Learnability and Cognition: the Acquisition of Argument Structure*. Cambridge, MA: The MIT Press, 2013.

[32] N. Chomsky, *The Minimalist Program*. Cambridge, MA: MIT Press, 2015.

[33] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in graphical models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 105–161.

[34] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Learning in graphical models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 301–354.

[35] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009.

[36] W. E, J. Lu, and Y. Yao, "The landscape of complex networks—critical nodes and a hierarchical decomposition," *Methods Appl. Anal.*, vol. 20, no. 4, pp. 383–404, 2013.

[37] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[38] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[39] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[40] ——, "Snowball — a small string processing language designed for creating stemming algorithms for use in Information Retrieval," http://snowball.tartarus.org/, 2014.

[41] S. Pinker, "Words and rules in the human brain," *Nature*, vol. 387, no. 6633, pp. 547–548, 1997.

[42] W. D. Marslen-Wilson and L. K. Tyler, "Dissociating types of mental computation," *Nature*, vol. 387, no. 6633, pp. 592–594, 1997.

[43] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak, "Quantifying the evolutionary dynamics of language," *Nature*, vol. 449, no. 7163, pp. 713–716, 2007.

[44] Y. Yang, W.-t. Yih, and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal: Association for Computational Linguistics, 2015.

[45] J. Yeon and L. Brown, *Korean: a comprehensive grammar*. London, UK: Routledge, 2011.

[46] T. Lundskær-Nielsen and P. Holmes, *Danish: An Essential Grammar*, 2nd ed. London, UK: Routledge, 2000.

[47] B. C. Donaldson, *Dutch: A Comprehensive Grammar*, 2nd ed. London, UK: Routledge, 2008.

[48] C. Rounds, *Hungarian: An Essential Grammar*. London, UK: Routledge, 2001.

[49] S. Starostin, A. Dybo, and O. Mudrak, *Etymological dictionary of the Altaic languages*. Leiden, The Netherlands: Brill, 2003.

[50] J. I. Hualde and J. Ortiz de Urbina, Eds., *A grammar of Basque*. Berlin, Germany: Mouton de Gruyter, 2003.

[51] G. Eryiğit and E. Adalı, "An affix stripping morphological analyzer for Turkish," in *Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, Innsbruck, Austria, 2004.