DoPA Meter — A Tool Suite for Metrical Document Profiling and Aggregation

Christina Lohr[1,2,3] **& Udo Hahn**[1,3]
[1]Jena University Language & Information Engineering (JULIE) Lab,
Friedrich Schiller University Jena, Germany
[2] Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
University Leipzig, Germany
[3]SMITH Consortium of the German Medical Informatics Initiative

Abstract

We present DoPA Meter, a tool suite for the metrical investigation of written language, that provides diagnostic means for its division into discourse categories, such as registers, genres, and style. The quantitative basis of our system are 120 metrics covering a wide range of lexical, syntactic, and semantic features relevant for language profiling. The scores can be summarized, compared, and aggregated using visualization tools that can be tailored according to the users' needs. We also showcase an application scenario for DoPA Meter.

1 Introduction

The way how we encode contents in natural language utterances gives rise to linguistic divisions into registers, genres, style levels, etc. (for a thorough distinction of these terms, see Lee (2001); Biber and Conrad (2019)) that follow functional communication requirements, e.g., ease of comprehension or adherence to the wording of social peer groups. The behavioral traits indicating such divisions are manifold and range from simple token frequencies, lexical choice options (synonyms, more specific vs. more general or sublanguage vs. layman terms), via syntactic variations (easy vs. complex sentence constructions) over to pragmatic distinctions (e.g., formal vs. informal language use). Many of NLP's most pressing applied research questions (e.g., hate and fake detection, communication biases relating to people's political, religious, racial, personal orientation) are considered to be flagged this way (Xiao et al., 2022).

In this paper, we address a large variety of such behavioral aspects of language use from a metrical perspective. None of these metrics is new, but their assembly and broad coverage in a coherent tool suite and modular software framework is. We also provide means for summarization, comparison and aggregation of results and their proper visualization.

2 Related Work

The tool-based computational analysis of behavioral traits of language use can be divided into three branches of research: (1) *readability* checkers with language complexity measures incorporating mostly surface-level syntactic and lexico-semantic features of utterances, (2) *stylometrics* tools with strong emphasis on powerful lexico-statistical metrics, and (3) *psychometrics* devices with mostly simple frequency-based computations complemented by dictionaries with psychologically typed lexical categories.

From the perspective of *readability* (for a survey, see Collins-Thompson (2014)), the DeLite system (vor der Brück et al., 2008) can be considered as one of the language profiling systems closest to the design goals and feature types of our system. Still, its main goal, as a readability checker, is much narrower than ours. DeLite identifies and highlights passages of text which are difficult to understand (together with reasons why this is the case). To reach this goal, DeLite comes with a wide range of shallow and deep features to score the readability of documents, which is also at the heart of our work. Deep features include, e.g., topological information from dependency trees for syntactic scoring (e.g., center embedding depth, phrasal fan-out ratios) and from semantic networks for semantic scoring (number of readings per lexical entry, number of propositions per sentence, semantic network connectivity). Altogether, 48 indicators for readability at the morphological, lexical, syntactic and semantic level can be calculated, averaged per document, and a global document readability score is finally computed by applying a k-nearest neighbor classifier. The system ran on German and English input data, yet has, to the best of our knowledge, never been made publicly accessible.

In the field of *stylometrics* (for a survey, see Neal et al. (2017)), Stylo (Eder et al., 2016) has

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 218–228
December 6-10, 2023

become a *de facto* standard for the quantitative study of writing style. STYLO is an R package equipped with powerful statistical analysis modules for analytics based on frequency measurements of character- and token-based n-grams (PoS n-grams etc., not supplied by default, require externally preprocessed input). STYLO comes in two flavors. Its API allows to configure a complete processing pipeline using traditional R scripting, while it also offers a rich graphical user interface (GUI) for non-technical users to run stylometric analyses and interpret their outcome without the need for elaborate programming experience.

The seamless integration of various analytical tools under a common programming framework (making use of R's core library but also extending it by various clustering algorithms and machine learning classifiers) and its public accessibility on GITHUB[1] make STYLO a landmark development for stylometric tooling. Yet, STYLO does not integrate any deeper lexical, syntactic and semantic processing going beyond textual surface computations (such as distance metrics, e.g., Burrows's Δ, very popular in the stylometric community).

The third stream of work emphasizes human lexical choice patterns in terms of the *psychometrics* of word use. Perhaps its most prominent representative is the *Linguistic Inquiry and Word Count* approach and its associated LIWC engine (Tausczik and Pennebaker, 2010).[2] LIWC's focus is on a categorically stratified dictionary resource (the current master dictionary comprises 6,400 words, word stems, and selected emoticons) with simple descriptive statistical tools though. LIWC reads documents word-by-word, matches each word with its dictionaries and outputs simple frequency-based lexical and PoS statistics. Overall more than 80 psychologically relevant categories ranging from linguistic ones (such as function vs. content words, parts of speech, tense markers) to psychological ones (such as Cognitive, Perceptual, and Biological Processes) are attached to single lexical entries and counted during text analysis.

LIWC was recently compared and outperformed by the SEANCE system (Crossley et al., 2017) which makes use of a range of newer, even more specialized dictionaries with a larger number of more expressive psychological categories and variables and a higher coverage of entries. Crossley et al. (2019) use a battery of independent systems for their experiments, each one highly specialized for computing different dimensions of readability, such as syntactic complexity (177 indices from the TAASSC system (Kyle and Crossley, 2018)), lexico-semantic frequency and richness (135 indices from the TAALES system (Kyle and Crossley, 2015)), text cohesion (over 150 indices from the TAACO system (Crossley et al., 2016)), and sentiment and social cognition scores (20 indices from the SEANCE system (Crossley et al., 2017)). Hence, roughly 500 individual scores have to be assembled from these stand-alone systems and combined in an umbrella system for result merging. Alternatively (not used by Crossley et al. (2019), but playing a prominent role in many recent readability studies), COH-METRIX[3] (Graesser et al., 2011) provides a multi-dimensional set of (psycho)linguistic and discourse features (version 3.0 incorporates 108 different indices).

Recently, Štajner et al. (2020) introduced COCO, an advanced system with cognitively plausible features, yet its focus is limited on conceptual complexity computation of texts. SENTSPACE (Tuckute et al., 2022) is a sentence-focused analysis engine rather close to the design goals of our work, which also uses a range of cognitively plausible lexical, syntactic and semantic features. However, it lacks classical stylometric and readability indices and is limited to analyses up to the single document level only. In contrast to this work, we aim at cross-document and cross-corpus analyses for more powerful register, genre and style analyses.

Despite the remarkable progress that has been made already—the proliferation of surface-level, linguistic and cognitive features under scrutiny, and the growing number of metrics making use of them—we observe a fundamental lack of integration of and abstraction from single counts and scores in these precursors. Accordingly, a major goal of our work is to provide reasonable summarization, comparison, and aggregation levels for single metrics so that divisions into registers, genres and styles can be computed on the fly based on the contributions of a wide range of linguistic layers (integrating lexical, syntactic, and semantic features) for complex collections of (multilingual) linguistic data in terms of (sets of) corpora.

[1] https://github.com/computationalstylistics/stylo

[2] The most recent version, LIWC2015, is available under http://liwc.wpengine.com/ and must be purchased for a modest fee for academic and industrial use.

[3] http://www.cohmetrix.com

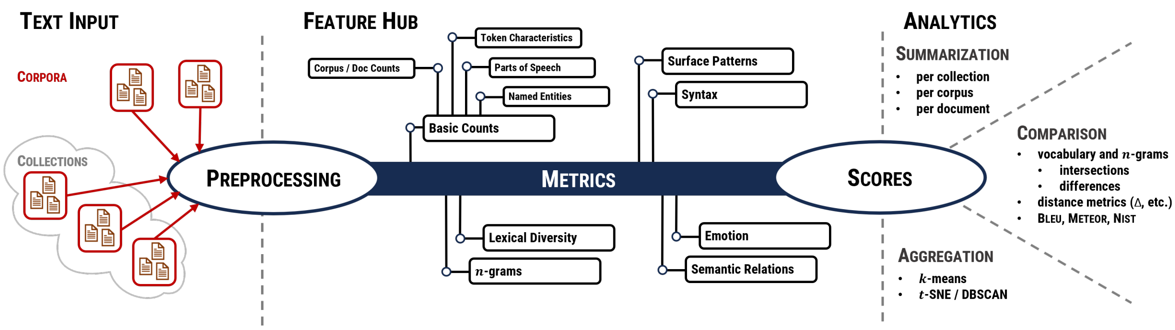

Figure 1: Overview of the building blocks of DoPA Meter

3 DoPA Meter's System Architecture

DoPA Meter is based on Python and spaCy[4] and supports all spaCy compatible language modules. Our system is publicly accessible via GitHub.[5] It is based on strict software engineering principles, such as modularity, easy resource maintenance and (re-)configuration (selection and augmentation of metrics and language resources, such as corpora and lexicons).

The three-layered architecture of DoPA Meter is depicted in Fig. 1. It consists of

- arbitrarily many *text corpora* that can also be grouped into *collections of corpora* which serve as textual input channel (including a preprocessing pipeline),
- the *feature hub* that elicits relevant features from the corpora for use by a large variety of *metrics*,
- and three *analytics* layers—apart from simple report generation (summaries of metrics-derived scores), we offer a comparison mode across documents and corpora, as well as cluster-based aggregation of results.

3.1 Input and Pre-processing

The input for DoPA Meter consists of a set of *text corpora* that can be bundled into *collections*, for convenience. Each corpus consists of single text files, the *documents*, each of which will automatically be pre-processed and split into *sentences* and *tokens*.

3.2 Feature Hub

The computation of features is divided into (1) simple *feature counts* whose results feed (2) a collection of *metrics*. We here distinguish micro statistics (at the document level) and macro statistics (at the corpus level).

The feature hub comprises sets of single features and groups them for better comprehensibility (see the discussion below and Table 3 in the Appendix). The computation of features allows for a *tailored* mode (configured by the user via choice options) or a *default* mode that takes all features into account.

3.2.1 Basic Counts

In order to get started we perform basic counts of sentences, tokens, types (vocabulary size), lemmata and characters using spaCy tooling (*Corpus/Doc Counts* in Fig. 1).[6]

In addition, *Token Characteristics*[7] comprise information about alphanumeric strings, lower/upper casing, etc. The counts of *Parts of Speech* (PoS) and *Named Entities* and their tagging are derived from spaCy's embedded language models and supply linguistically more informed feature sets.

3.2.2 n-grams

n-grams are sequential series of (configurable) n=*{1,2,3,...}* tokens or (PoS) tags. The scores calculate the ratios of n-grams for single documents and whole corpora or corpus collections.

3.2.3 Lexical Diversity

Lexical Diversity subsumes a group of 24 features borrowing from stylometric vocabulary metrics. Among others, this includes the common *type-token ratio (TTR)*, but also more sophisticated metrics such as *Guiraud's* R or *Herdan's* C. We also incorporate metrics which address the frequency spectrum of lexical items (e.g., *Sichel's* S) and ones capturing lexical distributions over the whole document (e.g., *Moving-Average TTR*). Last but not least, we also provide metrics for *lexical density* such as the ratio of function words. For surveys of metrics of lexical diversity, see Malvern et al. (2004); Evert et al. (2017).

[4] https://spacy.io
[5] https://github.com/dopameter/dopameter
[6] https://spacy.io/usage/linguistic-features
[7] https://spacy.io/api/token

3.2.4 Surface Patterns

Surface pattern metrics, also known as *Readability* scores, mainly focus on syllable counts, token and sentence length and thus target surface-level phenomena only. Among the large number of possible choices, we included into DoPA Meter 19 metrics, among them *Flesch-Kincaid*, *Dale–Chall* (for English, only), *SMOG*, *Gunning fog*, and the four *Wiener Sachtext formulas* (Bamberger and Vanacek, 1984) (for German, only). This feature class also contains a simple *Formality* score using PoS tags (Heylighen and Dewaele, 1999).

3.2.5 Syntax

Syntax-focused metrics account for the two major syntax representation formats: *dependency* and *constituency*. For dependency parsing, we exploit the transition-based dependency parser embedded in spaCy (Honnibal and Johnson, 2015), for constituency parsing we use the Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2019).

The *parse metrics* take general parse graph properties into account, such as the *average maximum depth* for each parse tree, i.e., the longest path from the root node to a leaf node, the *maximum fan-out* of each parse tree, i.e., the largest number of child nodes of a node in the entire parse tree, and the inverse *average out-degree centrality* value, i.e., the number of out-going edges, computed over all dependency graphs of all sentences of a document.

3.2.6 Semantic Relations

We here focus on lexico-semantic resources that provide a linkage between lemmas in terms of various semantic relations. Lexicons structured this way can be regarded as semantic networks. Our focus is on relations typically provided by WordNet-style specifications which feature synonymy, antonymy, taxonomy (hyponyms/hypernyms), and partonymy (parts and wholes).

Based on such knowledge-“heavy” resources we define several metrics that exploit the topological structures spanned in these semantic networks as instantiated by the lexical items we identify as lemmas of these lexicons within each sentence. Accordingly, we defined metrics which focus on *relational depth* by determining the minimal path length of each reading of each lemma within a document (i.e., the distance from the *top* node of the semantic network to the lemma) following taxonomic links (hypernymy or hyponymy links, only), sum up these individual length scores and average over the number of all the lemmas’ readings, and on *semantic richness*, i.e., for each (reading of the) lemma in a sentence, we determine all semantic relation instances (i.e., hypernyms, hyponyms, parts (is-part) and wholes (has-part), antonyms) it shares with other lemmas in the lexicon and average this number over all readings per lemma in the document. Scores and their averages are also available for each individual semantic relation only (e.g., the number of hyponyms of all instantiated lemmas).

3.2.7 Emotion

DoPA Meter supports scores for the eight fundamental emotional variables (valence, arousal, dominance, joy, anger, sadness, fear and disgust) based on dictionary look-ups incorporating the emotion lexicons from Buechel et al. (2020) in the JEmAS pipeline (Buechel and Hahn, 2016).[8]

3.3 DoPA Meter’s Analytics

3.3.1 Summarization Mode

In the summarization mode, statistical reports of the resulting scores are generated per document and corpus (collection), including common information, such as min/max values, means, quartiles, etc. This reporting mode describes fundamental quantitative characteristics in the feature hub and can already pinpoint at differences between documents and corpora that can be deeper explored by larger-scale clustering or classification algorithms.

3.3.2 Comparison Mode

The comparison mode points out differences or similarities between complete text corpora or user-defined subsets therefrom. It is based on a differential analysis of the corpus vocabulary, n-grams and the metrics targeting different levels of linguistic analysis mentioned above.

Besides the metrics already introduced, we also make use of well-known distance metrics from the field of stylometrics and authorship detection, e.g., *Burrows’* Δ (Burrows, 2002).

In addition to these stylometric computations, we incorporate scores originating from the field of machine translation, such as Bleu (Papineni et al., 2002), Meteor (Denkowski and Lavie, 2014) and NIST (Doddington, 2002).

3.3.3 Aggregation Mode

Going beyond the micro statistics at the single document and corpus level, the aggregation mode is

[8] https://github.com/JULIELab/JEmAS/releases

able to compute dependencies between different (sets of) corpora at the macro level of analysis. With varying configurations of features, *k*-means and *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) (van der Maaten and Hinton, 2008) with DBSCAN (Ester et al., 1996) are used as clustering algorithms at the moment. Our modular architecture, however, is open to extension by a wider range of additional clustering algorithms and other machine learning libraries.

4 DoPA Meter in Action

We now illustrate facets of the rich functionality of DoPA Meter. Our scenario features two languages, English and German, and a broad application domain (medicine) with six corpora (collections) from a wide range of genres (see Table 1):[9]

Corpus	Documents	Sentences	Tokens
de.Clin	3 497	145 870	1 649 156
de.PubMed	1 028	5 676	101 173
de.SocMed	4 000	30 943	433 999
de.Wiki	4 400	326 721	4 348 255
en.Clin	5 918	437 598	7 065 887
en.SocMed	3 601	13 168	172 927

Table 1: Quantitative data of the demo corpus collection

de.Clin is composed of several publicly available German clinical corpora: JSynCC (Lohr et al., 2018), Assess (Miñarro Giménez et al., 2019), Bronco (Kittner et al., 2021), GraSCCo (Modersohn et al., 2022), Ex4CDS (Roller et al., 2022), Cardio:De (Richter-Pechanski et al., 2023) and a set of X-ray reports (Dewald et al., 2023),
de.PubMed contains the German subset of PubMed abstracts featuring clinical cases,[10]
de.SocMed contains medical layman and expert expressions from a patient forum (Seiffe et al., 2020),
de.Wiki collects medical articles from Wikipedia including info-box data with an ICD-10 code,[11]
en.Clin incorporates public corpora supplied for the i2b2 and n2c2 challenge series,[12] and
en.SocMed combines English language Twitter corpora with biomedical content: Bear (Wührl and Klinger, 2022), CoVERT (Mohr et al., 2022), and BioClaim (Wührl and Klinger, 2021).

[9] Instructions how to build the corpora in order to reproduce our experiments can be found under `https://doi.org/10.5281/zenodo.10000771`

[10] `https://pubmed.ncbi.nlm.nih.gov/`, running the query `"Case Reports[Publication Type] AND GER[LA]"`

[11] `https://www.wikipedia.de/`

[12] `https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/`

Summarization Mode:

The boxplots from Figure 2 depict the results from surface-level formality scoring (based on Heylighen and Dewaele (1999)) in a visual way. Clinical documents, for both languages, are in the high end of formal language use, whereas social media language, not surprisingly, scores at the lower end, with news, Wikipedia, and PubMed in between.

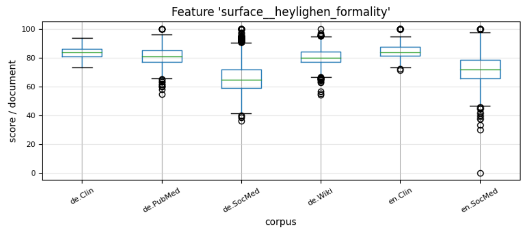

Figure 2: Surface Heylighen formality scores

Table 2 contains scores that illustrate corpus-based metrics from *Surface Patterns* (*Flesch Reading Ease* index), *Syntax* (depth of dependency parse trees (Dep-Depth)), and WordNet-based *Semantic Relations* (semantic richness of synonyms).

	Surface	**Syntax**	**Semantics**
Corpus	**Flesch**	**Dep-Depth**	**Synonym-Rich**
de.Clin	69.97	4.28	2.05
de.PubMed	59.91	4.75	3.45
de.SocMed	35.88	6.34	4.09
de.Wiki	87.68	4.74	3.10
en.Clin	85.59	4.98	0.80
en.SocMed	85.07	4.14	0.81

Table 2: Scores for *Flesch Reading Ease (Flesch)*, average maximum depth of dependency trees (*Dep-Depth*), and semantic richness of synonyms from WordNet (Synonym-Rich) (maxima in red, minima in blue)

Surprisingly, German Wikipedia texts are the hardest to understand, in a similar readability range with English clinical documents and social media chats. The German expert-layman data is by far the easiest to read. German clinical documents exhibit a higher readability than English ones.

The highest syntactic complexity in terms of parse tree depth is attributed to the German expert-layman corpus (expert statements seem to suffer from 'hard' syntax), with no substantial differences for the remaining corpora.

The German social media corpus (in contrast to the English one) is the richest in terms of synonyms, whereas both clinical corpora are semantically poor at that level (adhering to canonical medical terminology—the English one being even poorer than the German one). The medical German Wikipedia is in a similar range with German clinical and PubMed documents on that dimension.

Comparison Mode:
To highlight the lexical intersection among corpora, the heatmap in Fig. 3 is provided for 1-grams. The language division is obvious, yet the status of the German (medical) WIKIPEDIA is interesting insofar as it has a rather strong overlap with with German PUBMED and expert-layman social media data. Furthermore, German clinical reports share a remarkable portion of vocabulary with German PUBMED and, to a lesser degree though, with expert-layman interaction in social media.

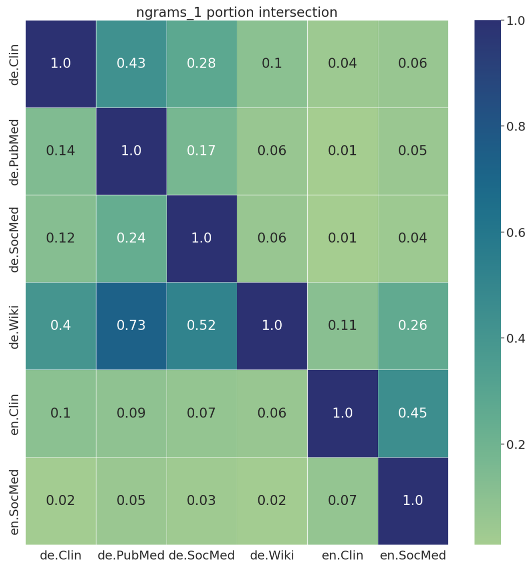

Figure 3: Vocabulary Intersection

Aggregation Mode:
Figure 4 depicts the distribution of the scores for formal token attributes, e.g., whether a token is alphanumeric or a punctuation mark, using T-SNE (van der Maaten and Hinton, 2008), thus mapping high-dimensional data onto two dimensions.

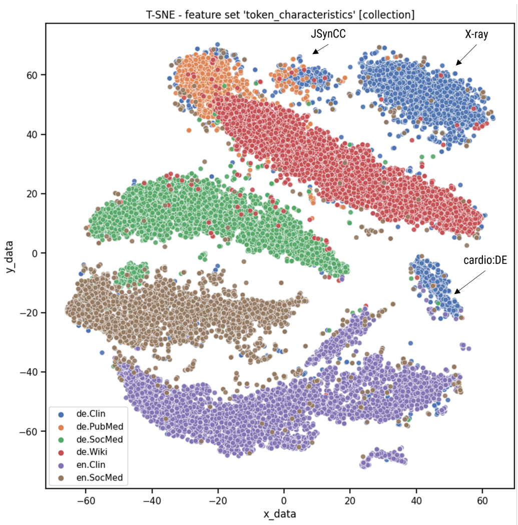

Figure 4: Clustering by token characteristics (1)

Again, the division between languages is obvious. There are clear differences between German (upper part of Fig. 4) and English language (lower part). Social media corpora (de.SocMed and en.SocMed) of both languages lie close to each other (green and brown area) as are the samples from PUBMED and WIKIPEDIA (orange and green parts). Yet, the samples of German clinical language are divided into three distinct clusters (blue dots, with labels for the three largest corpora; for more details, see Section D in the Appendix), parts of which are close to WIKIPEDIA and PUBMED, or even overlap with those from the English language.

All these observations indicate that none of the features in isolation is capable of properly predicting specific discourse categories, such as registers or text genres. Hence, a deeper exploration of dependencies between the features we measure seems more appropriate and DOPA METER might be a suitable toolkit for this endeavor.

5 Conclusions

We introduced DOPA METER, a toolkit for quantifying feature distributions at the lexical, syntactic and semantic dimension. We supply 120 metrics for scoring linguistic behavior at these axes. Scores can be summarized, compared, and aggregated using flexibly tailorable visualization tools.

DOPA METER's feature collection reflects one main design goal of our work, namely the integration of as many linguistic levels as possible, thus moving away from much more selective approaches in stylometrics and psychometrics. A second unique feature of our approach is its focus on lucid system architecture for flexible system engineering, i.e., easy maintainability and augmentation by new metrics and language resources (corpora, lexicons) in a coherent *all-in-one* system design. This contrasts with the proliferation of stylometric extensions spread over lots of local GITHUB links lacking further integration, on the one hand, and frozen system packages in the psychometric domain, on the other hand. The source code and its documentation are provided under the open MIT licence and our tool can be conveniently expanded and adapted to specific needs.

This way, DOPA METER may be useful as a metadata generator for documents and text corpora, with facilities for quantitative data description (scoring), comparison and aggregation. Such an approach may also pave the way towards an empirically sound way of routinely running NLP *data diagnostics* (Xiao et al., 2022).

References

Richard Bamberger and Erich Vanacek. 1984. *Lesen – Verstehen – Lernen – Schreiben*. Diesterweg.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including PAIS 2016 — Prestigious Applications of Artificial Intelligence.*, volume 2: Long Papers, pages 1114–1122, The Hague, The Netherlands, August 29 - September 2, 2016. IOS Press.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*, pages 1202–1217, [Seattle, Washington, USA,] July 5-10, 2020 (Virtual Event). Association for Computational Linguistics (ACL).

John Burrows. 2002. 'Delta' : a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: a survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2017. Sentiment analysis and social cognition engine (SEANCE) : an automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 49(3):803–821.

Scott A. Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Michael Denkowski and Alon Lavie. 2014. METEOR UNIVERSAL : language specific translation evaluation for any target language. In *WMT 2014 — Proceedings of the 9th Workshop on Statistical Machine Translation @ ACL 2014.*, pages 376–380, Baltimore, Maryland, USA, June 26-27, 2014. Association for Computational Linguistics (ACL).

Cornelia L. A. Dewald, Alina Balandis, Lena S. Becker, Jan B. Hinrichs, Christian von Falck, Frank K. Wacker, Hans Laser, Svetlana Gerbel, Hinrich B. Winther, and Johanna Apfel-Starke. 2023. Automated classification of free-text radiology reports: using different feature extraction methods to identify fractures of the distal fibula. *RöFo — Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 195(8):713–719.

George R. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *HLT 2002 — Human Language Technology Conference. Proceedings of the 2nd International Conference on Human Language Technology Research.*, pages 138–145, San Diego, California, USA, March 24-27, 2002. Morgan Kaufmann Publishers Inc.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R : a package for computational text analysis. *R Journal*, 16(1):107–121.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD '96 — Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining.*, pages 226–231, Portland, Oregon, USA, August 2-4, 1996. AAAI Press.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-VIEW-ALATION : a large-scale evaluation study of association measures for collocation identification. In *Electronic Lexicography in the 21st Century. eLex 2017 — Proceedings of the 5th Conference on Electronic Lexicography.*, pages 531–549, Leiden, Netherlands, 19-21 September 2017. Lexical Computing CZ s.r.o.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. COH-METRIX : providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Center "Leo Apostel", Free University of Brussels.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.*, pages 1373–1378, Lisbon, Portugal, 17-21 September 2015. Association for Computational Linguistics (ACL).

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.*, pages 3499–3505, Florence, Italy, July 28 - August 2, 2019.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.*, volume 1:

Long Papers, pages 2676–2686, Melbourne, Victoria, Australia, July 15-20, 2018. Association for Computational Linguistics (ACL).

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated German oncological corpus. *JAMIA Open*, 4(2):ooab025.

Kristopher Kyle and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Kristopher Kyle and Scott A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

David Y. W. Lee. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation.*, pages 1259–1266, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.

Jose A. Miñarro Giménez, Ronald Cornet, Marie Christine Jaulent, Heike Dewenter, Sylvia Thun, Kirstine Rosenbeck G, Daniel Karlsson, and Stefan Schulz. 2019. Quantitative analysis of manual annotation of clinical text samples. *International Journal of Medical Informatics*, 123:37–48.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO : the first publicly shareable, multiply-alienated German clinical text corpus. In *German Medical Data Sciences 2022 — Future Medicine: More Precise, More Integrative, More Sustainable! Proceedings of the Joint Conference of the 67th Annual Meeting of the GMDS & 14th Annual Meeting of the TMF*, number 296 in Studies in Health Technology and Informatics, pages 66–72, [Kiel, Germany,] 21-25 August 2022 (Virtual Event). IOS Press.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. COVERT : a corpus of fact-checked biomedical COVID-19 tweets. In *LREC 2022 — Proceedings of the 13th International Conference on Language Resources and Evaluation.*, pages 244-257, Marseille, France, June 20-25, 2022. European Language Resources Association (ELRA).

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):#86.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU : a method for automatic evaluation of machine translation. In *ACL '02 — Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*, pages 311–318, Philadelphia, Pennsylvania, USA, July 6-12, 2002. Association for Computational Linguistics (ACL).

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Mingyang He, Michael M. Allers, Anna S. Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, Julian Mierisch, Florian Borchert, Charlotte Schwind, Norbert Frey, Christoph Dieterich, and Nicolas A. Geis. 2023. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10:#207.

Roland Roller, Aljoscha Burchardt, Nils Feldhus, Laura Seiffe, Klemens Budde, Simon Ronicke, and Bilgin Osmanodja. 2022. An annotated corpus of textual explanations for clinical decision support. In *LREC 2022 — Proceedings of the 13th International Conference on Language Resources and Evaluation.*, pages 2317–2326, Marseille, France, June 20-25, 2022. European Language Resources Association (ELRA).

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones: resources for medical laymen to technical language and vice versa. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation.*, pages 6185–6192, Marseille, France, May 11-16, 2020. European Language Resources Association (ELRA).

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Greta Tuckute, Aalok Sathe, Mingye Wang, Harley Yoder, Cory Shain, and Evelina Fedorenko. 2022. SENTSPACE : large-scale benchmarking and evaluation of text using cognitively motivated lexical, syntactic, and semantic features. In *NAACL-HLT 2022 — Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations Session.*, pages 99–113, Seattle, Washington, USA, July 10-15, 2022 (and Virtual Event). Association for Computational Linguistics (ACL).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using T-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32(4):429–435.

Sanja Štajner, Sergiu Nisioi, and Ioana Hulpuş. 2020. CoCo : a tool for automatically assessing conceptual complexity of texts. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 7179–7186, Marseille, France, May 11-16, 2020. European Language Resources Association (ELRA).

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical TWITTER posts. In *BioNLP 2021 — Proceedings of the 20th Workshop on Biomedical [Natural] Language Processing @ NAACL-HLT 2021.*, pages 131–142, June 11, 2021 (Virtual Event). Association for Computational Linguistics (ACL).

Amelie Wührl and Roman Klinger. 2022. Recovering patient journeys: a corpus of biomedical entities and relations on TWITTER (BEAR). In *LREC 2022 — Proceedings of the 13th International Conference on Language Resources and Evaluation.*, pages 4439–4450, Marseille, France, June 20-25, 2022. European Language Resources Association (ELRA).

Yang Xiao, Jinlan Fu, Weizhe Yuan, Vijay Viswanathan, Zhoumianze Liu, Yixin Liu, Graham Neubig, and Pengfei Liu. 2022. DATALAB : a platform for data analysis and intervention. In *ACL 2022 — Association for Computational Linguistics: System Demonstrations.*, pages 182–195, Dublin, Ireland, May 22-27, 2022 (and Virtual Event). Association for Computational Linguistics (ACL).

A Ethical Considerations

DOPA METER uses a wide range of external resources, such as corpora, lexicons or terminology systems with potentially built-in biases. Users of DOPA METER should be sensitive towards potential pitfalls when analyzing data and reporting the results gathered with DOPA METER.

B Limitations

DOPA METER combines metrics, e.g., for readability or syntactic complexity, which are commonly used but often lack comparative evaluation. Hidden, and potentially unrecognized or unwarranted, dependencies between them should be carefully considered.

Despite our efforts to include at least two languages (English and German), the multilingual dimension needs further elaboration. When doing so one might encounter shortcomings or even gaps for particular languages (e.g., for readability formulae, corpora, terminologies or lexicons).

Finally, DOPA METER's aggregation component needs further extension by complementary clustering and ML classification algorithms.

C Grants

This work was supported by BMBF within the SMITH project under grants 01ZZ1803G and 01ZZ1803A such as the GeMTeX project under grant 01ZZ2314D.

D Fine-Grained Clustering of All Individual Corpora

The following figure provides a more detailed view of the data aggregated in Fig. 4.

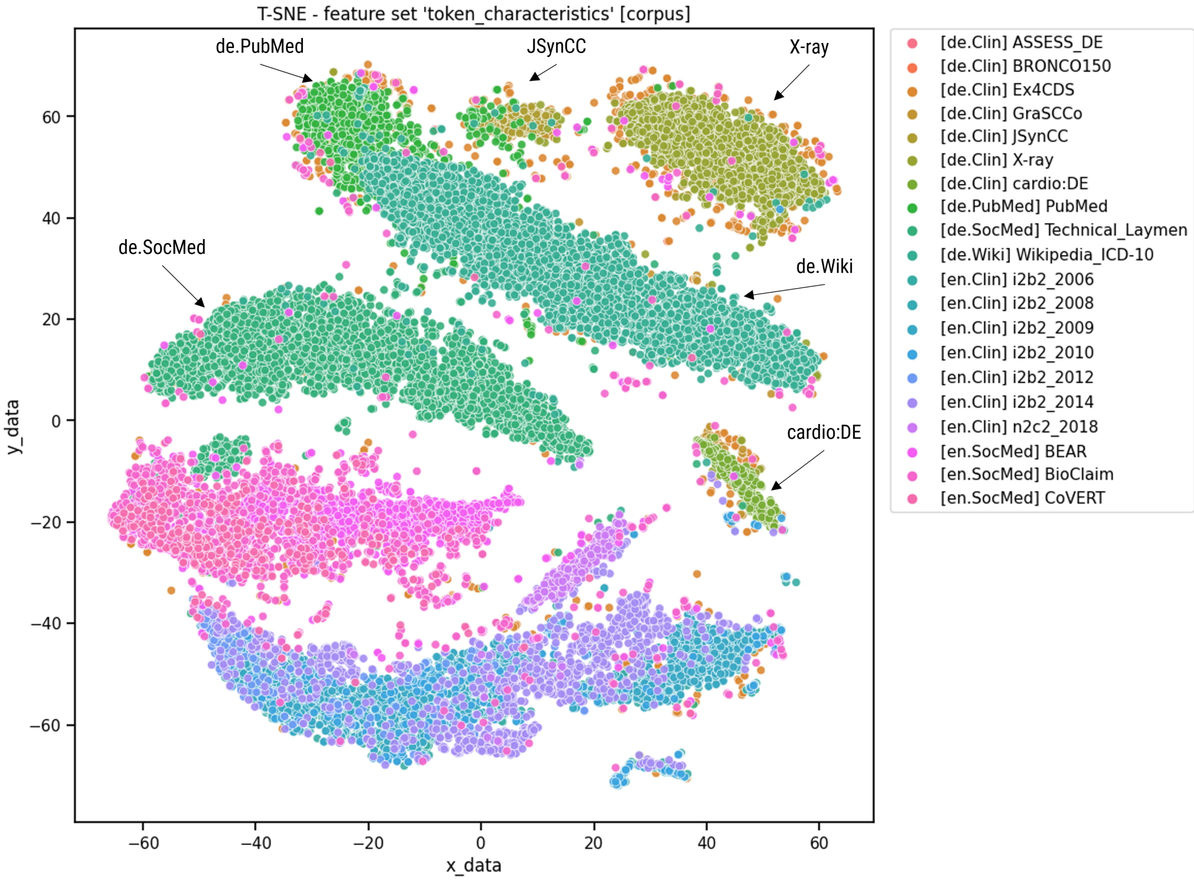

Figure 5: Clustering by token characteristics (2): Finer-grained visualization of Fig. 4

E Feature Hub Summary

Feature Hub	Metrics	Amount of Metrics			Modus / Analysis		
		German	*English*	*Mult.*	*Count*	*Metrics*	*Compare*
Corpus / Doc Counts	*characters*, *sentences*, *different_sentences*, *tokens*, *types*, *lemmata*	6	6	6	✓		
Token Charcteristics	*is_alpha*, *is_ascii*, *is_digit*, *is_lower*, *is_upper*, *is_title*, *is_punct*, *is_left_punct*, *is_right_punct*, *is_space*, *is_bracket*, *is_quote*, *is_currency*, *like_url*, *like_num*, *like_email*, *is_stop*	17	17	17	✓	✓	
Part of Speech	depends on spaCy language model German (`de_core_news_sm`): TIGER tagset (e.g., DET, NOUN, VERB, ADP, ...) English (`en_core_web_sm`): Onto Notes 5 (e.g., AUX, NOUN, VERB, PROPN, ...)	1	1	1	✓	✓	
Named Entities	depends on spaCy language model German (`de_core_news_sm`): WikiNER (only LOC, PERS, MISC, ORG) English (`en_core_web_sm`): WordNet 3.0 (e.g., DATE, LOC, PERSON, ORG)	1	1	1	✓	✓	
n-grams (tfidf)	depends on configuration of n and most frequent words, preferred: n={1,2,3}	1	1	1	✓	✓	✓
Lexical Diversity	*type_token_ratio*, *lexical_density*, *guiraud_r*, *herdan_c*, *dugast_k*, *maas_a2*, *dugast_u*, *tuldava_ln*, *brunet_w*, *cttr*, *summer_s*, *sttr*, *sichel_s*, *michea_m*, *honore_h*, *entropy*, *yule_k*, *simpson_d*, *herdan_vm*, *hdd*, *evenness*, *mattr*, *mtld*	23	23	23	✓	✓	✓
Surface Patterns	*avg_token_len_chars*, *avg_sent_len_tokens*, *avg_sent_len_chars*, *flesch_kincaid_grade_level*, *smog*, *coleman_liau*, *ari*, *forcast*, *gunning_fog*, *heylighen_formality* no default: *toks_min_three_syllables*, *toks_larger_six_letters*, *toks_one_syllable*, *syllables letter_tokens no_digit_tokens* only German: *flesch_reading_ease*, *wiener_sachtextformel_1*, *wiener_sachtextformel_2*, *wiener_sachtextformel_3*, *wiener_sachtextformel_4* only English: *flesch_reading_ease*, *dale_chall*	23	20	18	✓	✓	✓
Syntax - Dependency	AvgFan, MaxFan, AvgMaxDepth, AvgDepDist, MaxDepDist, AvgOutdegreeCentralization, AvgClosenessCentralization, occurrences of tree nodes (depending on spaCy language model)	8	8	8	✓	✓	
Syntax - Constituency	AvgMaxDepth, AvgFan, MaxFan, AvgNonTerminales_sent, AvgConstituents_sent, AvgTunits_sent, AvgLenConstituents, AvgLenTunits, AvgOutdegreeCentralization, MaxOutdegreeCentralization, AvgClosenessCentralization, MaxClosenessCentralization occurrences of tree nodes	13	13	13	✓	✓	
Emotion	valence, arousal, dominance, joy, anger, sadness, fear, disgust	8	8	8		✓	
Semantic Relations	*sem_rich_hypernyms*, *sem_rich_hyponyms*, *sem_rich_taxonyms*, *sem_rich_antonyms*, *sem_rich_synonyms*, *sem_rich_meronyms*, *sem_rich_holonyms*, *sem_rich*, *min_depths_avg*, *min_depths_min*, *min_depths_max*, *max_depths_avg*, *max_depths_min*, *max_depths_max*, *synsets_avg*, *senses_avg*, occurrences of synsets, occurrences of senses	18	18	18	✓	✓	✓
Amount of all Metrics		119	116	114			

Table 3: Summary of all *Feature Hubs* and all *Metrics* of DOPA METER