Gramfinder: Human and Machine Reading of Grammatical Descriptions of the Languages of the World

Harald Hammarström
harald.hammarstrom@lingfil.uu.se
Uppsala University
Uppsala, Sweden

ABSTRACT

The diversity of the world's approximately 7 000 languages embodies a wealth of information on the communication machinery inside our heads as well as the history of populations. Traditionally, data has been collected manually by humans reading grammatical descriptions of individual languages, but the number of languages and books is now far beyond human capacities. At the same time, over 37 000 books and articles relating to descriptive linguistics — most importantly grammatical descriptions of minority languages — have been digitized and are available for data mining research purposes. At present, technologies for automatic extraction of information from grammars can only target relatively formulaic facts with high accuracy. For the next stage of database curation, we are thus facing a trade-off between the labour-intensive high accuracy and high capability of human reading and the quick and objective but less capable data mining approaches. Given this situation, we present Gramfinder — a free and open search/extraction interface targeted to the specific domain and task. Gramfinder has a slim human-controllable term extraction component, the relevant document sorting and selection facilities, search result highlighting, quick access to the relevant section of full documents, bibliographic support, structured export and map facilities. This setup saves different amounts of time in dialectic human-computer database curation depending on the granularity and complexity of the sought-after items of information.

CCS CONCEPTS

• **Applied computing → Document management and text processing**.

KEYWORDS

data mining, computational linguistics, human-computer interaction, linguistic typology

ACM Reference Format:
Harald Hammarström. 2021. Gramfinder: Human and Machine Reading of Grammatical Descriptions of the Languages of the World. In *DaSH '21@KDD'21: Workshop on Data Science with Human-in-the-loop, August 15–16, 2021, Virtual Conference*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

DaSH '21, August 15–16, 2021, Virtual Conference

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION AND PREVIOUS WORK

The diversity of the world's approximately 7 000 languages embodies a wealth of information on the communication machinery inside our heads as well as the history of populations [5]. Although the interest in linguistic diversity has a long history, it is only in the last decade that large databases with cross-linguistic data have become available and practical to use for quantitative empirical studies. The most widely known open databases of linguistic structure include the World Atlas of Language Structures (WALS) (wals.info), Atlas of Pidgin and Creole Language Structures (APiCS) (apics.org), the South American Indigenous Language Structures (SAILS) (sails.clld.org) and PHOIBLE (phoible.org)[1].

Traditionally, data on the languages of the world have been collected manually by humans reading grammatical descriptions. For example, the below reference describes a language called Kagulu, whose grammatical properties are of interest for various linguistic research questions.

Petzell, Malin. (2008) *The Kagulu language of Tanzania: grammar, text and vocabulary* (East African languages and dialects 19). Köln: Rüdiger Köppe Verlag. 234pp.

However, the number of languages and books about them is now far beyond human capacities — there are over 37 000 descriptive items of literature comprising all the world's languages (of which some 12 000 are grammars describing over over 4 500 languages). Moreover, human reading, although ideal in many ways, is very labour-intensive and potentially involves subjectivity. A single linguistic subdomain (equivalent of, say, 10-30 discrete features) covering 50-300 languages is typically the size of a PhD thesis (e.g., Di Garbo [3], Krasnoukhova [19], Miestamo [24], van Lier [30]) although PhD work also includes design and analysis, in addition to data collection. Perhaps more relevant for comparisons is the time required for MA-level research assistants to collect data of this kind. For a fixed questionnaire of some 200 features [26], assistants within the Nijmegen Typological Survey [27] could collect data at a speed of up to 20 features per hour. This can be achieved only working language-by-language when presented with an electronic copy of a descriptive grammar for each language. The pace is slower if working feature-by-feature or if the relevant information is scattered across several publications. In other words, a significant amount of time is spent not so much on reading and interpreting per se, but on orientation and finding the relevant section.

[1] A fuller listing of available linguistic databases is provided at http://languagegoldmine.com/, accessed 2021-06-01.

Given the relatively novel appearance of the task of automatically extracting information from descriptive grammars[2], only a few embryonic approaches have appeared [13, 21, 31, 33, 34]. All of them require the user to submit a word, phrase or description of the sought after feature but retrieve the result with varying depths of analysis of the source documents and with varying amounts of training data required. As such, the techniques so far proposed are relatively transparent. Since the target application is databases with human-scrutizable information, black-box approaches, e.g., neural networks, are less in demand even if they were to produce higher accuracy in some sense.

Some linguistic features, e.g., `classifier`, `suffix(es)`, `preposition(s)`, `rounded vowel(s)` or `inverse`, are signalled by a specific keyword (or a small set thereof) and their presence/absense can indeed be inferred automatically with accuracy competing or exceeding that of humans [13, 31]. Under favourable circumstances the relative frequency of the term(s) in question can also be exploited [12]. However, other features of interest, such as whether the verb agrees with the agent in person, are expressed in a myriad of ways, and are not amenable to such simple routines. It may be noted that the important class of word-order features, which are among the easiest for a human to discern from a grammar, typically belong to the class of non-term-signalled features unless there is a specific formula such as SOV or N-Adj gaining sufficient popularity in grammatical descriptions. For many prospective practitioners, it is not immediately obvious where on the scale of feasibility a certain feature of interest falls. Different practitioners also have different goals and resources, so that precision and recall will be valued quite differently.

None of the work so far has elaborated on the role of human-computer interaction against the background of varied difficulty of extraction and different valuings of precision versus recall. Hence, we present Gramfinder — a boosted search / feature extraction interface for dialectic human-computer data mining tailored to the specific domain of grammatical descriptions.

2 DATA

The document collection for the search infrastructure described in this paper consists of over 37 000 digitized books and articles relating to descriptive linguistics. The most important subset is made up of some 12 000 grammatical descriptions (see Virk et al. [32]), but the collection also includes dictionaries, scoiolinguistic studies, phonologies, comparative studies, text collections, overviews, wordlists and bibliographies [15]. The collection comprises (1) out-of-copyright texts digitized by national libraries, archives, scientific societies and other similar entities, (2) texts posted online with a license to use for research, usually by university libraries and non-profit organizations (notably the Summer Institute of Linguistics), and (3) texts under publisher copyright where quotations of short extracts are legal. A listing of the collection can be enumerated via the open-access bibliography Glottolog (glottolog.org, [10]). For each reference pertaining to the present study, this catalogue features manually curated annotations of

(i) the language it is written in (the meta-language, usually English, French, German, Spanish, Russian or Mandarin Chinese, see Table 1),
(ii) the language(s) described in it (the vernacular, typically one of the thousands of minority languages throughout the world), and
(iii) the type of description (comparative study, description of a specific feature, phonological description, grammar, bibliography, sociolinguistic study, overview etc, see [15]).

The collection contains descriptive information of varying amounts for no less than 7 521 languages, very close to the total number of languages known (cf. Hammarström et al. [9]).

Table 1: Meta-languages of documents in the present collection.

Meta-language		# docs	# pages
English	eng	24 081	3 493 543
French	fra	3 632	666 187
German	deu	2 798	434 647
Spanish	spa	2 628	382 859
Portuguese	por	1 343	177 615
Russian	rus	993	267 263
Dutch	nld	662	92 447
Mandarin Chinese	cmn	537	131 687
Indonesian	ind	477	72 148
Italian	ita	402	71 552
…	…	…	…

The collection has been OCRed using ABBYY Finereader 14 using the meta-language as recognition language. The original digital documents are of quality varying from barely legible typescript copies to high-quality scans and even born-digital documents. The OCR correctly recognizes most tokens of the set meta-language, but, particular to our collection, most documents contain a fraction of tokens which do not belong to the set meta-language but to the minority language(s) being discussed. These tokens are typically recognized poorly, as expected from the dictionary/training-heavy, contemporary techniques for OCR. We cannot easily improve on the OCR on a scale relevant for the present collection but some post-correction of OCR output very relevant for the genre of linguistics is possible and advisable (see Hammarström et al. [16]). The bottom line is that searches for meta-language terms are relatively unimpaired by OCR but parsing of full sentences may falter.

3 GRAMFINDER

Against the background of the collection specifics, the target domain, copyright restrictions, and the demand for dialectic and explainable data mining support just explained, we have developed Gramfinder — a free and open tool for building cross-linguistic databases semi-automatically. More specifically, the design principles are as follows:

Metadata control: As explained above, the document collection is tied to a bibliography with full hand-curated metadata. This allows filtering and specialization for, e.g., year, language described, meta-language, type of description etc.

[2]We are not aware of any work on parallel problems in other scientific domains (cf. Firoozeh et al. [6], Nasar et al. [25]).

Notably, the same cannot be done in, e.g., Google Books due to lack of (or low accuracy) extensive metadata [18]. An example snapshot of metadata, references and filter possibilities derived from glottolog.org [10] is shown in Figure 1.

Document-object correspondence: An important subset of the documents in the collection, notably the grammatical descriptions, describe exactly one language. Formally, if D is the set of raw-text descriptions, each $d \in D$ mainly describes exactly one s in a set of entities (here: languages) S. Then if a term k describing a property of objects in S occurs in a document d to a significant degree, we can infer that the object s described in d actually has the property signalled by k. This is perhaps the most important characteristic for semi-automatically building a cross-linguistic database successfully. Some parallel premise presumably applies to other domains and texts, e.g., ethnographic descriptions, but we are unaware of related work that make essential use of it in other scientific fields. Large collections such as Google Books are unable to exploit the correspondence if there is no suitable metadata [23].

Explainable data mining: Current cross-linguistic databases are typically browsed by humans and/or modelled computationally in experiments that require a human understanding of the data at hand [1, 4, 7, 20]. For this reason, any data mining technique applied to the collection in this context must be explainable [8], i.e., a linguist must be able to go back and check how an algorithm arrived at a given judgment from the original text. For the purpose of Gramfinder (though this is likely to be extended to include other options in the future) we have implemented the term-spotting technique of Hammarström et al. [13] since this involves no deep parsing, no training data, no human tuning of thresholds and is fully transparent. It requires the user to supply a keyword or set of keywords (or individual such sets for each meta-language) associated with the feature of interest. Unfortunately, to simply look for the existence of a term is too naive. In many grammars, terms for grammatical features do happen to occur although the language being described, in fact, does not exhibit the feature. However, the more often the term occurs, the higher the probability that it is not "noise". Hammarström et al. [13] describe a simple technique in which a threshold for significance is automatically calculated for each document without the need for human intervention, supervision or other knowledge source. Figure 2 shows some sample data mining output for the terms `prefix` and `suffix` where the automatically calculated threshold t is shown along with the number of hits. Although no parameters need to be tweaked, the extraction term(s) do need to be supplied by a human. As explained in the introduction, the user may be able to find a suitable such set only by iterated trial and inspection depending on his/her familiarity with the literature and ultimate goals.

Inspection support: To allow the user a quick appreciation of the accuracy and potential value of a search/extraction procedure, Gramfinder offers normal search result lists with the relevant paragraph, highlighting and link to full text, familiar from any modern search engine (see Figures 2-3). In addition, since the results concern languages, results are shown aggregated by source (usually there is more than one source document per language) and shown on a map (using languages coordinates from glottolog.org, [10]), cf. Figure 4. In addition, the results can be optionally aggregated or sorted according to language family and/or geographical macroarea, as classified in the latter source.

Search infrastructure: The user may also switch off any data mining technique altogether and revert to a pure search interface. In this case, standard string search facilities from Information Retrieval [22], such as RegExps, capitalization, boolean operators and tokenization[3] are allowed. Naturally, sorting and filtering of results are also featured.

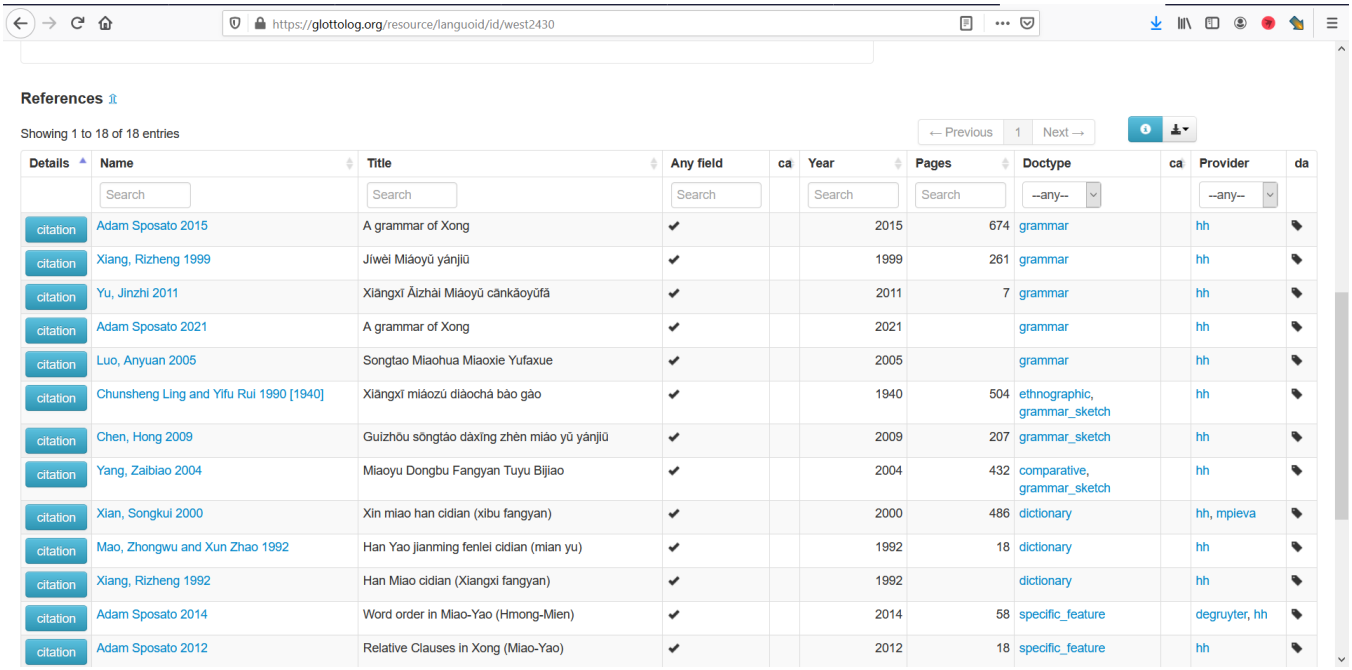
Export facilities: Search / extraction results are directly exportable in a CSV-format for any further operations (see Figure 4). Crucially, the relevant metadata need to be included in the export such as source, year, language, number of pages, place of publication and so on. This allows the user to do any (dis)aggregation desired, e.g., use the newest or longest description per language in place of the (default) majority vote.

Copyright compliance: Since a large fraction of the source texts are under active copyright, an open search/extraction tool cannot reproduce full-texts directly or indirectly. Hence, the search/extraction snippets have to stay within the legal size boundaries and links to the full-texts have to be mediated by password protection or the like (as needed per individual item). The out-of-copyright subsection of the grammatical descriptions are available for full browsing and download via Språkbanken (https://spraakbanken.gu.se/korp/?mode=dream, see [32]).

The Gramfinder interface differs from Google Books and similar efforts with its full metadata control and the inclusion of a data mining component. Consequently, Gramfinder can also provide export facilities with rich information that immediately constitutes a raw cross-linguistic database. It also extends existing general-purpose corpus browsing tools (such as Korp, see [2]) with domain-specific components such as language maps and language family support.

No formal user survey have yet been conducted, but the interface has been tested informally by a range of interested linguists on such topics as classifiers [17], gender and noun class [14], ejectives [29], laryngeal/creaky voice, ditransitive constructions, ideophones, language endangerment [35], partitives, prefix/suffix statistics [12], spatial frames of refences, directionals and locatives, word taboo, spontaneous nasalization, switch reference, polysynthesis and tone. Regarding the role of human in the loop in these efforts, the following can be observed. The data mining extraction for the features `classifier` [17], `noun class` and `masculine`, `feminine`, `neuter` gender [14] have been manually checked. Each datapoint could be checked more or less at a glance using Gramfinder, resulting in by far the largest databases (3000+ languages) in existence for these features. The lion's share of the speed is due to organizational help, source aggregation, highlighted snippets and quick links to the originals. Given the corrected data, the data mining technique can

[3] For Chinese, the Jieba https://github.com/fxsjy/jieba tokenizer is employed.



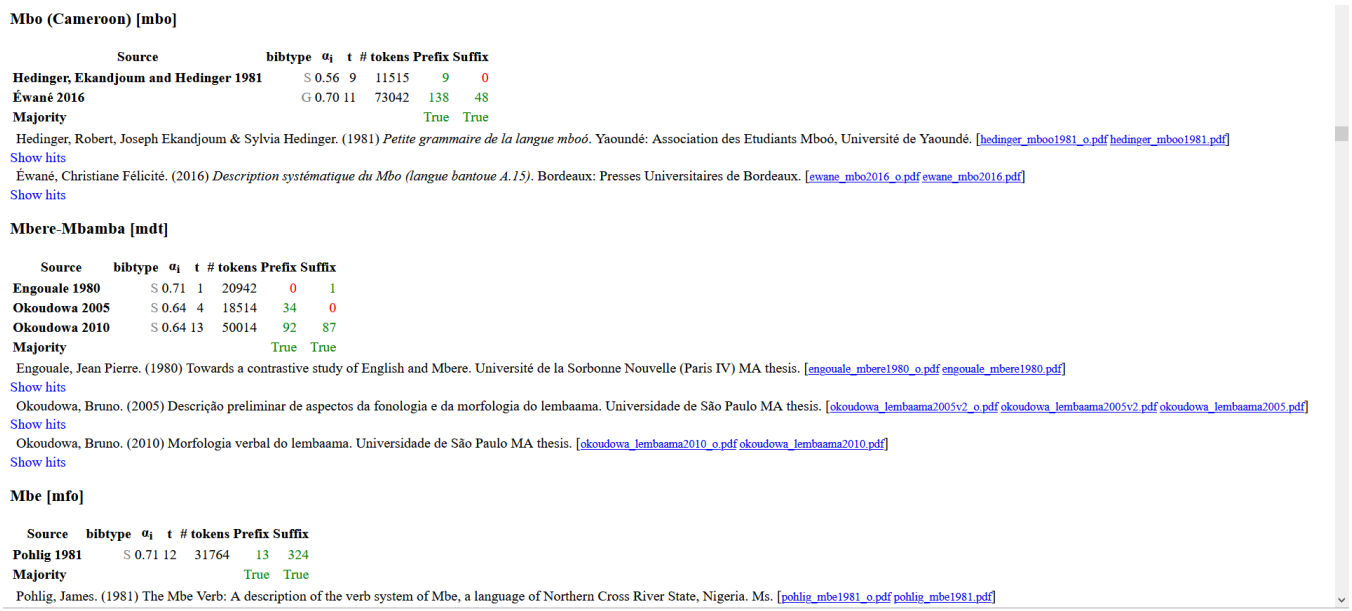
https://glottolog.org/resource/languoid/id/west2430

References

Showing 1 to 18 of 18 entries ← Previous 1 Next →

Details	Name	Title	Any field	ca	Year	Pages	Doctype	ca	Provider	da
	Search	Search	Search		Search	Search	--any--		--any--	
citation	Adam Sposato 2015	A grammar of Xong	✔		2015	674	grammar		hh	
citation	Xiang, Rizheng 1999	Jíwèi Miáoyǔ yánjiū	✔		1999	261	grammar		hh	
citation	Yu, Jinzhi 2011	Xiāngxī Āizhài Miáoyǔ cānkǎoyǔfǎ	✔		2011	7	grammar		hh	
citation	Adam Sposato 2021	A grammar of Xong	✔		2021		grammar		hh	
citation	Luo, Anyuan 2005	Songtao Miaohua Miaoxie Yufaxue	✔		2005		grammar		hh	
citation	Chunsheng Ling and Yifu Rui 1990 [1940]	Xiāngxī miáozú diàochá bào gào	✔		1940	504	ethnographic, grammar_sketch		hh	
citation	Chen, Hong 2009	Guizhōu sōngtáo dàxǐng zhèn miáo yǔ yánjiū	✔		2009	207	grammar_sketch		hh	
citation	Yang, Zaibiao 2004	Miaoyu Dongbu Fangyan Tuyu Bijiao	✔		2004	432	comparative, grammar_sketch		hh	
citation	Xian, Songkui 2000	Xin miao han cidian (xibu fangyan)	✔		2000	486	dictionary		hh, mpieva	
citation	Mao, Zhongwu and Xun Zhao 1992	Han Yao jianming fenlei cidian (mian yu)	✔		1992	18	dictionary		hh	
citation	Xiang, Rizheng 1992	Han Miao cidian (Xiangxi fangyan)	✔		1992		dictionary		hh	
citation	Adam Sposato 2014	Word order in Miao-Yao (Hmong-Mien)	✔		2014	58	specific_feature		degruyter, hh	
citation	Adam Sposato 2012	Relative Clauses in Xong (Miao-Yao)	✔		2012	18	specific_feature		hh	

Figure 1: Examples of references and metadata for the Western Xiangxi Miao language as it appears on glottolog.org.



Mbo (Cameroon) [mbo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Hedinger, Ekandjoum and Hedinger 1981	S	0.56	9	11515	9	0
Éwané 2016	G	0.70	11	73042	138	48
Majority					True	True

Hedinger, Robert, Joseph Ekandjoum & Sylvia Hedinger. (1981) *Petite grammaire de la langue mboó*. Yaoundé: Association des Etudiants Mboó, Université de Yaoundé. [hedinger_mboo1981_o.pdf hedinger_mboo1981.pdf]
Show hits

Éwané, Christiane Félicité. (2016) *Description systématique du Mbo (langue bantoue A.15)*. Bordeaux: Presses Universitaires de Bordeaux. [ewane_mbo2016_o.pdf ewane_mbo2016.pdf]
Show hits

Mbere-Mbamba [mdt]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Engouale 1980	S	0.71	1	20942	0	1
Okoudowa 2005	S	0.64	4	18514	34	0
Okoudowa 2010	S	0.64	13	50014	92	87
Majority					True	True

Engouale, Jean Pierre. (1980) Towards a contrastive study of English and Mbere. Université de la Sorbonne Nouvelle (Paris IV) MA thesis. [engouale_mbere1980_o.pdf engouale_mbere1980.pdf]
Show hits

Okoudowa, Bruno. (2005) Descrição preliminar de aspectos da fonologia e da morfologia do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2005v2_o.pdf okoudowa_lembaama2005v2.pdf okoudowa_lembaama2005.pdf]
Show hits

Okoudowa, Bruno. (2010) Morfologia verbal do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2010_o.pdf okoudowa_lembaama2010.pdf]
Show hits

Mbe [mfo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Pohlig 1981	S	0.71	12	31764	13	324
Majority					True	True

Pohlig, James. (1981) The Mbe Verb: A description of the verb system of Mbe, a language of Northern Cross River State, Nigeria. Ms. [pohlig_mbe1981_o.pdf pohlig_mbe1981.pdf]

Figure 2: Sample search/extraction output for the search terms `prefix` and `suffix`. For each language and corresponding grammatical descriptions, the number of hits is shown, alongside the automatically calculated threshold t (see Hammarström et al. 13 for details). The sources are spelled out with links to full-text and displayable hit snippets.

be formally evaluated. For the features `classifier` and `gender` the data mining tool out-of-the-box yields precision and recall rivalling humans. But for features such as `ideophones` or `noun class` recall is significantly lower, and the human needs to supplement the original search with a wider set of terms. For features such as `tone` or `locative` qualifications are needed to increase precision. `endangered` is a term which is associated with a degree such as "critically" and "highly" and trial and error is needed to strike the balance between the degree term and the closeness to "endangered" and its absense. The term "dying" is often used for "endangered"

North-Central Ju [knw]

Source	bibtype	α_i	t	# tokens	creaky	creaky vowel
Heikkinen 1986	P;W	1.00	1	9612	2	1
Heikkinen 1987	S	1.00	1	26503	1	1
Heine and König 2015	G	1.00	1	128913	1	1
König and Heine 2001	G	1.00	1	53001	1	1
Majority					True	True

Heikkinen, Terttu. (1986) Phonology of the !Xũ Dialect spoken in Ovamboland and Western Kavango. *South African Journal of African Languages* 6(1). 18-28. [heikkinen_xu-phonology1986.pdf heikkinen_xu-phonology1986_o.pdf heikkinen_xu-ovamboland-wkavango1987.pdf]

Show hits

- creaky

  [p. 2] n the open syllable: C -» r and q only in unstressed syllable, (or C -A r and q / in stressed syllable); CiV C2V-root: C2 -*■ only b,r,n,9, (m,s,x,w) The following vowels and diphthongs function in the open syllables: V -*• a single-quality vowel of one of the following sets: Oral Nasalized 20 S.-Afr. Tydskr. Afrikatale 1986, 6(1) The sign , underneath a vowel symbol indicates that the vowel is a **creaky** vowel. V -* VV, a double-quality vowel of one of the following sets: Oral Nasalized Fronted (unrounded) In the orthography used, the way of symbolizing diphthongs follows the direction taken in the new official orthography for Nama, another Namibian language which has diphthongs. Thus, for ei and bu, ai and au are preferred to ei and ou, which might also be chosen. The diphthong ue has bee

  [p. 2] -▶ (CO C2. 9m 'eat', firj 'see', 'rj|rj 'sit singular', in 'suck breast', rpm 'we inclusive', ijrj 'so that'; 'be long', na.Zm 'ostrich', zd',jVra 'be strong, not easily tom' This subtype of closed syllable is only apparently closed, for the nasal does the work of a vowel completely: it carries the tone, and it may also get breathy at the beginning of rising tones (or)). There are nasals which are **creaky**, just as some vowels are (iji). In most cases it is easy to imagine a suppressed vowel before the final nasal. However, it is difficult to see how any vowd could be present in for example the !Xu words for 'eat', 'we', and 'suck' (9m, ipm, ip); what vowel could be pronounced with closed lips? S ^ CiV C2. kx9am 'cut singular', kx?6m 'cut plural', J=9ag 'think', zdrj 'tip of upper lip', 'n^a

- creaky vowel

  [p. 2] n the open syllable: C -» r and q only in unstressed syllable, (or C -A r and q / in stressed syllable); CiV C2V-root: C2 -*■ only b,r,n,9, (m,s,x,w) The following vowels and diphthongs function in the open syllables: V -*• a single-quality vowel of one of the following sets: Oral Nasalized 20 S.-Afr. Tydskr. Afrikatale 1986, 6(1) The sign , underneath a vowel symbol indicates that the vowel is a **creaky vowel**. V -* VV, a double-quality vowel of one of the following sets: Oral Nasalized Fronted (unrounded) In the orthography used, the way of symbolizing diphthongs follows the direction taken in the new official orthography for Nama, another Namibian language which has diphthongs. Thus, for ei and bu, ai and au are preferred to ei and ou, which might also be chosen. The diphthong ue has bee

Heikkinen, Terttu. (1987) *An Outline of the Grammar of the !Xũ Language spoken in Ovamboland and West Kavango* (Suid-Afrikaanse Tydskrif vir Afrikatale 7 Byblad 1). Pretoria: African Language Association of Southern Africa. [heikkinen_xu-ovamboland-wkavango1987_o.pdf]

Figure 3: Example of an expanded (but within legal size) display of hits with highlighting for the search terms `creaky` and `creaky voice`.

This file shows the results of searching {grammar, phonology, grammar_sketch} written in Mandarin Chinese [cmn], German [deu], English [eng], French [fra], Italian [ita], Dutch [nld], Portuguese [por], Russian [rus], Spanish [spa] for:

Heading	[cmn] regexp(s)	[deu] regexp(s)	[eng] regexp(s)	[fra] regexp(s)	[ita] regexp(s)	[nld] regexp(s)	[por] regexp(s)	[rus] regexp(s)	[spa] regexp(s)
Tone	声调 OR 聲調	\W[T]onolog OR \W[T]onal OR \W[T]on\W OR \W[Tt]öne\W	\W[Tt]one[s]\W OR \W[Tt]onal OR \W[Tt]onolog	\W[Tt]on[es] OR \W[Tt]onal OR \W[Tt]onolog	\W[Tt]onolog OR \W[Tt]ona OR \W[Tt]ono	\W[Tt]oon\W OR \W[Tt]onal	\W[Tt]onolog OR \W[Tt]ona OR \W[Tt]on	\W[Тт]онол OR \W[Тт]онал OR \s[Тт]он\W OR \W[Тт]онем OR \W[Тт]она[мх] OR \W[Тт]оны OR \W[Тт]онов\W OR \W[Тт]ону\W	\W[Tt]onal OR \W[Tt]onolog OR \W[Tt]ono

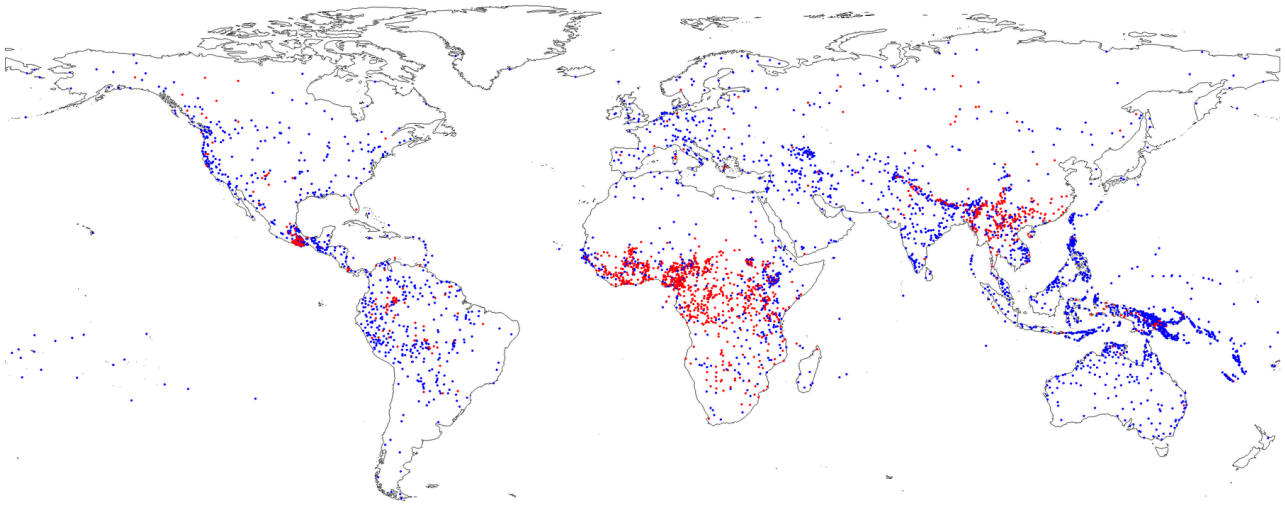
Download tab-separated file with consensus by language.

Download tab-separated file with assessment by source.

Results of extraction by language

Results of extraction by (sub-)family

Tone



False blue
True red

Figure 4: Examples regular expressions in different languages for the feature `tone`. The corresponding search/extraction results are returned on a map and in downloadble CSVs.

but experiments show that this word also occurs in example sentences so often that it is useless as an indicator of the vitality of the described language. The feature `spontaneous nasalization` is rare and here all documents (not only those which describe exactly one language) and all search hits (not only those with sufficiently many occurrences) are of interest. The term `partitive` is used for a variety of related phenomena, not all of which are relevant for a specific researcher. In the case at hand, the ability to restrict hits to the Uralic family with a keypress was decisive.

The search/extraction procedures can be further developed by standard NLP tools such as stemming, collocation recognition and vector-space semantics as well as further domain-specific document processing including ToC-recognition, section division [11], and bibliographical citation analysis Tkaczyk et al. 28.

An application has been submitted to the Centre for Digital Humanities Uppsala for durable hosting of Gramfinder.

4 CONCLUSION

There are good prospects for building cross-linguistic databases from the rich legacy of descriptive linguistic literature. The bulk of relevant documents are available in digital form for research purposes. Fully automated data mining on these documents is currently viable only for a minority of features of interest, yet the same data mining techniques can form the starting point for the remaining majority of features of interest. We present Gramfinder, a free and open search/data-mining infrastructure that responds to this demand. Through the seven design features metadata control, document-object correspondence, explainable data mining, inspection support, search infrastructure, export facilities and copyright compliance it aims to facilitate a dialectic iterated search/extraction procedure, speed up manual inspection/correction of results and enable rich CSV-exports and experimentation. The domain-specific support provides an edge over less targeted tools (such as Google Books). Informal usage confirms the speed-up in database curation and variability of user needs. The organization and document access alone provides an enormous time-save especially compared to fetching the corresponding documents in a library. The NLP and document analysis support can be developed further.

ACKNOWLEDGMENTS

This research was made possible thanks to the financial support of the From Dust to Dawn: Multilingual Grammar Extraction from Grammars project funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2017.0105 awarded to Harald Hammarström (Uppsala University) and the Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage (DReaM) Project awarded 2018-2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden.

REFERENCES

[1] Balthasar Bickel. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In *The Oxford Handbook of Linguistic Analysis* (2 ed.), Bernd Heine and Heiko Narrog (Eds.). Oxford: Oxford University Press, 901–923.

[2] Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp — the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 474–478.

[3] Francesca Di Garbo. 2014. *Gender and its interaction with number and evaluative morphology: An intra- and intergenealogical typological survey of Africa.* Ph.D. Dissertation. Stockholm University.

[4] Alexis Dimitriadis and Simon Musgrave. 2009. Designing linguistic databases: A primer for linguists. In *The Use of Databases in Cross-Linguistic Studies*, Martin Everaert, Simon Musgrave, and Alexis Dimitriadis (Eds.). Empirical Approaches to Language Typology [EALT], Vol. 41. Berlin: De Gruyter Mouton, 13–75.

[5] Nicholas Evans and Stephen Levinson. 2009. The Myth of Language Universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32, 5 (2009), 429–492.

[6] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering* 26 (2020), 259–291.

[7] Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Nature Scientific Data* 5, 180205 (2018), 1–10.

[8] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (December 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[9] Harald Hammarström, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg, and Bettina Speckmann. 2018. Simultaneous Visualization of Language Endangerment and Language Description. *Language Documentation & Conservation* 12 (2018), 359–392.

[10] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://glottolog.org. Accessed on 2021-05-20..

[11] Harald Hammarström. 2021. Inventory and Content Separation in Grammatical Descriptions of Languages of the World. In *Proceedings of 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021)*. Milton Keynes, UK: The Open University, in press.

[12] Harald Hammarström. 2021. Measuring Prefixation and Suffixation in the Languages of the World. In *Proceedings of The 3rd Workshop on Research in Computational Typology and Multilingual NLP*. Stroudsburg, PA: Association for Computational Linguistics (ACL), 81–89.

[13] Harald Hammarström, One-Soon Her, and Marc Tang. 2021. Term-Spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In *Proceedings of SLTC-2020*. NEJLT, 12pp.

[14] Harald Hammarström, One-Soon Her, Marc Tang, Olof Lundgren, Hilda Appelgren, and William Zetterberg. 2021. Automatically building a database of gender/noun class/classifiers from digitized grammatical descriptions. Paper presented at the Lund Gender Workshop, 12 Mar 2021, Lund University.

[15] Harald Hammarström and Sebastian Nordhoff. 2011. LangDoc: Bibliographic Infrastructure for Linguistic Typology. *Oslo Studies in Language* 3, 2 (2011), 31–43.

[16] Harald Hammarström, Shafqat Mumtaz Virk, and Markus Forsberg. 2017. Poor Man's OCR Post-Correction: Unsupervised Recognition of Variant Spelling Applied to a Multilingual Document Collection. In *Proceedings of the Digital Access to Textual Cultural Heritage (DATeCH) conference*. Göttingen: ACM, 71–75.

[17] One-Soon Her, Harald Hammarström, and Marc Tang. submitted. Introducing WACL: The World Atlas of Classifier Languages. *submitted* (submitted).

[18] Ryan James and Andrew Weiss. 2012. An Assessment of Google Books' Metadata. *Journal of Library Metadata* 12, 1 (2012), 15–22.

[19] Olga Krasnoukhova. 2012. *The Noun Phrase in the Languages of South America.* Ph.D. Dissertation. Radboud Universiteit Nijmegen.

[20] Stephen C. Levinson and Russell D. Gray. 2012. Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences* 16, 3 (2012), 167–173.

[21] Jayden L. Macklin-Cordes, Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao, and Erich R. Round. 2017. Robots who read grammars. Poster presented at CoEDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

[22] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

[23] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 6014 (2011), 176–182.

[24] Matti Miestamo. 2003. *Clausal negation: A typological study.* Ph.D. Dissertation. University of Helsinki.

[25] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics* 117 (2018), 1931–1990.

[26] Ger Reesink and Michael Dunn. 2012. Systematic typological comparison as a tool for investigating language history. In *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, Nicholas Evans and Marian Klamer (Eds.). Language Documentation & Conservation Special Publication, Vol. 5. Honolulu: University of Hawaii Press, 34–71.

[27] Hedvig Skirgård. 2014. The Nijmegen Typological Survey. Paper presennted in the MPI Nijmegen lunch talks series, Max Planck Institute for Psycholinguistics.

[28] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (Fort Worth, Texas, USA) *(JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3197026.3197048

[29] Matthias Urban and Steven Moran. 2021. Altitude and the distributional typology of language structure: Ejectives and beyond. *PLoS ONE* 16(2), e0245522 (2021), 1–36.

[30] Eva van Lier. 2009. *Parts of Speech and Dependent Clauses: A typological study*. Ph.D. Dissertation. Universiteit van Amsterdam.

[31] Shafqat Mumtaz Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic Extraction of Typological Linguistic Features from Descriptive Grammars. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, Kamil Ekštein and Václav Matoušek (Eds.). Lecture Notes in Computer Science, Vol. 10415. Berlin: Springer, 111–119.

[32] Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World's Languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Marseille, France, 871–877. https://www.aclweb.org/anthology/2020.lrec-1.109

[33] Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulagaria: NCOMA Ltd, 1247–1256.

[34] Søren Wichmann and Taraka Rama. 2019. Towards unsupervised extraction of linguistic typological features from language descriptions. First Workshop on Typology for Polyglot NLP, Florence, Aug. 1, 2019 (Co-located with ACL, July 28-Aug. 2, 2019).

[35] Roberto Zariquiey, Mónica Arakaki, and Harald Hammarström. 2021. Use of Terms Relating to Language Endangerment in the Descriptive Linguistic Literature. *In preparation* (2021).