elfen: A Python Package for Efficient Linguistic Feature Extraction for Natural Language Datasets

Maximilian Maurer
GESIS Leibniz Institute for the Social Sciences
Heinrich-Heine University Düsseldorf
maximilian.maurer@gesis.org

Abstract

A detailed understanding of the basic properties of text collections produced by humans or generated synthetically is vital for all steps of the natural language processing system life cycle, from training to evaluating model performance and synthetic texts. To facilitate the analysis of these properties, we introduce elfen, a Python library for efficient linguistic feature extraction for text datasets. It includes the largest set of item-level linguistic features in eleven feature areas: surface-level, POS, lexical richness, readability, named entity, semantic, information-theoretic, emotion, psycholinguistic, dependency, and morphological features. Building on top of popular NLP and modern dataframe libraries, elfen enables feature extraction in various languages (80 at the moment) on thousands of items, even given limited computing resources[1]. We show how using elfen enables linguistically informed data selection, outlier detection, and text collection comparison. We release elfen as an open-source PyPI package, accompanied by extensive documentation, including tutorials[2].

1 Introduction

While there is a dire need to understand our data at all levels, such as pre-training, fine-tuning, few-shot in-context learning, evaluation, and synthetically generated texts, there is a surprising lack of deep engagement with the basic properties of the data at hand. This is especially worrisome since works on spurious correlations in the data affecting model performance show that long-standing linguistic measures can provide insights into model behavior (Poliak et al., 2018; Liusie et al., 2022; Borah et al., 2023, inter alia). Similarly, recent works have emphasized the need for and promises of measuring data diversity (Nguyen and Ploeger, 2025). We argue that in line with these efforts, measuring the linguistic composition of texts is highly relevant, especially in the age of generative LLMs: Firstly, measuring linguistic properties of (pre-)training data at scale can give insights into downstream model behavior (e.g., Zhang et al., 2021). Secondly, given the popularity of benchmark datasets to assess improvements of ever bigger models, linguistic features can give insights into the comparability and the shortcomings of benchmark datasets (e.g., Gubelmann et al., 2024). Finally, given the rising prevalence of synthetic data, it becomes more and more important to understand its properties, be it to understand and detect it better (Dönmez et al., 2025) or to assess its utility for (pre-)training, especially given risks like model collapse (Shumailov et al., 2024).

While there are tools for linguistic feature extraction, most of them are focused on a specific area (e.g., lexical richness) or support a limited number and scope of features. Suppose a researcher wants a broad coverage of features in a given analysis. In that case, this causes difficulties, given that different tools require different preprocessing and can differ widely in how efficient they are, especially when dealing with large numbers of instances. There is a clear lack of availability of unified extraction tools providing a comprehensive number of features in different areas, and a way to efficiently extract them.

To fill this gap, we present elfen, a Python package to efficiently extract linguistic features for large numbers of text instances. Our contributions are fourfold: (1) We provide a tool with the largest collection of features ($1,061$), (2) most of which are extractable in 80 languages out of the box. (3) elfen provides efficient extraction (on

[1]For instance, the features for popular benchmarks can be extracted on consumer hardware in less than an hour. For more details, see Appendix D.

[2]We host the code at https://github.com/mmmaurer/elfen/, make it available through the GESIS Methods Hub at https://methodshub.gesis.org/library/methods/elfen/, and provide documentation and tutorials at https://elfen.readthedocs.io/en/latest/.

Library	elfen	LFTK	LIWC
Surface-Level	11	9	2
Lexical Richness	26	10	0
Readability	11	6	1
Named Entities	19	19	0
Information Theory	2	0	0
Emotion/Sentiment	36	0	8
POS	20	34	20
Psycholinguistics	78	3	33
Semantics	17	0	41
Morphology	798	0	0
Syntatice Dependencies	43	0	0
Reading Time Formulas	0	3	0
Total	1,061	84	105

Table 1: Comparison of the number of features implemented per feature area for `elfen` (`v1.2.4`), LFTK (Lee and Lee, 2023), and LIWC (Boyd et al., 2022). We keep the comparison to libraries with the same scope and goal as `elfen` . Due to different design choices regarding normalization by token, lemma, or sentence count, we only consider what they call *foundation* features in LFTK for this comparison. We count all non-*psycholinguistic* dictionary features as *semantic* for LIWC.

average 21.8% faster than comparable libraries) on tens of thousands of items. (4) `elfen` builds on popular libraries, allowing for easy integration into existing workflows and the multilingual coverage of it to grow with them.

In the following, we discuss existing tools and the contributions of `elfen` (Section 2), present the implementation and functionalities of it (Section 3), and showcase already existing and potential use cases of it (Section 4).

2 Related Work

Measuring characteristics of texts to compare them has a long history in (computational) linguistics and NLP, from early works trying to measure specific properties like lexical diversity (e.g., Yule, 1944) and readability (e.g., Mc Laughlin, 1969), to more recent advances trying to measure overall semantic similarity between texts (Corley and Mihalcea, 2005; Reimers and Gurevych, 2019)

In consequence, several tools for extracting such features have been developed, some of which are now outdated or no longer actively maintained (Graesser et al., 2004; Simig et al., 2022). Currently available, actively maintained tools can be categorized across two axes: (1) the scope of the features they provide, and (2) the units they operate on.

`Variationist` (Ramponi et al., 2024), for instance, mainly measures token, n-gram, and sequence occurrence frequencies on a (sub)corpus level, and, in turn, provides only a few corpus-level features. Conversely, there are tools focusing on text instance-level features, most of which focus on a single group of features, such as lexical richness (Shen, 2022) or readability[3] on a text-level, as well as tools for extracting token-level features such as `wn` (Goodman and Bond, 2021), a package for using open multilingual wordnets (Bond and Foster, 2013). More extensive tools like LFTK (Lee and Lee, 2023) and the commercial provider LIWC (Boyd et al., 2022) aim to cover broader sets of features on a text instance-level. While LFTK focuses on lexical richness and readability measures and frequencies of token-level features as characteristics of a given text, for example, the number of nouns or of named entities, LIWC mainly focuses on frequencies of words associated with certain semantic categories, for example, politics, or emotional or perceptive grounding.

2.1 Resource Gap

The fragmented coverage of feature groups, the lack of integration of token-level resources, such as psycholinguistic norms or emotion lexicons in open-source tools, and the fact that they are not optimized for datasets with large numbers of text instances make existing tools suboptimal for conducting analyses on typical NLP tasks and benchmark datasets, pre-training instances, and synthetically generated text collections. The package presented in this paper, `elfen` , thus aims to provide a unified tool for extracting an extensive number of text-level characteristics across feature areas, optimized for modern large NLP datasets. `elfen` provides a more comprehensive coverage across and within most feature areas, as we show in the comparison of `elfen` with LFTK and LIWC in Table 1, while being significantly faster than comparable tools when extracting large amounts of features.

3 Implementation and Functionality

3.1 Implementation

To allow for extensive analyses of datasets with tens of thousands of text items, we build `elfen` on top of Polars[4] for efficient parallel processing, and

[3] https://github.com/andreasvc/readability
[4] pola.rs

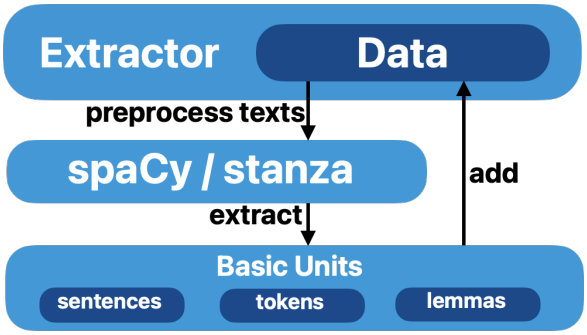

Figure 1: Schematic overview of preprocessing in `elfen`.

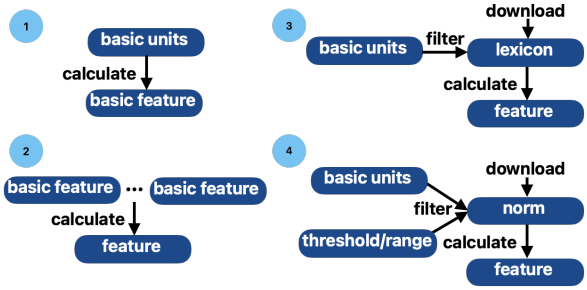

Figure 2: Schematic overview of the four types of feature extraction procedures.

spaCy (Honnibal et al., 2020) and stanza (Qi et al., 2020) to integrate with established NLP tools and pipelines.

As illustrated in Figure 1, `elfen` first preprocesses the texts in the dataframe used to initialize the `Extractor` with the backend model of choice from spaCy or stanza models. The extracted basic units, the sentences, tokens, lemmas, and syllables per text are stored in their respective columns in the dataframe, allowing for efficient downstream calculation of features based on these basic units.

From these basic units, `elfen` allows for the extraction of (implementation-wise) four types of features: (1) features directly computable from a basic unit (e.g., number of tokens), (2) features computable from multiple basic features (e.g., type-token ratio), (3) lexicon-based features requiring some basic unit and a lexicon to match the text with (e.g., the number of hedges), and (4) *norm*-based features requiring some basic unit and a *norm*, a lexicon including a measurement of a given property per basic unit (e.g., the *concreteness* of a given token).

As shown in Figure 2, (1) and (2) are directly calculated, utilizing parallel processing abilities from polars. This allows these features to be extracted in less than a second after initializing the `Extractor`,

Backend	Library	Runtime ± SD
`en_core_web_sm`	`elfen`	11.43 ± 0.51
	LFTK	15.46 ± 0.02
`en_core_web_md`	`elfen`	12.22 ± 0.26
	LFTK	15.92 ± 0.64
`en_core_web_lg`	`elfen`	11.99 ± 0.21
	LFTK	15.77 ± 0.36
`en_core_web_trf`	`elfen`	13.48 ± 0.23
	LFTK	15.64 ± 0.46

Table 2: Mean extraction time in seconds ± standard deviation (5 runs) for the first 100 items of the Stanford Sentiment Treebank (Socher et al., 2013) for all available features in `elfen` (1,061) and LFTK (220) using the same backend spacy models.

even for datasets with tens of thousands of items. For (3) and (4), if the license permits, the respective lexicons or norms will be automatically downloaded[5] and filtered according to the basic units present. For (4), if applicable, for example, if features based on *highly* concrete tokens are of interest, an additional filtering step is performed on the property measurement. Finally, the number of basic units fulfilling the filter criteria is counted, or a given property of their measurements is calculated (e.g., average concreteness of the tokens).

The optimized extraction of basic units, features, and filtering results in a considerable speedup over existing feature extraction tools. Given the same spaCy backbone models, `elfen` extracts all available 1,061 features on average 21.8% faster than the most comparable open-source library, LFTK (Lee and Lee, 2023), extracts all available 220 features (see Table 2).

3.2 Use Case-Driven Extraction

The main class of the package, the `Extractor`, is implemented with a special focus on the ease of use for various analysis scenarios. As illustrated in Figure 3, the `Extractor` provides the extraction in only a few lines of code of individual features, applicable in cases where researchers are interested in specific features, feature groups, in (the comparison of) a family of features, and all available features in `elfen`, for exploratory scenarios or when they

[5]The usage of the norms and lexicons is subject to different licenses. Complying with them and the citation guidelines is the user's responsibility. Some lexicons will need to be downloaded manually. Further details can be found in the documentation and the repositories' README.

```
# initializing extractor
extractor = elfen.Extractor(
                    data = df,
                    language = "en",
                    text_column = "text")
# extracting a single feature: ttr
extractor.extract("ttr")
# extracting a feature area/group: readability
extractor.extract_feature_group("readability")
# extracting all available features
extractor.extract_features()
```

Figure 3: Code examples of feature extraction capabilities. The `Extractor` here is initialized with a polars data frame `df`, which contains English text in the column *text*.

are interested in a comprehensive overview of the instances in a dataset.

3.3 Implemented Linguistic Features

`elfen` implements 1,061 features in eleven broad feature areas. Table 3 describes the feature areas and gives an example of a feature[6].

3.4 Multilingual Support

Given that `elfen` is using spaCy and stanza for preprocessing, we rely on the availability of language-specific models in them. SpaCy currently has 24 language-specific and one multilingual model available. Stanza provides 138 models in 80 languages.

All of the features except for the psycholinguistic norm-based, emotion, and semantic features are language-independent[7]. The emotion lexicons are available in 108 languages, psycholinguistic features are currently available in English, German, French, and Italian, and semantic features are available in all languages supported by the `wn` package (currently 34).

3.5 Analysis and Processing Utilities

Given `elfen`'s focus on linguistic analyses of text datasets, in addition to the extraction, we provide useful utilities for downstream analyses. These include extracted feature **rescaling** to specified ranges, **normalization** to have a mean of 0 and a standard deviation of 1, or by the number of tokens, lemmas, or sentences. We also provide functionality to extract local (within a given instance) and global (across the whole dataset) token and lemma frequencies.

4 Evaluation

To illustrate the usefulness of `elfen`, we discuss existing work already using it in three categories. We additionally outline three broad analysis scenarios showcasing how `elfen` provides insights in LLM-related research.

4.1 Existing Work Using `elfen`

4.1.1 Analysis of Human and LLM Behavior

`elfen` enables the analysis of human and language model behavior, and its connection to performance. It has, for example, already been used to assess linguistic factors in the human perception of gendered style of texts (Chen et al., 2025). Similarly, Falk and Lapesa (2025) to assess linguistic factors in annotation uncertainty in humans and models on morals and values. While Falk and Lapesa (2025) have a particular focus on human label variation and its connection to model uncertainty, in principle, any model-internal or output-derived metrics could be substituted to assess connections to structural characteristics of the texts. Thus, **`elfen` facilitates analyses connecting human and language model behavior with linguistic structure.**

4.1.2 Authorship and Stylistic Analysis

Tasks such as authorship attribution and stylistic analysis (e.g., Sari et al., 2018; Ayele et al., 2024) naturally use linguistic features due to their inherent need for interpretability. Zeng et al. (2025) use `elfen` as one potential interpretable component in their explainable authorship verification method. As this exemplifies, **`elfen` provides interpretability in tasks where it is vital, and can be integrated in respective systems for such tasks.**

4.1.3 Detection and Analysis of LLM-Generated Text

Another natural avenue of work is the analysis and detection of LLM-generated content. A major limitation of prior works in this line of work is the lack of access to extensive corpus statistics (Wu et al., 2025), which `elfen` alleviates. Parfenova et al. (2025), for instance, use `elfen` to

[6] For a more detailed description including references for the implemented features, see Appendix A. Our documentation provides additional information on each feature at https://elfen.readthedocs.io/en/latest/feature_overview.html

[7] We provide a periodically updated overview of which features are available in which language at https://elfen.readthedocs.io/en/latest/multilingual_support.html

Feature Area	Description	Example
Surface-level	Structural characteristics of a text	Number of tokens
Readability	Reading complexity; how hard to read a text is	Flesch reading ease
Psycholinguistics	Cognitive, social, or sensorimotor groundings of words	Number of highly concrete tokens
POS	Parts-of-speech in the text	Number of nouns
Morphology	Grammatical/lexical properties of words in a text	Number of plurals
Information theory	Redundancy and formulaicity of a text	Shannon entropy
Lexical richness	How lexically diverse is a text	Type-token ratio (TTR)
Syntactic Dependencies	Predicate-argument relations in a text	Number of adverbial modifiers
Semantics	Polysemy and ambiguity	Number of hedges
Named entities	Reference to entities with a proper name	Number of organizations
Emotion/Sentiment	Emotion or sentiment evoking or related words	Number of high arousal tokens

Table 3: Overview of feature areas with example features.

analyse convergence patterns in multi-agent annotation. Similarly, we show `elfen`'s utility for the case of LLM-written arguments, both for extensive analyses and detection scenarios (Dönmez et al., 2025). Thus, **`elfen` enables light-weight and interpretable detection and analyses of LLM-generated synthetic text and language model behavior.**

4.2 Exemplified Use Cases

To further illustrate the broad range of analyses `elfen` enables, we discuss three exemplary analysis steps that may be applied to many use cases: (1) Dataset comparison, (2) linguistically-informed targeted sampling, and (3) outlier detection.

We showcase these use cases[8] on two popular language understanding benchmark datasets, MMLU-Pro (Wang et al., 2024) and BigBench Hard (Suzgun et al., 2022).

4.2.1 Dataset Comparison

To understand dataset and domain effects at each step of the NLP life cycle, it is beneficial to understand in depth where datasets differ. This is particularly relevant for the generalization of training and test data: To train and test models for a given task and draw insights on models' capabilities, especially for benchmarking, we ideally want data to be as diverse as possible to reduce the influence of confounders. Suppose there are multiple datasets for a given task that differ structurally. It may be beneficial to either use the most diverse dataset or use multiple complementary datasets to get more robust results and more informative insights on model behavior and capabilities.

Two angles of assessment here are (1) a comparison of the overall and feature-area-wise correlation structure for a coarse overview, and (2) a comparison of how individual features are distributed for the datasets to assess fine-grained differences.

For (1), a natural option is to inspect correlation matrix heatmaps and similarity measures between (sub)matrices. As Figure 4 illustrates for BigBench Hard and MMLU-Pro, the correlation structure between two datasets on the same task can differ quite drastically, both overall and in specific areas, like morphological structures. This is reflected in similarity measures between the correlation (sub)matrices. For example, the Mantel correlation (Mantel, 1967) for morphological features (0.430, $p < 0.001$) is substantially lower than for surface-level features (0.925, $p = 0.005$)[9].

Given the particular differences in morphological features, it may be interesting for researchers to look more closely into such features, along the lines of (2). Such an analysis yields insights like MMLU-Pro showing considerably more variability in its usage of plural nouns ($\mu = 0.4$, $\sigma = 0.5$) than BigBench Hard ($\mu = 0.0$, $\sigma = 0.0$)[10].

Differences like these may be expected, given that BigBench Hard integrates different problems into a given template format per subtask, and, in MMLU-Pro, each instance has a specific problem-instruction combination. `elfen` helps to quantitatively confirm such intuitions, which illustrates how **`elfen` facilitates linguistic comparisons of datasets along given axes of interest**.

[8]The code for the exemplified use cases is available as commented notebooks at https://github.com/mmmaurer/elfen-examples. The enriched datasets are available at https://huggingface.co/collections/mmmaurer/enriched-language-understanding-benchmarks

[9]For a full table with Mantel correlation results, see Appendix E.1.

[10]For full statistics for morphological features, see Appendix E.2

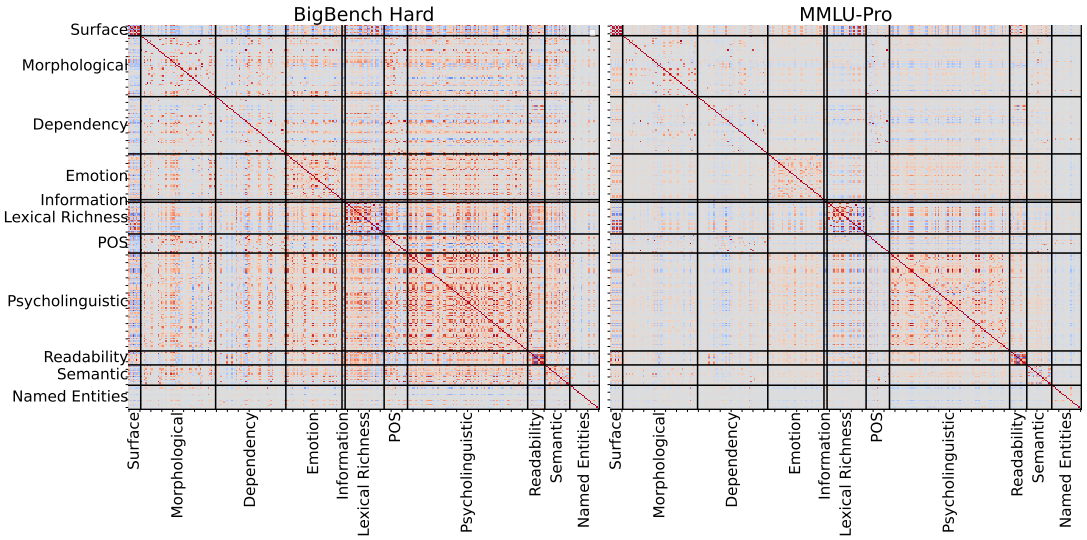

Figure 4: Heatmap of the correlation structures of BigBench Hard and MMLU-Pro for the eleven feature areas.

4.2.2 Targeted Sampling

Given the potential effects of structurally very different texts in a dataset, looking into respective samples for targeted comparisons or shot selection in few-shot scenarios can be beneficial. Following the latter example, suppose we want to ensure the examples for computer science problems in our shot selection include texts with relatively many tokens (> 50) and a high relative frequency of nouns (> 0.25). `elfen` provides the respective statistics of the subset overall, allowing for targeted sampling. For comparison, given that only 10.7% of instances in the computer science problems of MMLU-Pro fulfill these desiderata, the likelihood of having at least one such instance in a random sample is 0.203 for two shots and 0.365 for four shots. As this exemplifies, **`elfen` enables targeted sampling to use subsets of datasets with specific characteristics for downstream experiments.**

4.2.3 Outlier Detection

For linguistic bias-aware error analyses, it can be beneficial to understand whether a given model behaves differently for datapoints that are *outliers* in (a subset) of their structural characteristics. The features `elfen` provides can be used to run quick analyses to identify such outliers and inspect downstream effects when training or testing on them.

To showcase this, we construct linguistic fingerprints of the instances of MMLU-Pro by concatenating their respective features. We then use the local outlier factor to determine such outliers. Figure 5 shows a t-SNE projection of the fingerprints, showing that the identified outliers are either isolated or on the edges of clusters.

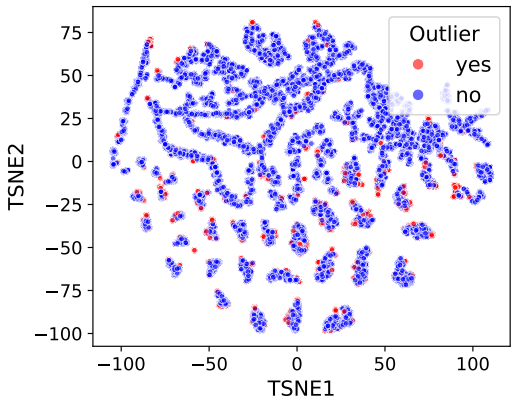

Figure 5: t-SNE projection of MMLU-Pro linguistic fingerprints showing outliers.

A researcher interested in the sensitivity of different models to confounding characteristics could, for example, use such identified outliers to test whether there are systematic differences in how models perform on these instances specifically. In line with the previous subsection, if such differences are present, few-shot selection of such outliers could be tested as a way to address this.

Overall, this shows how **`elfen` helps to identify instances in datasets warranting special attention in experiments and downstream analyses.**

5 Conclusion

We presented `elfen`, a Python package to efficiently extract linguistic features for text datasets, building on existing NLP preprocessing libraries and established resources. `elfen` provides the most

extensive number of features of comparable tools, spanning eleven feature areas. We showcased the versatility of `elfen` on prior work that already used it, and on three generalizable analysis use cases.

Ethical Considerations and Limitations

While the features provided in `elfen` are grounded in linguistic theory and draw on rigorously motivated and collected external resources, they should not be viewed as perfect or absolute properties, but rather as potentially noisy proxies for the underlying structures. This is mainly due to three limitations of the theories and resources `elfen` builds on:

Firstly, not all features can be expected to transfer across languages. For instance, Mandarin often relies on syntactic order rather than inflectional morphology to encode grammatical relationships. Compared with many Indo-European languages, the same values for some features may lead to very different conclusions. **We thus encourage researchers to critically engage with what `elfen`-derived features measure and reveal about linguistic realization when comparing them across languages.**

Secondly, we rely on existing tokenizers. While these tokenizers may be expected to work virtually perfectly in *well-behaved* text, they may not work as well in the presence of linguistic and orthographic variation such as dialects and sociolects (Wegmann et al., 2025). If this is not taken into account in the interpretation of results, this can lead to wrong inferences. This is clearly particularly problematic when the object of study includes the behavior of (groups of) humans. **Given the risk of flattening or misrepresenting groups or their language, we urge researchers using `elfen` to carefully assess whether off-the-shelf tokenizers can handle the variation present in their data.**

Thirdly, external measurements such as psycholinguistic norms or affective dictionaries are subject to limitations that are passed down to the features in `elfen` based on them. Besides limitations in the way ratings are collected (Mohammad, 2018a; Delatorre et al., 2019), the main concern is that most of them are collected from Western, well-educated, rich, and politically liberal[11], *WEIRD* (Henrich et al., 2010), study participants and their language variants, causing a bias in both the selection of lexical items and the measurements (Siew et al., 2025). Finally, aggregated ratings may flatten individual differences (c.f. Knupleš et al., 2023; Paisios et al., 2023), resulting in a simplified picture of the complex reality of language perception. While these are problems outside of the scope of `elfen` itself, and we continually update the included resources, **we urge caution for the inferences researchers make from psycholinguistic and affective features, as they may result from a WEIRD viewpoint on a limited number of lexical items, particularly for European languages.**

[11] *Liberal* here refers to the usage of the term in the US political landscape.

Acknowledgements

We thank Gabriella Lapesa and Vigneshwaran Shankaran for their helpful comments on earlier versions of this manuscript. We thank the members of the Computational Social Science department at GESIS, and early adopters for feedback on usability, errors, and useful missing features in early versions of `elfen`.

References

Jonathan Anderson. 1981. Analysing the Readability of English and Non-English Texts in the Classroom with Lix. *Seventh Australian Reading Association Conference*, pages 1–13.

Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, and 10 others. 2024. Overview of pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification condensed lab overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 231–259, Cham. Springer Nature Switzerland.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Patrick Bonin, Aurélie Méot, and Aurélia Burgiska. 2018. Concreteness ratings for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods*, 50:2366–2387.

Angana Borah, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2023. Measuring spurious correlation in classification: "clever hans" in translationese. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 196–206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51:467–479.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911.

John Bissell Carroll. 1964. *Language and Thought*. Prentice-Hall.

Hongyu Chen, Neele Falk, Michael Roth, and Agnieszka Falenska. 2025. "feels feminine to me": Understanding perceived gendered style through human annotations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31447–31468, Suzhou, China. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall and. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

P. Delatorre, A. Salguero, C. León, and A. Tapscott. 2019. The impact of context on affective norms: A case of study with suspense. *Frontiers in Psychology*, 10:1988.

Veronica Diveica, Penny M. Pexman, and Richard J. Binney. 2023. Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods*, 55(2):461–473.

Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. 2025. AI argues differently: Distinct argumentative and linguistic patterns of LLMs in persuasive contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34583–34614, Suzhou, China. Association for Computational Linguistics.

Daniel Dugast. 1978. Sur quoi se fonde la notion d'etendue theoratique du vocabulaire? *Le francais Modern*, 46(1):25.

Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921, Vienna, Austria. Association for Computational Linguistics.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2024. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33(1):21–48.

Pierre. Guiraud. 1954. *Les caracteères statistiques du vocabulaire : essai de méthodologie*. Presses universitaires de France, Paris.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83.

Gustav Herdan. 1955. A new derivation and interpretation of Yule's 'Characteristic' K. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 6:332–339.

Gustav Herdan. 1964. *Quantitative Linguistics*. Butterworths.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Philipp Kanske and Sonja A. Kotz. 2010. The leipzig affective norms for german: A reliability study. *Behavior Research Methods*, 42(4):987–991.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability

Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Institute for Simulation and Training.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the nature of disagreements on mid-scale ratings: A case study on the abstractness-concreteness continuum. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–86, Singapore. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Moira Linnarud. 1987. Lexis in composition: a performance analysis of swedish learners' written english. *Studies in Second Language Acquisition*, 9:254 – 256.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 78–84, Online only. Association for Computational Linguistics.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52:1271–1291.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220.

Heinz-Dieter Mass. 1972. Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

G. Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad. 2018b. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Maria Montefinese, Elena Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) in italian. *Behavior Research Methods*, 46:887–903.

Maria Montefinese, David Vinson, Gabriella Vigliocco, and Ettore Ambrosini. 2019. Italian age of acquisition norms for a large set of words (itaoa). *Frontiers in Psychology*, Volume 10 - 2019.

Pascale Moreira and Yuri Bizzoni. 2023. Dimensions of quality: Contrasting stylistic vs. semantic features for modelling literary quality in 9,000 novels. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 739–747, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Dong Nguyen and Esther Ploeger. 2025. We need to measure data diversity in nlp – better and broader. *Preprint*, arXiv:2505.20264.

Dimitri Paisios, Nathalie Huet, and Elodie Labeye. 2023. Addressing the elephant in the middle: Implications of the midscale disagreement problem through the lens of body-object interaction ratings. *Collabra: Psychology*, 9(1):84564.

Angelina Parfenova, Alexander Denzler, and Jürgen Pfeffer. 2025. Emergent convergence in multi-agent LLM annotation. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 206–225, Suzhou, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages

180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alan Ramponi, Camilla Casula, and Stefano Menini. 2024. Variationist: Exploring multifaceted variation and bias in written language data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 346–354, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Brian J. Richards and David D. Malvern. 1997. *Quantifying lexical diversity in the study of language development*. University of Reading, Faculty of Education and Community Studies.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sabine Schröder, Teresa Gemballa, Steffie Ruppin, and Isabelle Wartenburger. 2011. German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, 44:380–394.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Herbert S. Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547.

Cynthia S. Q. Siew, Feria Chang, and Jin Jye Wong. 2025. Investigating the effects of valence, arousal, concreteness, and humor on words unique to singapore english. *Journal of Cognition*.

Daniel Simig, Tianlu Wang, Verna Dankers, Peter Henderson, Khuyagbaatar Batsuren, Dieuwke Hupkes, and Mona Diab. 2022. Text characterization toolkit (TCT). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 72–87, Taipei, Taiwan. Association for Computational Linguistics.

Edward H. Simpson. 1949. Measurement of Diversity. *Nature*, 163.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Mildred C. Templin. 1957. *"Certain Language Skills in Children: Their Development and Interrelationships"*, ned - new edition edition, volume 26. University of Minnesota Press.

Alessandra Vergallito, Marco Alessandro Petilli, and Marco Marelli. 2020. Perceptual modality norms for 1,121 italian words: A comparison with concreteness and imageability scores and an analysis of their impact on word processing tasks. *Behavior Research Methods*, 52:1599–1614.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is sensitive to language variation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10958–10983, Vienna, Austria. Association for Computational Linguistics.

Bodo Winter, Gary Lupyan, Lynn K Perry, Mark Dingemanse, and Marcus Perlman. 2024. Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, 56(3):1640–1655.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

George U. Yule. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

Peter Zeng, Pegah Alipoormolabashi, Jihu Mun, Gourab Dey, Nikita Soni, Niranjan Balasubramanian, Owen Rambow, and H. Schwartz. 2025. Residualized similarity for faithfully explainable authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15824–15837, Suzhou, China. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

A Full Description of Features

In the following, we provide a more detailed description of the available features per feature area, including references to relevant literature introducing or describing them.

Surface-Level Features provide structural characteristics of the texts. The package provides extraction of the sequence length (characters; both with and without whitespaces), number of tokens, sentences, types, lemmas, long words (over six characters), and token frequencies on an item level. Based on those, the number of tokens per sentence, characters per sentence, and average word length can be extracted.

Readability Features were proposed to measure the complexity of texts. Following the `readability` python package[12], the package extracts the Gunning fog index, ARI, Flesch reading ease, and Flesch-Kincaid grade level (Kincaid et al., 1975), the Cole-Liau index (Coleman and Liau, 1975), SMOG (Mc Laughlin, 1969), LIX (Björnsson, 1968), and RIX (Anderson, 1981). Additionally, the package provides extraction of the basic features necessary for calculating these readability scores: The number of syllables in a text, words with only one syllable, and words with more than two syllables.

Psycholinguistic Norm Features measure words' cognitive, social, and sensorimotor grounding. We use concreteness norms[13] (Brysbaert et al., 2014), i.e. how concrete or abstract is a given word, word prevalence norms (Brysbaert et al., 2019), i.e. how well-known/-used is a word, Age-of-Acquisition norms (Kuperman et al., 2012), i.e. at what age do children learn a given word, Socialness norms (Diveica et al., 2023), i.e. how socially relevant is a words' meaning, Iconicity norms (Winter et al., 2024), i.e. to which degree the sound of a word reflects its' meaning, and Sensorimotor norms (Lynott et al., 2020), i.e. how connected a words' meaning is to perceptual modalities (e.g. visual) and action effectors (e.g. arm/hand). Per item and for each norm, the package implements the extraction of the average rating of all tokens from the item in the norm lexicon, their average standard deviation in the ratings, the number of tokens with a high rating (upper third of the Likert scale), a low rating (lower third of the scale), and the number of tokens with a particularly high standard deviation (such that the ratings span over multiple thirds of the scale).

While the norms are collected for individual words without context, these features are included to measure potential effects of the presence of words with a particular grounding or ambiguity thereof.

Part-of-Speech Features. Per item, the package provides extraction of the number of tokens per universal dependencies POS tag (de Marneffe et al., 2021), the number of lexical tokens (nouns, verbs, adjectives, and adverbs), and the POS variability (number of different POS tags relative to the number of tokens).

Lexical Richness Measures provide information about how lexically diverse a given text is. Intuitively, the more lexically rich a text is, the more different words a text contains. Following the `lexicalrichness` python package (Shen, 2022), per item, `elfen` allows for the extraction of the type-token ratio (TTR) (Templin, 1957), root TTR (Guiraud, 1954), corrected TTR(Carroll, 1964), Herdan's C(Herdan, 1964), Summer's TTR, Dugast's Uber index(Dugast, 1978), Maas' TTR (Mass, 1972), Yule's K (Yule, 1944), Herdan's V_m (Herdan, 1955), Simpson's D (Simpson, 1949), mean segmental TTR (Richards and Malvern, 1997), moving average TTR (Covington and and, 2010), measure of textual lexical diversity (MLTD, McCarthy and Jarvis, 2010), and the hypergeometric distribution diversity (HD-D, McCarthy and Jarvis, 2007, 2010). Additionally, the lo-

[12] https://github.com/andreasvc/readability

[13] All of the cited norms here are in English. Find the full list of currently supported languages per psycholinguistic dimension, including references in Appendix B. We regularly add more.

cal and global numbers of hapax (dis)legomena (i.e. the number of words per item that occur only once/twice per item/globally in the dataset), Sichel's S (Sichel, 1975), and the lexical density, i.e., the percentage of lexical tokens (Linnarud, 1987) are extractable.

Morphological Features. `elfen` allows for the extraction of the number of tokens with a given morphological feature for all available universal dependencies morpho-syntactic features (de Marneffe et al., 2021).

Information-Theoretic Features. As a measure of redundancy or formulaicity, following Moreira and Bizzoni (2023), `elfen` implements the compressibility of the text per item. The compressibility is defined as the bit length of the compressed text divided by the bit length of the uncompressed text. `elfen` implements the average token Shannon entropy per item to measure predictability.

Dependency Features provide information about the morphosyntactic realizations of predicate-argument structures. `elfen` implements the number of dependency relation types (according to Universal Dependencies, de Marneffe et al., 2021), the number of noun chunks in the text, the tree width, i.e. the maximum number of nodes in the subtree of a token, the tree depth, i.e the maximum distance of a token from the root of the dependency tree, the tree branching factor, i.e. the average number of children of a token, and the ramification factor, i.e. the mean number of children per level in the dependency tree.

Semantic Features. To measure the impact of token-level ambiguity/polysemy on the text, we extract Open Multilingual Wordnet synsets (Bond and Foster, 2013) for all nouns, adjectives, and verbs using the `wn` python library (Goodman and Bond, 2021). Given these synsets, per item, we extract the average size of the synsets, the number of tokens with a large synset (more than four senses), and the number of tokens with a small synset (less than three senses) for nouns, adjectives, and verbs, respectively, and overall.

We extract the number of hedges[14] (i.e., words expressing speaker uncertainty; e.g., *might*, *presumably*, or *maybe*) and the hedge-token ratio per item as a measure of the presence of uncertainty expressions in the text.

[14] https://github.com/words/hedges

Named Entity Features. Per item, we extract the number of named entities overall and per entity type (e.g. names, locations, organizations, etc.).

Emotion and Sentiment Features. To measure the effects of the occurrence of words commonly associated with/evoking a given emotion or sentiment, we use the NRC-VAD lexicon (Mohammad, 2018a) for valence, arousal, and dominance, the NRC emotion intensity lexicon (Mohammad, 2018b) for the emotion intensity per basic emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), and the NRC word-emotion association lexicon (Mohammad and Turney, 2010, 2013) for sentiment. Per item and emotion dimension, we extract the average rating of all tokens from the item in the emotion lexicon, the number of tokens with a high rating, and the number of tokens with a low rating. For sentiment, per item, we extract the number of positive and negative sentiments, and the difference between them, normalized by the total number of tokens in the item.

B Norms/Lexicons per Language

Table 4 gives an overview of the availability of currently (`v1.2.4`) supported languages per psycholinguistic variable.

Variable	Language	Reference
Concreteness	DE	Kanske and Kotz (2010)
	EN	Brysbaert et al. (2014)
	FR	Bonin et al. (2018)
	IT	Montefinese et al. (2014)
Age of Acquisition	DE	Schröder et al. (2011)
	EN	Kuperman et al. (2012)
	IT	Montefinese et al. (2019)
Sensorimotor	EN	Lynott et al. (2020)
	IT	Vergallito et al. (2020)

Table 4: Psycholinguistic norms included in `elfen v1.2.4`.

C Additional Code Examples

Figure 6 gives an additional code example for normalization and rescaling. Figure 7 gives a code example of utilities included in `elfen`.

D Extraction of MMLU-Pro and BigBench Hard

We extract both MMLU-Pro and BigBench Hard on an Apple MacBook Pro with 24GB RAM and an Apple M4 chip. Table 5 shows the preprocessing

```
# rescale "n_tokens" to a range between 0 and 1
extractor.rescale("n_tokens",
                  minimum = 0,
                  maximum = 1)
# token-normalize "n_entities"
extractor.token_normalize("n_entities")
# normalize all features to a mean 0 std 1
extractor.normalize("all")
```

Figure 6: Code examples of token and zero-mean normalization, and rescaling.

```
# list all external resources available in elfen
elfen.list_external_resources()
# get a bibtex string for all resources
print(elfen.get_bibtex())
```

Figure 7: Code examples for utilities.

and extraction times with the number of instances per dataset.

Dataset	Size	Preprocessing	Extraction
MMLU-Pro	12,032	86.07s	812.86s
BigBench Hard	6,511	67.25s	453.96s

Table 5: Preprocessing and extraction times for MMLU-Pro and BigBench Hard.

E Full Results Use Cases

This section presents the full results for the use cases in section 4.

E.1 Full Mantel Results

Table 6 provides the full Mantel test results of the analysis use case in Section 4.2.1.

E.2 Full Morphological Feature Comparison

Table 7 provides the full descriptive statistics for the analysis of morphological features in the use case presented in Section 4.2.1.

Feature Group	Mantel	p-value
Surface	0.925	0.005
Morphological	0.430	0.000
Dependency	0.304	0.000
Emotion	0.444	0.000
Lexical Richness	0.735	0.000
POS	-0.026	0.785
Psycholinguistic	0.706	0.000
Readability	0.958	0.000
Semantic	0.555	0.000
Named Entities	0.519	0.004
All	0.481	0.000

Table 6: Mantel correlations including p-values per feature area between the correlation matrices of BigBench Hard and MMLU-Pro.

	BigBench Hard				MMLU-Pro			
Feature	μ	σ	min	max	μ	σ	min	max
n_NOUN_Number_Plur	0.00	0.00	0.00	0.01	0.05	0.05	0.00	0.67
n_VERB_VerbForm_Inf	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.33
n_PRON_Number_Sing	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.22
n_VERB_VerbForm_Part	0.00	0.00	0.00	0.02	0.03	0.04	0.00	0.50
n_PROPN_Case_Nom	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
n_PROPN_Number_Sing	0.00	0.00	0.00	0.03	0.05	0.06	0.00	0.62
n_VERB_Mood_Ind	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.25
n_PRON_Case_Acc	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.14
n_PRON_PronType_Prs	0.00	0.00	0.00	0.01	0.01	0.03	0.00	0.23
n_PUNCT_PunctType_Dash	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.22
n_NOUN_Number_Sing	0.00	0.00	0.00	0.03	0.19	0.08	0.00	1.00
n_PRON_Case_Nom	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.18
n_PUNCT_PunctType_Peri	0.00	0.00	0.00	0.01	0.05	0.03	0.00	0.33
n_DET_Number_Sing	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.10
n_PRON_PronType_Dem	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.17
n_PRON_PronType_Ind	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
n_VERB_Aspect_Prog	0.00	0.00	0.00	0.01	0.01	0.03	0.00	0.50
n_ADJ_Degree_Pos	0.00	0.01	0.00	0.05	0.06	0.05	0.00	0.50
n_PRON_Reflex_Yes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11
n_PRON_Gender_Masc	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.15
n_CCONJ_ConjType_Cmp	0.00	0.00	0.00	0.04	0.02	0.03	0.00	0.36
n_VERB_Person_3	0.00	0.00	0.00	0.01	0.01	0.03	0.00	0.33
n_PRON_Person_1	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.21
n_PRON_PronType_Art	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.33
n_VERB_Tense_Pres	0.00	0.00	0.00	0.01	0.03	0.04	0.00	0.50
n_DET_Definite_Def	0.00	0.00	0.00	0.01	0.07	0.05	0.00	0.33
n_PUNCT_PunctType_Quot	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.36
n_PUNCT_PunctType_Comm	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.37
n_PRON_Gender_Neut	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.17
n_VERB_Aspect_Perf	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.50
n_PROPN_Number_Plur	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.17
n_PRON_Gender_Fem	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.18
n_VERB_Number_Sing	0.00	0.00	0.00	0.01	0.01	0.03	0.00	0.33
n_VERB_Tense_Past	0.00	0.00	0.00	0.01	0.03	0.03	0.00	0.50
n_PUNCT_PunctSide_Fin	0.00	0.00	0.00	0.04	0.01	0.03	0.00	0.33
n_PUNCT_PunctType_Brck	0.00	0.01	0.00	0.07	0.02	0.04	0.00	0.46
n_ADJ_Degree_Sup	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.25
n_PRON_Poss_Yes	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.17
n_PRON_Person_3	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.18
n_NUM_NumType_Card	0.00	0.00	0.00	0.02	0.05	0.06	0.00	0.68
n_PRON_PronType_Rel	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.20
n_PRON_Person_2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.22
n_DET_Definite_Ind	0.00	0.00	0.00	0.01	0.03	0.04	0.00	0.50
n_ADJ_Degree_Cmp	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.25
n_DET_Number_Plur	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.12
n_VERB_VerbForm_Fin	0.00	0.00	0.00	0.01	0.03	0.03	0.00	0.40
n_PRON_Number_Plur	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.21
n_PUNCT_PunctSide_Ini	0.00	0.00	0.00	0.04	0.01	0.02	0.00	0.33

Table 7: Full overview of statistics (mean μ, standard deviation σ, $\min$ and $\max$) comparing BigBench Hard and MMLU-Pro on morphological features.