Syntactic Language Change in English and German: Metrics, Parsers, and Convergences

Yanran Chen
Natural Language Learning Group (NLLG),
University of Mannheim, Germany
yanran.chen@uni-mannheim.de

Wei Zhao
University of Aberdeen, UK
wei.zhao@abdn.ac.uk

Anne Breitbarth
Ghent University, Belgium
anne.breitbarth@ugent.be

Manuel Stoeckel
Text Technology Lab (TTLab),
Goethe University Frankfurt, Germany
manuel.stoeckel@em.uni-frankfurt.de

Alexander Mehler
Text Technology Lab (TTLab),
Goethe University Frankfurt, Germany
mehler@em.uni-frankfurt.de

Steffen Eger
Natural Language Learning Group (NLLG),
University of Mannheim, Germany
steffen.eger@uni-mannheim.de

Many studies have shown that human languages tend to optimize for lower complexity and increased communication efficiency. Syntactic dependency distance, which measures the linear distance between dependent words, is often considered a key indicator of language processing difficulty and working memory load. The current paper looks at diachronic trends in syntactic language change in both English and German, using corpora of parliamentary debates from the last c. 160 years. We base our observations on five dependency parsers, including the widely used Stanford CoreNLP as well as 4 newer alternatives. Our analysis of syntactic language change goes beyond linear dependency distance and explores 15 metrics relevant to dependency distance minimization (DDM) and/or based on tree graph properties, such as the tree height and degree variance. Even though we have evidence that recent parsers trained on modern treebanks are not heavily affected by data 'noise' such as spelling changes and OCR errors in our historic data, we find that results of syntactic language change are sensitive to the parsers involved, which is a caution against using a single parser for evaluating syntactic language change as done in previous work. We also show that syntactic language change over the time period investigated is largely similar between English and German for the different metrics explored: only 4% of cases we examine yield opposite conclusions regarding upwards and downtrends of syntactic metrics across German and English. We also show that changes in syntactic measures seem to be more frequent at the tails of sentence length distributions. To our best knowledge, ours is the most comprehensive analysis of syntactic language change using modern NLP technology in recent corpora of English and German.

1. Introduction

Many studies have shown that human languages are being optimized in the direction of lower complexity and higher efficiency in terms of communication (Gibson et al. 2019; Ferrer-i Cancho et al. 2022; Degaetano-Ortlieb and Teich 2022). Among others, syntactic dependency distance, i.e., the linear distance between syntactically dependent words, is often perceived as an indicator of language processing difficulty and working memory load in humans (Temperley 2007; Liu 2008; Zhu, Liu, and Pang 2022; Juzek, Krielke, and Teich 2020). The corresponding optimization

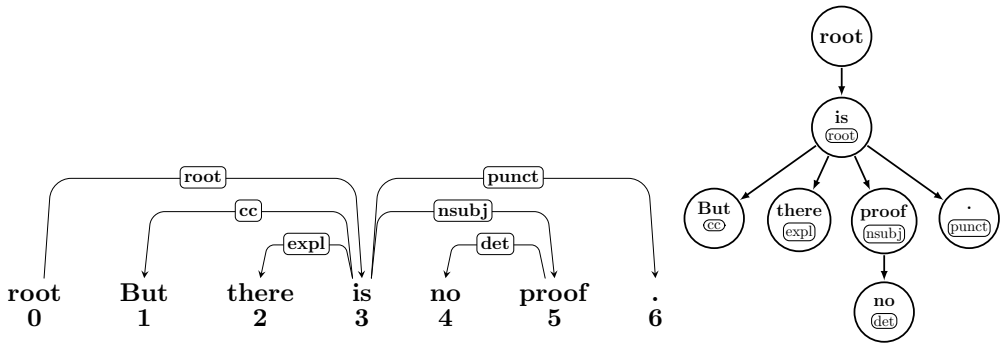


Figure 1: Dependency relations of sentence "But there is no proof." in linear order (left) and tree graph (right).

principle is well-established, i.e., dependency distance minimization (DDM), which refers to the tendency/preference of syntactically related words being placed closer to each other.

A large array of works have confirmed the existence of DDM in natural languages (e.g., Temperley (2007); Futrell, Mahowald, and Gibson (2015); Juzek, Krielke, and Teich (2020); Zhang, Li, and Zhang (2020)). On the one hand, studies suggest that the mean dependency distance (or its variants) of a real sentence is shorter than that of random baselines (e.g., sentences with random word order) and this may be a universal property of most natural languages (Gildea and Temperley 2010; Futrell, Mahowald, and Gibson 2015; Ferrer-i Cancho et al. 2022). On the other hand, there is also evidence of DDM in diachronic investigations of language change i.e., the mean dependency distance is decreasing over time (Lei and Wen 2020; Liu, Zhu, and Lei 2022; Juzek, Krielke, and Teich 2020; Zhang and Zhou 2023; Zhu, Liu, and Pang 2022; Krielke 2023).

The recent diachronic investigations involve using dependency parsers to automatically parse sentences extracted from a corpus covering a large time span, and observing how dependency distance in the parsed outputs varies over time (e.g., Lei and Wen (2020); Juzek, Krielke, and Teich (2020); Zhang and Zhou (2023)). The main advantage of such studies over approaches which leverage human annotated treebanks (Gulordava and Merlo 2015) is that they do not require expensive manual dependency annotations and thus allow for exploring arbitrary large diachronic corpora. However, to our best knowledge, all of them rely on a single dependency parser, mostly the the Stanford CoreNLP parser (Manning et al. 2014), which was published 9 years ago. In this work, we first inspect the parsers' legitimacy in our historical use case, as they are mostly only trained and evaluated on modern treebanks, which cover a small recent time period and may exhibit substantially less data noise (e.g., OCR errors) compared to historical corpora (§4). Subsequently, we investigate *whether different parsers yield the same trends regarding language change (§6.1)*.

Dependency relations can also be structured as a rooted directed tree, as shown in Figure 1 (right). Previous works only focus on liner dependency distance; instead, we also explore metrics beyond linear dependency distance, which are relevant to DDM and/or based on tree graph properties, such as the tree height and degree variance (§5).

Further, the majority of relevant studies are for English, with the exception that Krielke (2023) investigates DDM (a.o.) in both English and German (from 1650 to 1900). She compares the trends in scientific and general language, where the texts for general language span multiple domains, such as news and fiction. In our work, we focus on a homogeneous and directly comparable genre of political debates in both English and German, stretching from the 1800s to beyond the 2000s.

Our main research questions are:

- RQ1 (parsers): Are parsers trained on modern treebanks reliable to parse (our) historical data, which is affected by OCR and spelling error changes — especially the German data?
- RQ2 (parsers): When predicting trends of syntactic language change, can one rely on the predictions of a single parser?
- RQ3 (languages): Do English and German mostly change similarly regarding syntax (convergence) or do they have divergent patterns of syntactic language change?
- RQ4 (metrics): How do English and German change when looking at syntactic dependency graph properties beyond mean dependency distance?

We will make our code+data public upon acceptance.[1]

2. Related Work

The current paper connects to language change with a focus on syntactic change based on DDM.

Language Change. Language change has been researched for a long time along multiple dimensions (and in multiple communities), including semantic, syntactic, morphological change etc. **Semantic change** investigates the changes in word meaning over time. For example, Giulianelli, Del Tredici, and Fernández (2020) document semantic change in the words "boy" and "girl": Girl used to be a term for young people of either sex, while boy described male servants; the present meaning of "girl" and "boy" were first observed in the 15th and 16th centuries, respectively. Earlier works on studying semantic change relied on methods like latent semantic analysis (Sagi, Kaufmann, and Clark 2011) or treated semantic change detection as a diachronic word sense analysis problem (Mitra et al. 2014; Frermann and Lapata 2016). More recent studies have leveraged static or contextualized embeddings (Bamler and Mandt 2017; Bamman, Dyer, and Smith 2014; Chen et al. 2023; Tahmasebi, Borin, and Jatowt 2021; Rother, Haider, and Eger 2020; Ma, Strube, and Zhao 2024), jointly trained across varying periods (Bamler and Mandt 2017; Bamman, Dyer, and Smith 2014; Di Carlo, Bianchi, and Palmonari 2019; Dubossarsky et al. 2019; Haider and Eger 2019), or independently trained for different periods (Hamilton, Leskovec, and Jurafsky 2016). The latter scenario then involves remapping embeddings for different periods into one shared vector space (Hamilton, Leskovec, and Jurafsky 2016) or inducing the second-order embeddings, which represents the words' distance to a reference vocabulary (Eger and Mehler 2016; Rodda, Senaldi, and Lenci 2017). **Morphological change** focuses on how word formation and inflection change over time (Anderson 2015). For example, it has been shown that in most Germanic languages, including English, case morphology has gradually declined over time (Weerman and Wit 1999). Moscoso del Prado Martín and Brendel (2016) explore case changes in Icelandic, using historical corpora annotated with part-of-speech and other morphological information. The rules of verbal inflection have been explored for many languages. For example, more frequently used English verbs have undergone faster regularization (Lieberman et al. 2007);[2] verb inflection tends to become weaker in German (Carroll, Svare, and Salmons 2012) and Dutch (Knooihuizen and Strik 2014) over time. Zhu and Lei (2018) further confirm inflection decay, i.e., languages become less inflectional, in Modern English, using entropy-based

1 https://github.com/cyr19/syntaxchange

2 Note that on the contrary, highly frequent tokens in many cases resist regularization. This is known as the conservation effect (Bybee 1985). In German, for instance, lower-frequency strong verbs are historically more likely to adopt the weak inflection, which has a higher type frequency. Opposed to that, highly frequent verbs tend to preserve, and in some cases even extend, irregular inflection, as discussed e.g. in (Nübling 1998).

algorithms. **Syntactic change** has been discussed for changes in specific syntactic patterns and/or metrics reflecting language complexity. For instance, Krielke (2021); Krielke et al. (2022); Krielke (2023); Degaetano-Ortlieb and Teich (2022) investigate syntactic shifts in scientific English and/or German, observing several phenomena, such as the tendency towards heavy noun phrases and a simpler sentence structure with decreasing usage of clauses over time. Similar phenomena were also observed in Halliday (2003); Banks (2008); Biber and Gray (2011, 2016). One of the widely adopted approaches to model language complexity revolves around syntactic dependency distance. Studies have indicated that humans tend to position syntactically related words closer together (Tily 2010; Gulordava and Merlo 2015; Liu, Xu, and Liang 2017; Zhang and Zhou 2023); this hypothesis/phenomenon is called DDM.

Dependency Distance Minimization. The most relevant approaches to ours are those based on sentential dependency structures, where (a.o.) the dependency distance and the corresponding principle (i.e., DDM) have been researched extensively. From a **synchronic** perspective, many studies have shown that the mean dependency distance (*mDD*) (or variants) of a real sentence is shorter than that of some random baselines: for example, Ferrer-i Cancho (2004) and Liu (2007) observed that the mean dependency distance of Romanian/Czech and Chinese sentences is significantly shorter than that of random sentences, respectively. Liu (2008); Futrell, Mahowald, and Gibson (2015) provide large-scale evidence of DDM, which suggests that DDM may be a universal property of human languages. They use treebanks of different languages from various linguistic families to perform the tests and show that the dependency distance of real sentences is shorter than chance. More recently, Ferrer-i Cancho et al. (2022) analyze the optimality (based on dependency distance) of 93 languages from 19 linguistic families. Their results imply that 50% of human languages have been optimized to a level of at least 70%. Compared to English, German has been less optimized. This finding is in line with Gildea and Temperley (2010), who observe that DDM has a much weaker impact on German than on English.

On the other hand, **diachronic** investigation of dependency distance has received an increasing interest in the past few years (e.g., Zhang and Zhou (2023); Zhu, Liu, and Pang (2022); Liu, Zhu, and Lei (2022)). Tily (2010) may be the first relevant work. He finds that the total dependency distance of English sentences decreased from 900 to 1500 (from Old English to Early Modern English). As the corpora only contain constituency annotations, he automatically converts them into dependency annotations for analysis. Gulordava and Merlo (2015) compare Latin / ancient Greek from different time periods using manually annotated dependency treebanks, which contain around 1k-10k sentences for each period, finding that languages from the earlier periods exhibit lower DDM levels against the optimum than those from the later periods.

More recent studies leverage dependency parsers to automatically parse sentences extracted from diachronic corpora, and then base their observations on the parsing results (Lei and Wen 2020; Juzek, Krielke, and Teich 2020; Liu, Zhu, and Lei 2022; Zhu, Liu, and Pang 2022; Zhang and Zhou 2023; Krielke 2023). An important benefit of such approaches is that they do not require expensive dependency annotations; therefore, arbitrary corpora beyond treebanks can serve as the research resource. Lei and Wen (2020); Liu, Zhu, and Lei (2022) investigate the diachronic trend of dependency distance based on the State of the Union Addresses corpus,[3] which contains the political speeches of 43 U.S. presidents from 1790 to 2017. They show that the *mDD* and its normalized variant (*nDD*) (Lei and Jockers 2020) of American English sentences are decreasing over time. While Lei and Wen (2020) observe a consistent downward trend across different sentence length groups, Liu, Zhu, and Lei (2022) additionally divide the sentences of $\leq$ 10 words (the shortest sentence group in Lei and Wen (2020)) into two subgroups: sentences of 0-4 and 5-10 words, finding the existence of an anti-DDM phenomenon for the 0-4 group, i.e.,

3 https://www.presidency.ucsb.edu/

DDM does not apply to short sentences because of the constraints by other language optimization principles (Ferrer-i Cancho and Gómez-Rodríguez 2021). Zhu, Liu, and Pang (2022); Zhang and Zhou (2023) compare the diachronic trend of *mDD* and/or *nDD* for the past 100-200 years across varying text genres, based on the COHA corpus,[4] which contains texts in American English across 4 domains: News, Magazine, Fiction, Non-fiction. Their findings are often inconsistent. E.g., Zhu, Liu, and Pang (2022) find a downtrend of *mDD* only for the Fiction genre but an upward trend for the News and Magazine genres during the period 1900-2009. In contrast, Zhang and Zhou (2023) observe a continuous decrease in *mDD* and *nDD* for the Magazine genre spanning from 1815 to 2009. Nevertheless, both of their results imply that the diachronic trends of *mDD* vary for different genres. Juzek, Krielke, and Teich (2020) explore diachronic shifts in syntactic patterns for scientific English, leveraging the Royal Society Corpus (Fischer et al. 2020), which contains scientific publications from 1665 to 1996. They also observe a decrease in *mDD* over time. As follow-up work, Krielke (2023) comprehensively investigates language optimization for scientific English and German from 1650 to 1900, in comparison to that for general language. She finds that while the *mDD* for scientific language decreased over time, *mDD* for general-domain language does not show a decrease over time during the same period.

However, all those approaches have three main limitations: (1) all of them rely on one single parser, mostly the Stanford CoreNLP parser (Manning et al. 2014) with an exception that Krielke (2023) leverages the UDPipe parser (Straka 2018). In our work, we explore whether different parsers yield the same outcomes regarding our diachronic investigations. (2) Some of the works may report biased results due to improper sentence grouping. We show a clear trend of shorter sentences appearing more frequently in later periods, consequently, the average sentence length per decade is decreasing over time. Zhang and Zhou (2023) do not differentiate sentences by lengths and report the diachronic trends on all sentences. Liu, Zhu, and Lei (2022); Lei and Wen (2020) divide the sentences into groups having 1-10 (0-4, 5-10), 11-20, 21-30, and 31+ words, and analyze trends for each group separately. However, we demonstrate that even when restricting the maximum difference in sentence lengths to 3 tokens within a group, a slight downtrend in sentence length over time persists (refer to §7). Given the substantial length disparity within a group in their works (e.g., 10 for group 11-20), and lacking information on the length distribution in their data, it is hard to tell whether the observed downtrend of *mDD* is the effect of decreasing sentence length or really due to DDM. Those facts may cast doubts on the legitimacy of their findings. (3) Almost all of the works focus on English change, except for Krielke (2023), who compares the changes in scientific and general English and German. The dataset for general languages used in Krielke (2023) contain texts in multiple genres like fiction and news. Nevertheless, as Zhu, Liu, and Pang (2022); Zhang and Zhou (2023) suggest, trends of *mDD* may differ for various text genres. Instead, in this work, we focus on one specific genre, namely *political debates*, across both languages.

3. Data Preparation

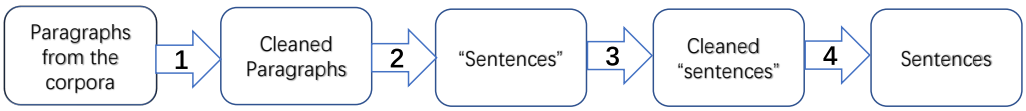


Figure 2: 4-step pipeline for sentence extraction from the corpora: 1. *paragraph-level preprocessing*, 2. *sentence segmentation with Spacy*, 3. *postprocessing*, 4. *filtering*.

4 https://www.english-corpora.org/coha/

Corpora. In this work, we use corpora consisting of political debates and speeches in German and English to ensure a fair comparison. For **English**, we select the **`Hansard`** corpus, which comprises the official reports of UK parliament debates since 1803. To compile the data for the time period 1803—2004, we extract XML files from the `Hansard` archive,[5] retaining every section labeled as "debate". Data from 2005 to April 2021 is obtained from a Zenodo repository.[6] For **German**, we utilize the **`DeuParl`** corpus published by Beese, Pütz, and Eger (2022);[7] it contains plenary protocols from both the Reichstag (the former German parliament, until 1945) and the Bundestag (the current German parliament), covering the period from 1867 to 2022.

Hansard		DeuParl	
"The Earl of Liverpool [...] [238 stance, when he [...]	Blankets 18,800 13,500 6,360	Ich will Ihnen z. B. sagen, ist [...]	(Bravo! Rechts.) Reichstag des [...] 20. Sitzung am 27. März 1867.

Table 1: Examples of problematic data. Line breaks in the table signify actual breaks in the data, with problematic texts highlighted in red.

Preprocessing. To extract sentences from the corpora, we employ a 4-step preprocessing approach, outlined in Figure 2. Initially, we conduct preprocessing at the paragraph level, with the aim of providing cleaner inputs for the sentence tokenizer. For example, as Table 1 (first column) shows, in the original data of `Hansard`, the page numbers are often not excluded from the sentence due to OCR errors, which occupy an individual line ("[238"). In `DeuParl`, the text is often split into lines by periods, no matter whether they serve as the ending punctuation. As shown in Table 1, "z.B." (English: "e.g.") is accidentally split into two lines (third column); a similar case exists for "27. März" (English: "27. March") in the last column. With paragraph-level preprocessing, we delete the line break and the page numbers (for example) before applying the sentence segmentation. In the case of `DeuParl`, which is in plain text format, we first divide each file into paragraphs of 50 lines, in order to ensure not to exceed the maximum input length of the sentence tokenizer. For `Hansard`, the paragraphs are naturally organized within the XML files. Subsequently, we utilize Spacy (Honnibal and Montani 2017) to segment the paragraphs into sentences. The next step involves postprocessing on the segmented paragraphs, primarily focused on correcting errors stemming from the sentence tokenizer. Furthermore, we find inconsistent usage of semicolons as sentence-ending punctuation. For example, in the UD treebanks, semicolons are occasionally permitted as the concluding punctuation of sentences; at other times, multiple individual sentences connected by semicolons are treated as one sentence. However, the sentence tokenizer tends to split the sentences by semicolons. With simple postprocessing, we manage to correct such segmentation errors by concatenating sentences separated by semicolons. Finally, we apply a filtering process to exclude texts that may not qualify as complete and standalone sentences; this is achieved through the following simple and intuitive rules:

- Sentences must start with a capitalized character.

5 https://www.hansard-archive.parliament.uk/
6 https://evanodell.com/projects/datasets/hansard-data/ (v3.1.0)
7 The original corpus was published in Walter et al. (2021). The version we use contains further clean-up.

- Sentences must end with a period, or a question mark, or an exclamation mark.
- Sentences must contain a verb based on the part-of-speech tags.
- The number of (double) quotation marks must be even.
- The number of left brackets must be equal to that of right brackets.

It is possible that such preprocessing/filtering biases the results later discussed to some extent, as some sentences are ignored through it; however, directly applying sentence segmentation to the original data leads to incomplete sentence segments or data from other parts, e.g., a budget table, as Table 1 (second column) shows, which can result in larger biases for the mismatch with the definition of the dependency parsing task, i.e., parsing the syntactic structure of a sentence, and consequently affect the metrics calculated on a sentence level.

3.1 Validation & Correction

To validate the feasibility of our preprocessing pipeline and study the influence of data noise on metrics such as dependency distance (discussed in §5.1), for each corpus, we apply the pipeline to five randomly sampled paragraphs from each decade, resulting in approximately 150-800 sentences per decade. We then manually review ten randomly selected outputs for each decade and make corrections if any issues are identified. Additionally, the data from the 2020s is merged into the 2010s, given that only partial data is available for the former.

text	date	is sent	has issues	issues	origin of issues	correction
Ich bitte, daß diejenigen Herren, welche für den Fall der Annahme des Z 33b in demselben die Worte „oder an Druck m anderen öffentlichen Orten" aufrecht erhalten wollen, sich von ihren Plätzen erheben.	1883-4-6	TRUE	TRUE	extra material; symbol; spelling	ocr; ocr; historic	Ich bitte, dass diejenigen Herren, welche für den Fall der Annahme des § 33b in demselben die Worte „oder an anderen öffentlichen Orten" aufrecht erhalten wollen, sich von ihren Plätzen erheben.

Table 2: Illustration of the formatted annotation.

Annotation. We first check the **German** sentences from `DeuParl`. Two annotators with an NLP background, one male faculty member who is a German native, and one female PhD student who speaks German fluently, check the outputs. In this phase, we do not have annotation guidelines; the annotators are asked to freely make comments on each text following their intuition and make corrections if they identify any issues in the text.[8] Subsequently, by inspecting the comments and comparing them to the original PDF files,[9] we standardize and format the annotation scheme. As Table 2 shows, the formatted annotation has 5 columns:

- **(1) is_sent**: A **binary** annotation ('TRUE'/'FALSE') for sentence identification, i.e., whether the text is a sentence (minor issues in the sentence, such as space or spelling errors, are allowed).

8 The annotation is done with Google spreadsheet.
9 https://www.reichstagsprotokolle.de/

- **(2) has_issues**: A **binary** annotation ('TRUE'/'FALSE') about the existence of any issues (only if (1) is 'TRUE').
- **(3) issues**: A sequence of labels for the **category of the issues** identified, e.g., spelling errors (only if (1) and (2) are 'TRUE').
- **(4) origin_of_issues**: A sequence of labels for the **origin of the issues**. For example, if a spelling error is a consequence of OCR errors, then the "origin" of it should be "OCR". The labels here are aligned in a one-to-one manner with those in (3) (only if (1) and (2) are 'TRUE').
- **(5) correction**: The **corrected sentence** (only if (1) and (2) are 'TRUE').

Category	**Example**	**Correction**
Spelling	Das ist die Mehrheit; die Diskussion ist geschlof-fen.	[...] geschlossen.
Space	Der EVG-Vertrag spreche von der „westlichen Verteidigung", und im Protokoll der NATO-Staaten sei vom Zusammenschluß der w e s t europäischen Länder die Rede.	[...] westeuropäischen [...]
Missing Material	Ich bin der Meinung — und viele an uns herangekommene Klagen lassen auch darauf schließen — , dass die Auffassung des Reichss-parkommissars, dass das Personal der Deutschen Reichspost voll ausgelastet, aber nicht überlastet sei, nicht richtig ist.	add "—"
Extra Material	[...] daß durch den Bund zweierlei Recht für die Norddeutschen geschaffen werden soll, (Sehr richtig!) daß gewissermaßen zweierlei Klassen von Norddeutschen geschaffen werden sollen, (Sehr gut!) eine Selekta, die vermöge ihrer Gesit-tung [...]	delete "(Sehr gut!)" and "(Sehr richtig!)"
Punctuation & Symbol	Ich bitte, daß diejenigen Herren, welche für den Fall der Annahme des Z 33b in demselben die Worte „oder an Druck m anderen öffentlichen Orten" aufrecht erhalten wollen, sich von ihren Plätzen erheben.	[...] § 33b [...]

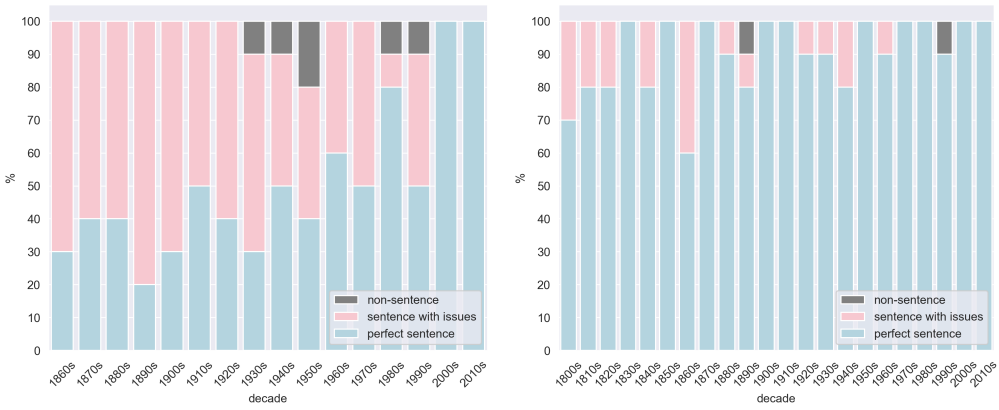
Table 3: **Categories of issues**; each is shown with an example and the corresponding correction. The parts having a certain issue or missing in the sentences are in red.

We classify the **issues** into 5 categories: *Spelling*, *Space*, *Missing Material*, *Extra Material* and *Punctuation & Symbol*, as shown in Table 3. Additionally, we split the **origins of issues** into 4 categories: (1) *Historic*, which indicates the issues are due to the differences in historical and modern language. For example, 'daß' (modern spelling: 'dass') is a spelling issue with the *historic* origin. (2) *OCR*, which indicates the issue is an OCR error. For example, the spelling issue in Table 3 results from an OCR error ('geschloffen' vs. 'geschlossen'). (3) *Genre*, which represents issues that exist due to the characteristics of the text genre. In the text of political debates, the language changes between written and spoken language frequently, and also, there exist interjections within a sentence, as shown in the fourth row of Table 3 (Extra Material). (4) *Preprocessing*, which denotes that the issues stem from our preprocessing pipeline, e.g.,

extra space. For details regarding those categories, we refer to Figure 18 in the appendix, which demonstrates our annotation guidelines following the refined annotation scheme.

Next, two annotators annotate the **English** text from `Hansard`, following the annotation scheme defined above. As the original PDF files of `Hansard` are inaccessible, we cannot identify the origins of the issues; thus, this annotation is omitted for `Hansard`. Among the annotators, one is a male faculty member, while the other is a female Master's student, both of whom speak English fluently and work in the NLP field.

To check the **agreement** among annotators, 40 German sentences are jointly annotated, 20 for the sentence identification task and 20 for the issue identification/correction task. For English, 30 sentences are commonly annotated for all annotation subtasks. For each binary annotation task, we calculate Cohen's Kappa (Cohen 1960) for inter-agreement, whereas for the correction task, we compute the edit distance (Levenshtein 1965) between sentences. Let s be the original sentence, and s'_a and s'_b be the sentences corrected by annotators A and B respectively. We calculate the edit distance (i) between s and s'_a ($ED(s, s'_a)$), (ii) between s and s'_b ($ED(s, s'_b)$), and (iii) between s'_a and s'_b ($ED(s'_a, s'_b)$). If the corrections are similar to each other, (i) and (ii) should be close while (iii) should be small. Overall, the annotators obtained a decent level of agreement, with ~ 0.7 on binary tasks and the edit distances of (i) = 3.88, (ii) = 3.96, and (iii) = 0.54 for the correction.

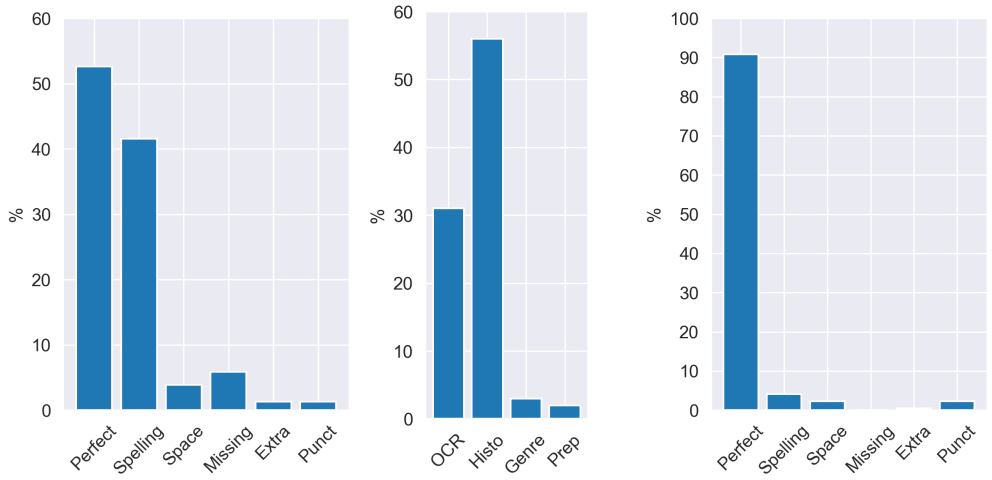


(a) Deuparl (b) Hansard

Figure 3: Distribution of the texts identified as perfect sentences (without any issues), sentences with issues, and non-sentences over time.

Results. We show the distribution of the texts identified as perfect sentences (without any issues), sentences with issues, and non-sentences in Figure 3. We observe that: (1) German data has more issues compared to English. For the older time periods, e.g., the first 3 decades in each corpus, 60-70% German sentences are found to contain issues, while there are only 20-30% English sentences found to be problematic. (2) Only a small portion of sentences were identified as non-sentences. Specifically, 6 out of 160 texts (~3.75%) from `DeuParl` and 2 out of 220 texts (~0.1%) from `Hansard` are found not to be sentences. (3) The proportion of problematic sentences is decreasing over time; for both languages, all sentences after 2000 are flawless.

Next, we present the percentages of the sentences having a specific issue in Figures 4a and 4c. Note that one sentence can have multiple issues, so the sum of the percentages does not need to equal 100%. In English data, 90% of the sentences are perfect — only 1%-2% of the sentences have *Spelling*, *Space*, or *Punct* issues. In contrast, only half of the German sentences are issue-free, with over 40% of the sentences having *Spelling* issues and 1%-5% of the



(a) **Deuparl** - issues (b) **Deuparl** - origins (c) **Hansard** - issues

Figure 4: (a)/(c): Percentage of perfect sentences and sentences with a specific issue to all texts identified as sentences. (b): Percentage of sentences with issues from a specific origin to all sentences containing issues.

sentences containing other issues. Furthermore, we find that *Historic* origin issues dominate in the problematic German sentences—over 55% of those contain at least one *Historic* origin issue, followed by *OCR* origin—~30% of the flawed sentences have issues raised from OCR errors.

In summary, our annotations suggest that English data from `Hansard` is substantially less problematic than the German data from `Deuparl`. Notably, the most prevalent issues in the German data, namely historical spelling and OCR spelling errors, rarely appear in the English data. This leads us to question whether the `Hansard` Corpus has already undergone preprocessing like spelling normalization. Through contact with the UK Parliament enquiry services,[10] we received feedback indicating that the data has not been preprocessed regarding spelling normalization. As for OCR errors, a statement on their website[11] explicitly mentions that OCR errors are curated once identified. As a consequence, we conclude that English has undergone smaller changes in terms of spelling in the past 200 years compared to German, making German the more interesting (and much less researched) language from a historical perspective.

3.2 Dataset Construction

We apply the preprocessing pipeline introduced above to get up to 200k sentences for each decade and language. To make the English and German data comparable, we limit the time period for English to the decade 1860-2020. Next, we tokenize the sentences and do part-of-speech tagging using Stanza (Qi et al. 2020). The length of sentences is then the number of tokens after tokenization. Figures 19 (in the appendix) and 5 show the distribution of sentence lengths and the average sentence length over time. We see that shorter sentences appear more often in the later periods and thus the average sentence length is overall decreasing over time for both languages. For instance, the average sentence length in DeuParl decreases from ~35 to ~20 in the time

10 https://www.parliament.uk/site-information/contact-us/
11 https://hansard.parliament.uk/about

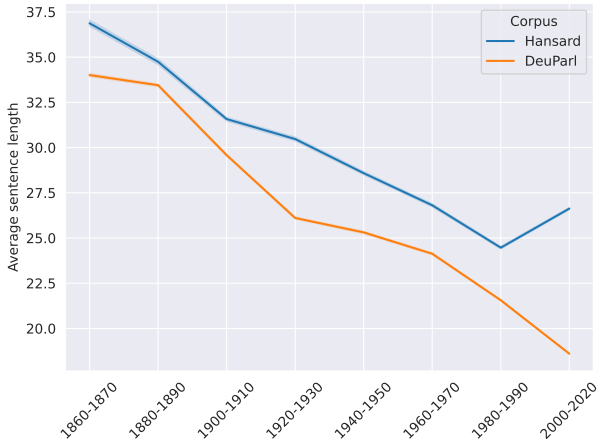


Figure 5: Average sentence length per decade group over time.

window 1860-1870 to 2000-2020. The only exception is the last subtrend for Hansard, where the average length increases by ∼2 tokens. Metrics like *mDD* are highly sensitive to sentence length (Lei and Jockers 2020), thus, it is important to control for sentence length when constructing the dataset for observation.

Final Dataset. We select sentences with varying number of tokens for observation: 5, 10, 15, 20, 30, 40, 50, 60, and 70, covering short, middle and long sentences. We aim to observe syntactic changes in sentences of the same length (i.e., same amount of tokens). Nevertheless, to obtain sufficient sentences for each length and time period, we set an offset of 2 tokens for sentence grouping, e.g., sentences of a length from 10 to 12 are grouped together; in addition, we consider decade groups with each spanning 2 decades (except for the last one, which spans 3 decades). For both English and German, the final dataset contains 450 sentences per length and decade group. These sentences are randomly sampled from a total of 32,400 preprocessed sentences for each language.

4. Dependency Parsers

Unlike previous approaches, which solely relied on a single parser, typically the Stanford CoreNLP Parser (Manning et al. 2014), we will base our observations on various parsers with the two most popular design choices: transition-based and graph-based. Transition-based parsers sequentially predict the arc step by step based on the current state, while graph-based parsers globally optimize the dependency tree, aiming to find the highest-scored one. The parsers used in this work include: (1) **transition-based**: CoreNLP (Manning et al. 2014), StackPointer (Ma et al. 2018); (2) **graph-based**: Biaffine (Dozat and Manning 2016), Stanza (Qi et al. 2020), TowerParse (Glavaš and Vulić 2021) and CRF2O (Zhang, Li, and Zhang 2020). Details of the parser configurations are provided in §9.1 in the appendix.

4.1 Evaluation

To explore the reliability of the parsers used, we evaluate them on (1) existing **UD treebanks**, (2) the treebanks built upon the two target corpora of this work (denoted as "**Target Treebanks**"), and (3) adversarially attacked treebanks concerning the most severe issues identified in §3.

Metrics. We calculate the **Unlabeled Attachment Score (UAS)** and **Labeled Attachment Score (LAS)** for the parser performance. UAS and LAS are originally defined as the percentage of tokens with correctly predicted heads and head+relations respectively; in this scenario, the input of the parser is sentences pre-tokenized by humans. On the other hand, Zeman et al. (2018) test parser performance in a real-world scenario, i.e., parsing the raw text without any gold standard (e.g., sentence segmentation, tokenization, part-of-speech tags, etc.). The metrics are then re-defined as the harmonic mean (F1) of precision P and recall R, where P is the percentage of the correct head (+relation) to the number of predicted tokens, while R is the percentage of the correct head (+relation) to the number of tokens in the gold standard. As we will use the parsers on sentences without tokenization, we argue that *the re-defined metrics on raw texts are more indicative in our case*. For all evaluations, we adopt the evaluation script from the CoNLL18 dependency shared task (Zeman et al. 2018).[12]

Evaluation on UD Treebanks. To ensure the correctness of the training, implementation, and usage of the parsers, we first evaluate the parsers on the test sets of the existing UD treebanks, for which we choose Universal Dependencies (UD) v2.12,[13] and compare the results to Zeman et al. (2018). Although our evaluation setup, parsers, and the version of the UD treebanks are different from theirs, comparable (or better) results are expected since (1) the discrepancy in UD versions is small,[14] (2) most of the tested parsers are published later than the shared task and (3) the parsers do not need to split the text into sentences by themselves here.

	UAS		LAS	
	UD	TARGET	UD	TARGET
CoreNLP*	78.6±5.5	82.9 (+4.3)	73.4±6.5	78.7 (+5.3)
Stanza*	88.2±6.2	88.0 (-0.2)	84.9±8.0	85.0 (+0.1)
TowerParse*	87.1±4.5	**92.8 (+5.7)**	82.8±5.9	**90.3 (+7.5)**
StackPointer	85.7±5.2	84.2 (-1.5)	81.3±6.2	80.2 (-1.1)
CRF2O	87.2±4.6	86.6 (-0.6)	83.5±5.4	82.8 (-0.7)
Biaffine	**91.6±2.6**	90.1 (-1.5)	**88.5±3.3**	87.2 (-1.3)

(a) English

	UAS		LAS	
	UD	TARGET	UD	TARGET
CoreNLP*	74.0±2.9	74.4 (+0.4)	67.3±2.9	69.6 (+2.3)
Stanza*	84.3±4.1	87.6 (+3.3)	78.8±4.4	82.0 (+3.2)
TowerParse*	86.2±2.0	89.7 (+3.5)	80.3±1.3	84.5 (+4.2)
StackPointer	83.6±1.3	86.9 (+3.3)	78.2±1.6	80.8 (+2.6)
CRF2O	78.5±1.5	81.9 (+3.4)	70.2±2.4	72.9 (+2.7)
Biaffine	**87.0±2.5**	**90.8 (+3.8)**	**81.7±1.8**	**84.5 (+2.8)**

(b) German

Table 4: UAS and LAS on **`UD`** and **`Target`** treebanks. We show the macro average and the standard deviation of the metrics over the UD treebanks. The best performance in each criterion is bold and the second best one is underlined. We mark the parsers trained on mismatched data (with respect to UD versions or treebanks) with *. Values in brackets indicate the absolute differences.

Results. In §9.2 in the appendix, we confirm the legitimacy of the used parsers by showing that the performance of our used parsers is mostly superior to those in Zeman et al. (2018). Next, we show the macro average and the standard deviation of the UAS and LAS over the UD treebanks in Table 4 (column "**UD**"). Biaffine consistently outperforms the others across different languages in terms of both UAS (en: 92% vs. 79%-88%; de: 87% vs. 79%-86%) and LAS (en: 89% vs. 73%-85%; de: 82% vs. 70%-80%). Furthermore, it also exhibits the strongest robustness across various treebanks, as demonstrated by the smallest average standard deviation (~2.55 percentage points

12 https://universaldependencies.org/conll18/evaluation.html; this script ignores the subtypes of the relations (e.g., for relation 'acl:relcl' (relative clause modifier), only 'acl' is considered.), but here, we consider the whole relations.

13 We exclude the treebank "GUMReddit" for English, as there are many nonsensical annotations.

14 Most changes are about fixing a tiny portion of incorrect annotations, e.g., see change logs on https://github.com/UniversalDependencies/UD_English-PUD/blob/master/README.md and https://github.com/UniversalDependencies/UD_German-GSD/blob/master/README.md.

(pp)) over different metrics and languages. It is noteworthy that CoreNLP, the parser leveraged by all other related approaches, is the worst in our evaluation, according to all criteria.

Stanza trained on slightly mismatched data outperforms the other fine-tuned ones on average for English. Nevertheless, it is the least robust one according to its high standard deviations (6.2 pp vs. <5.5 pp in UAS; 8.0 pp vs. <6.5 pp in LAS). Interestingly, Towerparse, which was trained on mismatched data in terms of treebanks and versions, performs similarly to Biaffine on German treebanks, only ~1 pp lower UAS/LAS. However, we observe that it tends to predict multiple roots for one sentence and produce cycles in the dependency tree graphs, violating the UD dependency definition. In our evaluation, it generates 440 multi-root predictions and 1,498 cycles, while the others rarely make such errors (only Stackpointer predicts multiple roots 28 times among the other parsers). Besides, it also has to skip 28 sentences due to the limitation of its implementation.[15] CRF2O performs on par with the second-tier parsers on English data but ranks as the second-worst on German data. The reason may be that it is not capable of parsing non-projective relations, which occur more frequently in German compared to English, as shown in Gómez-Rodríguez and Ferrer-i Cancho (2017). With its built-in evaluation function, which skips the non-projective sentences in the gold standard (~3k out of ~23k sentences), the UAS and LAS raise from <80% to >90% on German treebanks.

Evaluation on Target Treebanks. We leverage the dependency annotations from a private unpublished work accessible to us, which collects human annotations by manually correcting the automatic parsing results; it includes annotations in UD format for 111 sentences from Hansard and 163 from DeuParl from different time periods. Sentences were cleaned up before parsing. There were two annotators involved, who annotated independently and in case of disagreement resolved them. However, there is no agreement reported and it is unclear how reliable these human annotations are, as they correct the output from the parser, potentially leading to bias in correction ("priming effect").

Results. We show the evaluation results on the Target treebank in Table 4 (column "**Target**"). The values in the brackets indicate the absolute difference in metrics between the Target treebank and the UD treebanks. We observe that the Target treebank mostly does not negatively affect the parsers' performance. All parsers exhibit improved performance on the **German** Target treebank compared to the UD treebank, with an increase ranging from 0.4 to 4.2 pp. For **English**, the maximal performance drop is observed with StackPointer and Biaffine (1.1-1.5 pp), while CoreNLP and TowerParse achieve better results, with an increase ranging from 4.3 to 7.5 pp. Stanza and CRF2O display only minor fluctuations (<1 pp). Hence, our results weakly support the parsers' proficiency in the target domain, considering potential bias in human annotations.

Evaluation on Adversarial Treebanks. To inspect the effect of data noise on the parsers, we generate two adversarial datasets concerning the most two prevalent issues in German data identified in §3, i.e., historical spelling and OCR-raised spelling errors. We perturb the merged German test set to create the adversarial treebanks. To perform the historical spelling attack, we chose 23 candidate words from our previous annotations and replaced them with their historical spellings in the test set. This resulted in 1,971 sentences that were affected by the attack. For (2), we randomly replace 10%, 30%, or 50% of the characters in 1 or 2 tokens within a sentence with random characters, resulting in 6 attacking levels; we produce 2000 sentences per attack level.

Results. We illustrate the absolute difference in metrics between the sentences before and after attacking in Figure 21 in the appendix. Both types of attacks consistently have a greater influence

15 See the authors' explanation about skipping sentences in footnote 7 in Glavaš and Vulić (2021).

on LAS compared to UAS. This can be explained by the fact that LAS is more sensitive to perturbations because it considers both attachment and labeling accuracy, whereas UAS only considers attachment accuracy. Even if the attachment remains correct, changes to the words themselves (such as spelling or character replacement) can result in incorrect labels. As depicted in Figure 21a, **historical spelling** has a slight negative impact on the parsers' performance, with a difference of less than 0.07 pp in LAS and that of up to around 0.05 pp in UAS. Among the parsers, Biaffine is the most stable one, followed by TowerParse; CoreNLP exhibits the lowest robustness.

We present the results for **OCR spelling attack** in Figure 21b. OCR spelling errors have a slightly larger impact, degrading parser performance up to around 0.12 pp in LAS and 0.09 pp in UAS, when 50% characters of 2 tokens within a sentence are perturbed. Again, Biaffine is the most robust against OCR attack, followed by CRF2O and TowerParse, which demonstrate similar sensitivity to Biaffine when the attack level is high. At the highest attack level, CoreNLP, StackPointer, and Stanza show similar sensitivity. However, at low attack levels, StackPointer and Stanza demonstrate better robustness compared to CoreNLP. In real-world cases, OCR spelling errors often involve only one character within a sentence. For example, in the sentence "... wollen *Tie* nun das letzte nehmen." (from our annotation in §3), the only OCR error is in "*Tie*", which should be corrected to "*Sie*". Hence, Stanza and StackPointer are still superior to CoreNLP in real-world scenarios.

4.2 Summary

All parsers exhibit robustness and show minor (or no) degradation both under adversarial attacks and the target data. Among them, Biaffine shows the best performance in standard evaluations as well as the highest robustness under adversarial conditions. CoreNLP and CRF2O are often the worst ones in our evaluation; besides, they are unable to predict crossing edges, which we deem as an important indicator of language change. Non-projective sentences seem to substantially degrade CRF2O's performance; thus, we exclude it from further experiments. CoreNLP is retained, however, for comparison purposes. Therefore, *we will use five parsers to perform the large-scale analysis in §6: CoreNLP, StackPointer, Biaffine, TowerParse, and Stanza.*

5. Metrics

Unlike most of the relevant approaches that only focus on the linear dependency distance (e.g., Lei and Wen (2020); Liu, Zhu, and Lei (2022); Zhu, Liu, and Pang (2022); Zhang and Zhou (2023)), we examine a large array of metrics that capture statistical patterns driven by graph theory or linguistic phenomena. In the following, we outline the definition and usage of these metrics over a German sentence—"Das hat alles sehr gut geklappt!" (extracted from the UD GSD treebank).

Root Distance (d_{root}). d_{root} is the index of the word connecting to the pseudo root in a sentence. An example is shown in Figure 6, where d_{root} is 6 (from root to 'geklappt'). Lei and Jockers (2020) consider d_{root} as a normalizing factor when computing the mean of dependency distances for all dependency pairs in a sentence. Liu, Zhu, and Lei (2022) showed that d_{root} correlates positively with dependency distance over time in English; specifically, d_{root} increases over time for long sentences while decreasing over time for short sentences.

Mean Dependency Distance (mDD) & Normalized Mean Dependency Distance (nDD). Following Lei and Wen (2020); Zhang and Zhou (2023); Lei and Jockers (2020), we consider two statistical

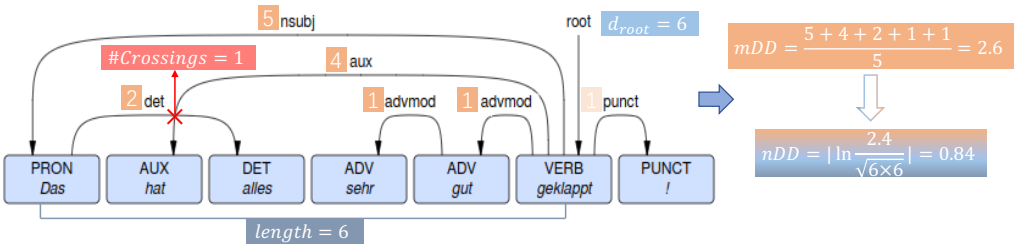


Figure 6: Illustration of observing d_{root} and ***#Crossings***, along with calculating ***nDD*** and ***mDD*** using equation 1 and 2, with the example sentence "Das hat alles sehr gut geklappt!". Arrows point from head to dependent tokens. Values with an orange background indicate the dependency distance between a dependent-head pair.

features of dependency distances, namely *mDD* and *nDD* given by:

$$mDD = \frac{1}{n} \sum_{(i,j)\in D} |i - j| \tag{1}$$

$$nDD = \left| \log \left(\frac{mDD}{\sqrt{d_{root} \times \text{length}}} \right) \right| \tag{2}$$

where $|i - j|$ represents the absolute difference between the position indices of a dependency pair (i, j), D is a set of dependency pairs in a sentence (punctuation and root dependencies are excluded), length is the number of words in a sentence, and n is the size D. In Figure 6, *mDD* and *nDD* are 2.6 and 0.84 for example.

Number of Crossings (#Crossings). A crossing is said to be observed when two dependency relations overlap. Ferrer-i Cancho and Gómez-Rodríguez (2016) show that *#Crossings* correlates positively with dependency distance in 21 out of 30 languages, while Liu, Xu, and Liang (2017) find that machine-generated artificial texts without crossings produce shorter dependency distances compared to those with crossings. We count the number of crossings in a sentence, ignoring punctuations. An example is shown in Figure 8: the dependency relation between "Das" and "alles" crosses over the relation between "geklappt" and "hat", i.e., *#Crossings* = 1.

Number of leaves (#Leaves). We count the number of leaves (*#Leaves*) of a dependency tree, excluding punctuations—see an example in Figure 7 (left).

Tree Height (Height) & Longest Path Distance (Height$_{dependency}$). *Height* is the longest-path distance from the root node to a leaf, while $Height_{dependency}$ is the weighted longest-path distance where an edge over two nodes is weighted by the dependence distance between the nodes. An example can be observed in Figure 7 (left): *Height* = 3, $Height_{dependency}$ = 13.

Depth Variance (depthVar) & Depth Mean (depthMean). We denote the depth of a node as the path length from the node to the root. *depthMean* and *depthVar* represent the mean and variance of the depths of all nodes in a tree, as illustrated in Figure 7 (middle).

Tree Degree (treeDegree) & Degree Variance (degreeVar) & Degree Mean (degreeMean). We measure the degree of a node as the number of outgoing nodes attached to that node, i.e., by

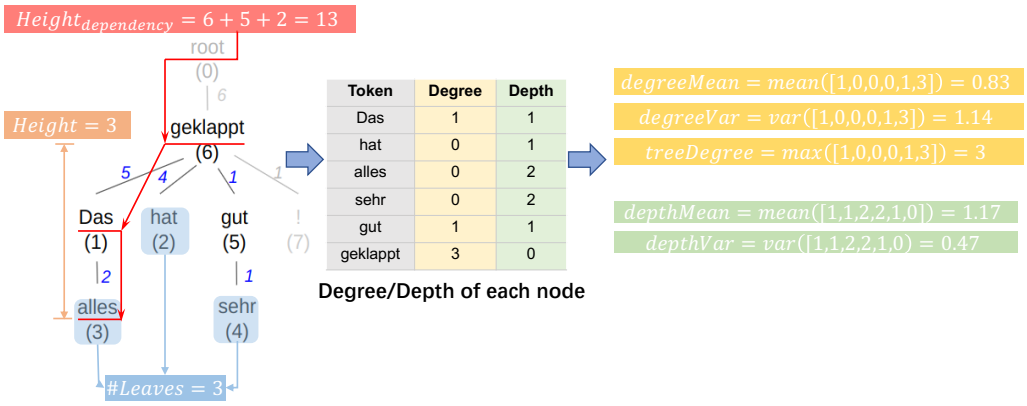
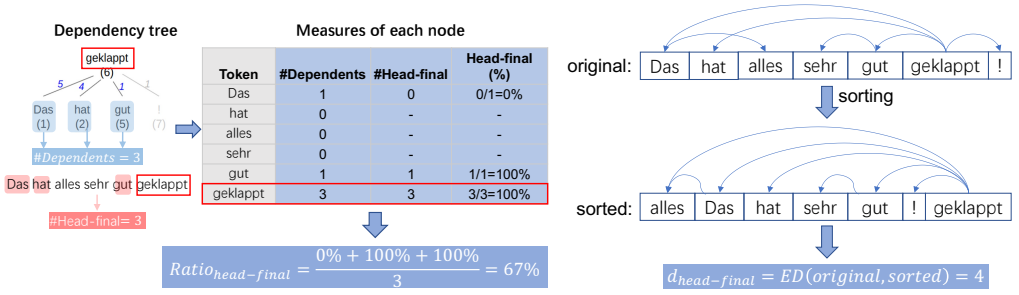


Figure 7: Illustration of calculating/observing of ***Height***, $\boldsymbol{Height_{dependency}}$, ***#Leaves***, ***treeDegree***, ***degreeMean***, ***degreeVar***, ***depthMean*** and ***depthVar*** with the example sentence "Das hat alles sehr gut geklappt!". Values in brackets are the position indices of the tokens. Values beside edges indicate the dependency distance of the corresponding dependency pairs.

counting the dependents of a head word in a dependency tree. The higher the degree of a node, the more dependents a head word has. This metric is related to dependency distance—as both long dependency distances and head words with high degrees can complicate the syntactic structure.We also consider statistical features: the mean (*degreeMean*) and variance (*degreeVar*) of the degrees of all nodes, as well as the maximum of all nodes' degrees (*treeDegree*).



(a) $Ratio_{head\text{-}final}$ (b) $d_{head\text{-}final}$

Figure 8: Illustration of calculating/observing $\boldsymbol{Ratio_{head\text{-}final}}$ and $\boldsymbol{d_{head\text{-}final}}$ with the example sentence "Das hat alles sehr gut geklappt!". Arrows point from head to dependent tokens.

Head-final Ratio ($Ratio_{head\text{-}final}$) & Head-final Distance ($d_{head\text{-}final}$). For a dependency pair, if a head follows its dependent, then the pair is termed as a head-final pair. Otherwise, if a head precedes its dependents, then the pair is considered head-initial. In linguistics, this head directionality is profiled as a crucial parameter for classifying languages (Liu 2010). In Figure 8a, for each head in a sentence, we calculate the percentage of its head-final pairs to all pairs with that head (head-final(%)), and then average the head-final(%) over all heads in that sentence, termed as $Ratio_{head\text{-}final}$. We note that the sentence "Das hat alles sehr gut geklappt!" not only includes head-final dependency pairs but also head-initial pairs, e.g., the head 'Das' precedes its dependent 'alles.' Here, we consider how distant this sentence is, with a mix of head-initial and head-final

pairs, compared to the same sentence but permuted to have only head-final pairs. This results in the metric $d_{head\text{-}final}$ that is the Levenshtein edit distance (Levenshtein 1965) between the original and permuted sentences.

Random Tree Distance ($d_{randomTree}$). We randomly permute the dependency tree structure in a sentence and then compute the tree edit distance between the original and permuted trees using the algorithm of Zhang and Shasha (1989).

5.1 Sensitivity of metrics to data noise

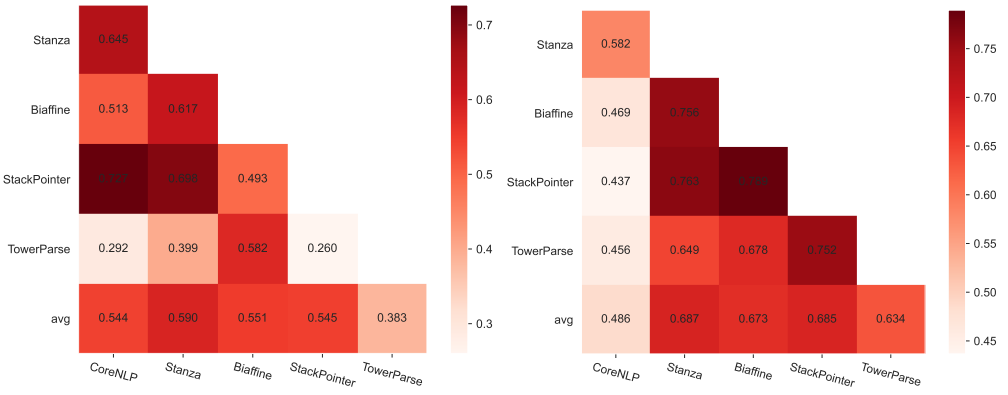
While our results in §4 indicate that the parsers perform decently on the target corpora, it remains uncertain whether the metrics derived from their parsing results are affected by data noise and to what extent. Since we will inspect the trends of the metrics over time, it is important to ensure that the data noise does not disrupt the rankings of those metrics. In §9.3 in the appendix, we demonstrate the insensitivity of our metrics to data noise by showing that the metrics for the clean data are highly correlated to those for the noisy data (overall 0.926 Spearman).

6. Analysis

Following Zhu, Liu, and Pang (2022); Liu, Zhu, and Lei (2022), we leverage the Mann Kendall (MK) trend test (Mann 1945) to indicate the diachronic change of the concerned metrics. The MK trend test is widely used to detect trends in time series. It has three outputs: "increasing", "decreasing", and "no trend"; an increasing or decreasing trend is detected only if it is significant with a p-value < 0.05. We perform it for the 9 sentence lengths and 15 metrics, totalling $15 \times 9 = 135$ test cases.

We first explore the dependence of the trends on parsers in §6.1, as the relevant approaches rely on a single parser (e.g., Zhu, Liu, and Pang (2022); Liu, Zhu, and Lei (2022)). Then, we compare the syntactic changes between English and German in the remaining sections. Specifically, we inspect the overall similarity of syntactic changes between English and German in §6.2, with a focus on sentence length in §6.3 and a particular metric in §6.4. Finally, we analyze different and similar trends between the two languages in §6.5 and §6.6 respectively.

6.1 What is the dependence of language change trends on parsers?



(a) English (b) German

Figure 9: Cohen's Kappa of MK results based on the dependency relations predicted by different parsers. Deeper colors denote higher agreement and vice versa. We show the average agreement for each parser in the last rows.

The outputs of MK can be seen as the labels of a three-way classification task. Therefore, we calculate Cohen's Kappa (Cohen 1960) to indicate the agreement between each parser pair. The values in Figure 9 are derived by calculating Cohen's Kappa across the 135 MK outputs based on different parsers for each language. We then average the agreements over all parser pairs containing a certain parser to achieve the average agreement for each parser, shown in the last rows in Figure 9.

Most parsers demonstrate a moderate agreement with others (0.4-0.6). On **English** data, as illustrated in Figure 9a, the highest agreement of 0.73 is observed between CoreNLP and StackPointer, succeeded by Stanza and StackPointer (0.70). Among the parser pairs, TowerParse and StackPointer obtain the lowest agreement of 0.26. Stanza obtains the highest average agreement with others (0.59), whereas TowerParse has the lowest one (0.38). For **German** (Figure 9b), the highest agreement is achieved between Biaffine and StackPointer (0.79), followed by Stanza and StackPointer (0.76); Stanza again obtains the highest average agreement (0.69), while CoreNLP least agrees with others on German data (0.44-0.58). In summary, our observations indicate that for both languages, ***the use of different parsers may lead to various conclusions about the diachronic trends of those metrics***, despite their decent performance observed in §4. This raises concerns about previous approaches which solely rely on one parser.

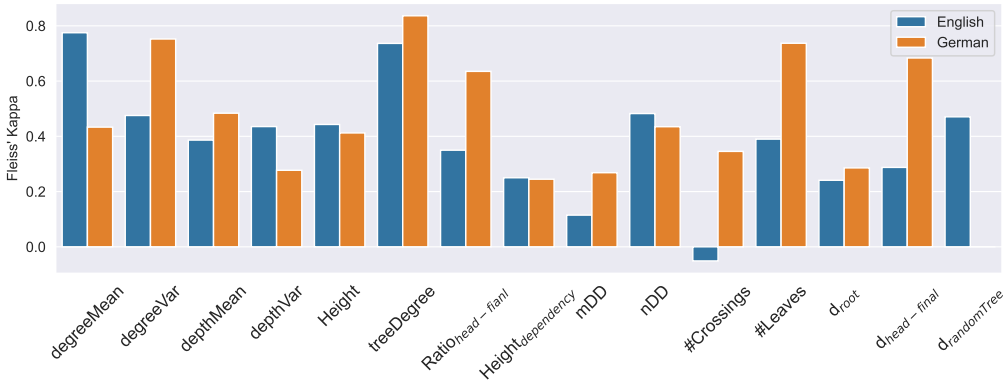


Figure 10: Fleiss' Kappa of the parsers on each metric.

Next, we examine the parsers' agreement on a trend for each metric. We regard the MK result based on each parser as an 'annotation' for the corresponding trend; thus, for each trend, we have 5 different 'annotations', allowing for calculating the Fleiss' Kappa (Fleiss 1971) among them. For each metric, we have 9 'annotation' instances, resulting from the 9 sentence lengths. As Figure 10 shows, most values range from ~0.2 to ~0.5, implying a consistently low level of agreement among the parsers across the metrics. *treeDegree* is the only metric where the parsers obtain a high agreement across both languages (~0.8 Fleiss' Kappa). Among the metrics, parsers show the second least agreement on trends in *mDD* for English (~0.1 Fleiss' Kappa). This further questions the previous works in reporting diachronic trends of dependency distance based on a single parser, as results may change when using a different parser.

As a result, we employ a **majority vote** approach to identify the most likely stable trends in the metrics: if the increasing/decreasing trends can be detected based on at least 3 out of 5 parsers, we consider them valid; an exception is made for *#Crossings*, where we relax the threshold to 2 out of 5 due to CoreNLP's inability to predict crossing dependencies. The final trends for the 135 cases are listed in Table 6 in the appendix, which we use for further analysis.

6.2 Are syntactic changes more alike or distinct in English and German?

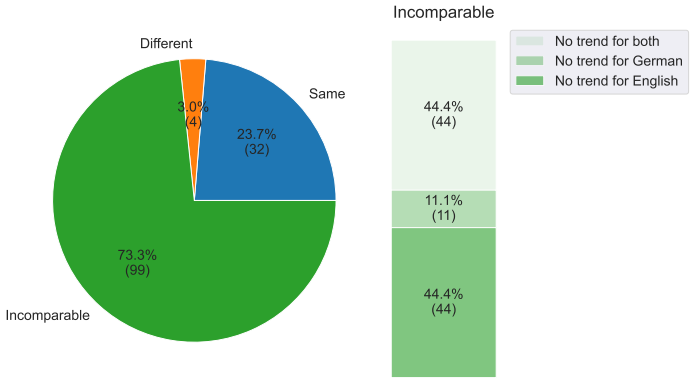


Figure 11: Left: percentages of the same, different and incomparable trends between English and German. Right: Percentages representing incomparable trends due to 'no trend for English,' 'no trend for German,' and 'no trend for both' to all incomparable trends. Values in brackets indicate the counts of trends.

We compare the 135 trends between English and German for the 15 metrics and 9 sentence lengths. As shown in Figure 11 (left), among all the trends, 99 cases (73.3%) are incomparable since there are no significant trends for German or English or both; 32 cases (23.7%) have the same trends, while only 4 cases (3%) show different trends, which may suggest a more uniform ("convergence"), rather than diverse, syntactic change between English and German. Among the incomparable trends, as shown in Figure 11 (right), 44 cases (44.4%) are the result of no significant trends for both English and German, 11 cases (11.1%) lack significant trends for German, and 44 cases (44.4%) have no significant trends for English. From this perspective, the German language has undergone more significant syntactic changes compared to English.

6.3 Which sentence lengths exhibit the most distinct or similar syntactic changes between English and German?

We count the number of metrics showing different/same/incomparable trends between English and German for each sentence length, displayed in Figure 12. We observe that: (1) for sentences of lengths 5-7, 10-12, 30-32 and 70-72, 1 out of 15 metrics exhibits different trends, whereas there are no different trends for the other lengths; (2) *sentences of length 5-7 behave most similarly*, shown with the highest number of metrics with the same trends (8 out of 15) among sentence lengths, followed by the longest sentences of length 70-72, which have 7 out of 15 metrics with the same trends; (3) the number of metrics with the same trends increases from 2 to 7 as sentence lengths go from 30-32 to 70-72. We also note that the distribution of the incomparable trends (green bars) seem to follow a Gaussian distribution, where there are more incomparable trends for the sentences of middle lengths (e.g., 15-22), while less incomparable trends exist for the shortest and longest sentences. Interestingly, as we show in Figure 19, sentences of middle lengths dominate in the corpora; this may suggest that syntactic changes are more likely to occur in sentences of extreme lengths (i.e., longer and shorter sentences) rather than in sentences of common lengths.

6.4 Which metrics exhibit the most distinct or similar trends between English and German?

We count the number of sentence lengths showing different/same trends between English and German for each metric, visualized in Figure 13. Only one metric shows both same and different

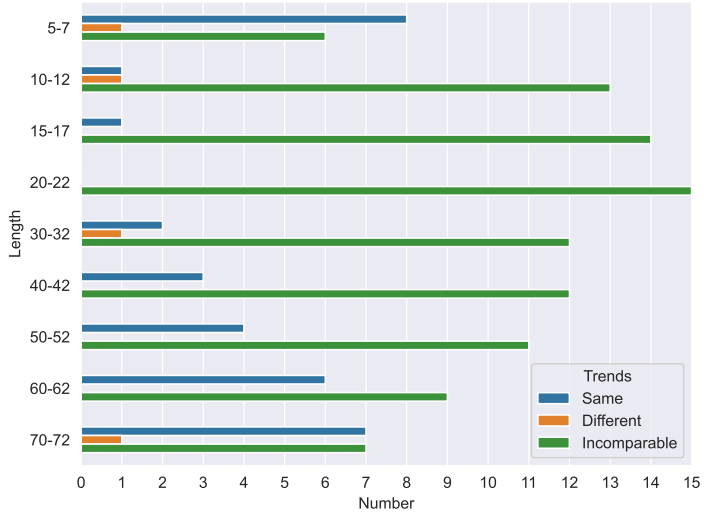


Figure 12: Number of same/different/incomparable trends for each **sentence length**. Larger numbers indicate more metrics exhibiting the same/different/incomparable trends between English and German for a specific sentence length group.

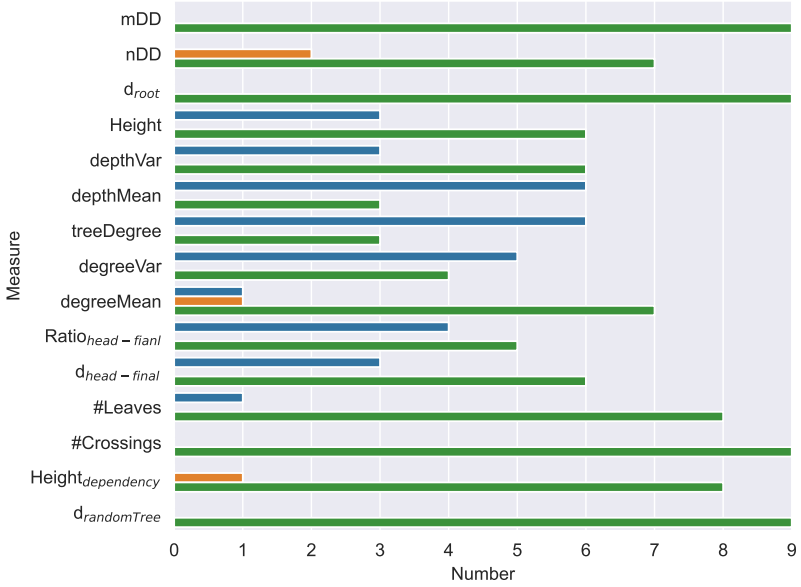
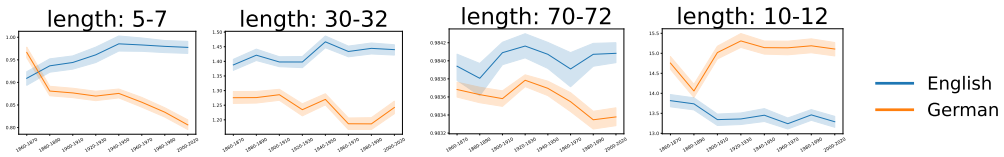


Figure 13: Number of same/different/incomparable trends for each **metric**. Larger numbers indicate more sentence lengths exhibiting the same/different/incomparable trends between English and German for a specific metric.

trends for various sentence lengths, namely *degreeMean*. *nDD behaves most differently*—for 2 out of 9 lengths, it exhibits different trends—followed by *degreeMean* and $Height_{dependency}$, whose

trends are different for one length. *depthMean and treeDegree are the most similar ones regarding changes in English and German*; for 7 out of 9 sentence lengths, they exhibit the same trends between English and German. *#Leaves* and *degreeMean* have the same trend only for one length. Among the others, *Height*, *depthVar*, *degreeVar*, $d_{head\text{-}final}$ and $Ratio_{head\text{-}final}$ behave similarly for 3-5 out of 9 sentence lengths.

6.5 What are the differences in syntactic changes between English and German?



(a) *nDD* (b) *degreeMean* (c) $Height_{dependency}$

Figure 14: Metrics showing different diachronic trends based on MK between English (blue) and German (orange), averaged per decade group and over the 5 parsers. We show 95% confidence intervals.

We visualize the 4 distinct diachronic trends of metrics between English and German in Figure 14. It shows the metrics for English (blue) and German (orange) over time, averaged over the 5 parsers per decade group. For ***nDD*** (Figure 14a), the trend goes up for English but down for German across both sentence lengths. Figure 14b shows the trends of ***degreeMean*** for sentences of length 70-72. Overall, while *degreeMean* increases for English over time, it decreases for German, suggesting that, on average, words govern more dependents over time for long sentences in English but fewer in German. $\boldsymbol{Height_{dependency}}$ of sentences having 10-12 words generally increases for German over time but decreases for English, as shown in Figure 14c, indicating a tendency of the sum dependency distance from the root to a leaf word becoming larger in German but smaller in English for sentences of that length.

An interesting observation when closely examining the four different trend patterns in Figure 14 is that for two (or even three) out of four cases, subtrends look quite similar across English and German. Witness *degreeMean* as a point in case: there is a sequence of four subtrends: downward, upward, downward, upward in both English and German, where the German pattern seems to be lagged compared to the English one (i.e., German seems to mimic the English pattern). This is a limitation of our overall trend analysis which takes a global pattern into account, disregarding local subpatterns (cf. Simpson's paradox (Simpson 1951)).

6.6 What are the similarities in syntactic changes between English and German?

Based on the MK trend test results, we identify 32 trends for 9 metrics that demonstrate simultaneous increase or decrease in both English and German. We show the average of those metrics per decade group over the 5 parsers in Figure 15; blue lines represent the trends for English and orange lines for German.

For the metrics assessing dependency trees vertically, namely ***Height*** (Figure 15a), ***depthVar*** (Figure 15b) and ***depthMean*** (Figure 15c), the trends are similar: they decrease over time for short sentences having 5-7 and/or 10-12 words but increase for long sentences like those of length 60-62 and 70-72. In contrast, the degree-relevant metrics, which assess trees horizontally, increase for short sentences but decrease for long sentences: in Figure 15e, 15d and 15f, we observe an increase in ***degreeVar***, ***treeDegree***, and ***degreeMean*** for sentences having 5-7 and/or 10-12 words, as well as

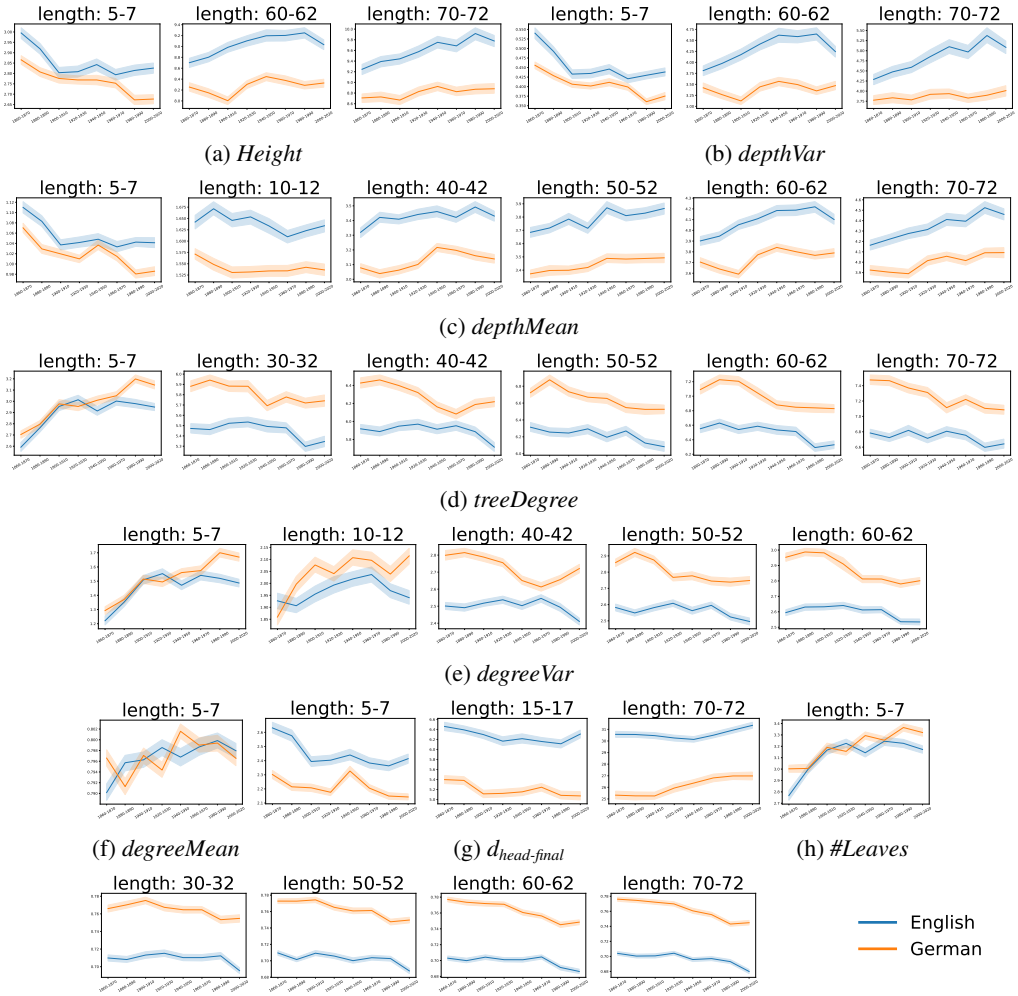
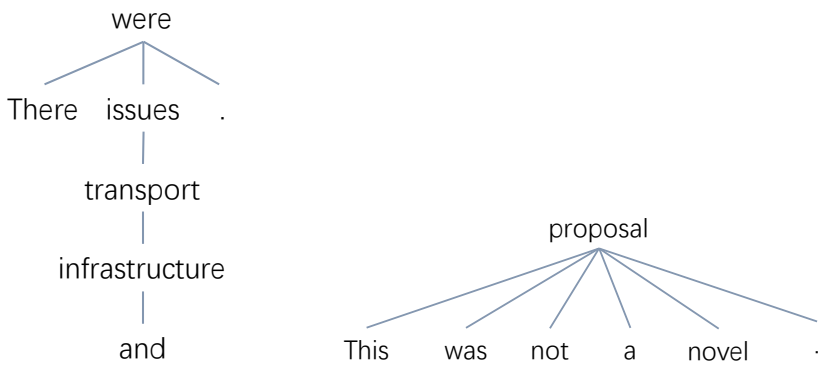


Figure 15: Metrics showing the same diachronic trends based on MK between English (blue) and German (orange), averaged per decade group and over the 5 parsers. We show 95% confidence intervals.

a decrease in both *treeDegree* and *degreeVar* for longer sentences of lengths from 40-42 to 60-62 words. All of the above suggests that generally, for both English and German, the dependency trees of shorter sentences with ≤10-15 words become shorter and wider over time, while those of longer sentences having ≥30 words become taller and narrower over time. Short and wider tree structures have a higher chance that a word governs multiple dependents compared to the taller and narrower ones and vice versa. An example for comparing between the shorter and wider vs. the taller and narrower dependency trees is given in Figure 16, where both sentences have 7 tokens.

The trends of $\boldsymbol{d_{head\text{-}final}}$ and $\boldsymbol{Ratio_{head\text{-}final}}$ are illustrated in Figure 15g and 15i respectively. A larger $d_{head\text{-}final}$ of a sentence indicates greater distance from the corresponding fully head-final sentence (where all dependency pairs of that sentence are head-final); thus, it suggests that the



(a) Taller and narrower: *Height*=5, *treeDegree*=2.

(b) Shorter and wider: *Height*=2, *treeDegree*=5.

Figure 16: Dependency trees for sentences of length 7: taller and narrower vs. shorter and wider. Predicted by Stanza. Both sentences taken from the Hansard corpus: left: 'There were transport and infrastructure issues.'; right: 'This was not a novel proposal.'

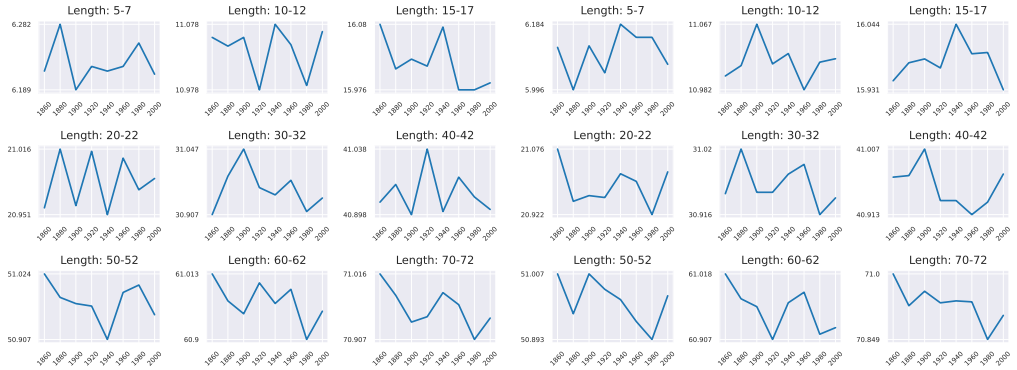
sentence is more head-initial. $d_{\textit{head-final}}$ is expected to rise with declined $\textit{Ratio}_{\textit{head-final}}$ to reflect language becoming more head-initial and vice versa. We see that: (1) $d_{\textit{head-final}}$ decreases for sentences with 5-7 and 15-17 words but increases for those having 70-72 words, implying that short sentences become more head-final, while long sentences become more head-initial; (2) $\textit{Ratio}_{\textit{head-final}}$ decreases over time for sentences having ≥ 30 words, also suggesting a tendency for longer sentences being more head-initial over time.

Figure 15h displays the trends of ***#Leaves*** for sentences having 5-7 words, where we see an increase over time for both languages. This suggests a tendency that more words never serve as a head and a head governs more dependents in short sentences, which is consistent with the changes in overall dependency tree structures, i.e., they become shorter and wider for short sentences.

7. Discussion

Effect of sentence length. As mentioned in §3.2, it is important to control the sentence length for diachronic syntactic observations as it may be a confounding variable.[16] Thus, we compute the diachronic trend of the average sentence length per decade group for each length group in our data, visualized in Figure 17. Even if the length difference within one group is up to 3 tokens in our data, we still often see a decrease in average sentence length over time, especially for the longer sentences with more than 50 tokens (last rows in Figure 17). However, the difference in the average sentence length within one length group is only up to ~0.1 token, which we deem as an acceptable level. The reasons for this are the following: (1) Intuitively, such a small difference in sentence length is unlikely to influence our metrics substantially. For instance, it takes at least one additional token to increase the maximal possible *Height* and *treeDegree*. (2) We observe trends of metrics that counter the development of sentence lengths, e.g., that longer sentences have a higher chance to produce taller dependency trees compared to shorter sentences; however, we see uptrends in *Height* in Figure 15a for sentences with 60-70 words, whose lengths are overall decreasing over time for both languages. This suggests that such a minor decrease in sentence

16 On the other hand, reducing sentence length over time may be one core aspect of syntactic simplification, e.g., splitting one long sentence into two shorter sentences.



(a) English (b) German

Figure 17: Average sentence length in our observation data over time. We show the maximal and minimal average lengths on the y-axes.

length does not substantially impact the observations on syntactic changes. (3) The subtrends in metrics do not often follow the subtrends in sentence length. For instance, in 6 out of 8 metrics in Figure 15 for sentences of length 5-7 (*Height*, *depthVar*, *depthMean*, *treeDegree*, *degreeVar*, $d_{\textit{head-final}}$), the corresponding trends behave consistently across both languages from the time window 1860-1870 to 1880-1890, even when German sentences become longer while English sentences become shorter during that period.

Scale of syntactic changes. We note that several of our syntactic trends, even if they are statistically significant, are small in scale. For example, *degreeMean* in Figure 14b changes from slightly below 0.984 to slightly above 0.984 in English over the past 160 years (not all trends are on such low scales, however: $\textit{Height}_{\textit{dependency}}$ decreases from slightly below 14 to slightly above 13 in the same time period in English (Figure 14c), for example). On the one hand, this makes sense as syntax is expected to change much slower, e.g., compared to lexical or semantic changes. On the other hand, this may also mean that some of our identified changes may hardly be noticeable (by linguist experts) in actual language or will more pronouncedly manifest themselves only in the future.

About DDM. Although some previous studies have shown that the *mDD* is decreasing over time, in this work, we do not see direct evidence supporting this. In fact, *mDD* for German is increasing over time across 7 out of 9 lengths in our experiment, while it is only decreasing for English sentences having 70-72 tokens (see Table 6 in the appendix). Overall, our observations on *mDD* are more in line with Krielke (2023) and Zhu, Liu, and Pang (2022), who base observations on the sentences of identical lengths instead of grouping sentences by a range of lengths (e.g., Lei and Wen (2020)). The former mostly find no trend or uptrend in *mDD* for 'general' text domains (vs. scientific language) including news, magazines etc. Further, some syntactic structures have been argued to appear under the pressure of DDM. E.g., Liu, Xu, and Liang (2017) demonstrate that dependency trees with more hierarchical depth and fewer parallel words are preferred since they are in a better position to produce shorter dependency distances; this is in accordance with our observations where the trees are becoming taller and narrower for longer sentences. On the other hand, head-initial dependencies are expected to produce shorter dependency distances than the head-final ones (Liu, Xu, and Liang 2017); we do observe that longer sentences are becoming more head-initial as well. Interestingly, the discrepancy in the trends between shorter and longer sentences also seems to 'support' the anti-DDM phenomenon in short sentences

(Ferrer-i Cancho and Gómez-Rodríguez 2021). The above facts prompt an intriguing question on what is the causal relationship between these observed syntactic changes and DDM—we defer this investigation to future research.

8. Conclusion

This study delved into diachronic language changes in English and German. We comprehensively examined 15 metrics, including the mean dependency distance, across extensive corpora containing political debates. By analyzing the agreement among five dependency parsers, including the commonly used Stanford CoreNLP, we provided novel insights into the importance of used parsers in such language change research. Notably, we discovered a distinctive uptrend in the mean dependency distance in German, which is consistent across various parsers and sentence lengths. We also found that German and English seemed undergo similar syntactic changes in the two political corpora examined, with only few instances showing opposing trends. We also observed more syntactic diachronic variability in sentences of extreme lengths, e.g., below or above mean sentence lengths. Our results may be relevant in downstream tasks, e.g., for text generation systems that map across historic time frames, e.g., diachronic summarizers or MT systems (Zhang, Ouni, and Eger 2023; Thai et al. 2022).

Possible limitations of this work are the following: (1) The text genre explored was restricted to political debates, in order to make the subcorpora comparable, but this may be an important factor affecting the measures. While there are studies showing a trend towards colloquialization in the parliamentary debates of some languages (Hiltunen, Räikkönen, and Tyrkkö 2020; Hou and Smith 2021), this has in fact not yet been studied for German. It is possible that one factor in the increase of dependency distance in the present study was in fact one such colloquialization trend, namely that an earlier more nominal style (typical of German written registers, especially legalese, replacing finite clauses with nominal constructions) gave way to a more natural/colloquial 'verbal' style with more finite clauses. This can be illustrated with the example in (1). (1a) is an attested example from the corpus. Such complex nominal structures might over time have been replaced by finite clauses as in the constructed example in (1b).

(1) a. die Beschlüsse, welche das Haus bei §55 **durch die Annahme der Kommissionsvorschläge** gefaßt hat. (1876-11-21)
'the resolutions, which the house has taken regarding §55 through adoption of the proposals by the committee'

b. die Beschlüsse, die das Haus bei § 55 gefasst hat, **als es die Kommissionsvorschläge angenommen hat**.
'the resolutions, which the house has taken regarding §55 when it adopted the proposals by the committee "

We will look at this possibility more closely in a future study. (2) The time depth of the used diachronic corpora only covers the last 200 years—which may be too short to draw conclusions about syntactic change, which tends to take several centuries to unfold (e.g. Kroch 2001). (3) We did not thoroughly investigate the origins of the changes in measures. For instance, Juzek, Krielke, and Teich (2020) indicate that the downtrend in dependency distance over time is because dependency relations with shorter distances appear more frequently in the later periods and vice versa, while Liu, Zhu, and Lei (2022) show that longer sentences have more dependency relations with decreasing dependency distances compared to the short ones, which results in the discrepancy in the trends for short and long sentences.

Our study has been mostly formal—investigating the role of parsers and metrics on the identification of language change in English and German—but remains preliminary regarding the detection of actual syntactic change. Future work should complement our analysis in this respect.

References

Anderson, Stephen R. 2015. Morphological change. *The Routledge handbook of historical linguistics*, pages 264–285.

Bamler, Robert and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, PMLR.

Bamman, David, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Association for Computational Linguistics, Baltimore, Maryland.

Banks, David. 2008. *The development of scientific writing. Linguistic features and historical context*. Equinox.

Beese, Dominik, Ole Pütz, and Steffen Eger. 2022. FairGer: Using NLP to measure support for women and migrants in 155 years of german parliamentary debates.

Biber, Douglas and Bethany Gray. 2011. The historical shift of scientific academic prose in english towards less explicit styles of expression. *Researching specialized languages*, 47(11).

Biber, Douglas and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Bybee, Joan L. 1985. *Morphology*. Benjamins, Amsterdam/Philadelphia.

Ferrer-i Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 70 5 Pt 2:056135.

Ferrer-i Cancho, Ramon and Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.

Ferrer-i Cancho, Ramon and Carlos Gómez-Rodríguez. 2021. Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ferrer-i Cancho, Ramon, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. 2022. Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308.

Carroll, Ryan, Ragnar Svare, and Joseph C Salmons. 2012. Quantifying the evolutionary dynamics of german verbs. *Journal of Historical Linguistics*, 2(2):153–172.

Chen, Maximillian, Caitlyn Chen, Xiao Yu, and Zhou Yu. 2023. FastKASSIM: A fast tree kernel-based syntactic similarity metric. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–231, Association for Computational Linguistics, Dubrovnik, Croatia.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Degaetano-Ortlieb, Stefania and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific english. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Di Carlo, Valerio, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, AAAI Press.

Dozat, Timothy and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Association for Computational Linguistics, Florence, Italy.

Eger, Steffen and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Association for Computational Linguistics, Berlin, Germany.

Fischer, Stefan, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The royal society corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, European Language Resources Association, Marseille, France.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Frermann, Lea and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gibson, Edward, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407.

Gildea, Daniel and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Association for Computational Linguistics, Online.

Glavaš, Goran and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Association for Computational Linguistics, Online.

Gómez-Rodríguez, Carlos and Ramon Ferrer-i Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Phys. Rev. E*, 96:062304.

Gulordava, Kristina and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala University, Uppsala, Sweden, Uppsala, Sweden.

Haider, Thomas and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of New High German poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222, Association for Computational Linguistics, Florence, Italy.

Halliday, Michael AK. 2003. On the language of physical science. In *Writing science*. Routledge, pages 59–75.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Association for Computational Linguistics, Berlin, Germany.

Hiltunen, T., J. Räikkönen, and J. Tyrkkö. 2020. Investigating colloquialization in the british parliamentary record in the late 19th and early 20th century. *Language sciences*, 79:1–16.

Honnibal, Matthew and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hou, Liwen and David A. Smith. 2021. Drivers of English syntactic change in the Canadian parliament. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 51–60, Association for Computational Linguistics, Online.

Juzek, Tom S, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific english. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119.

Knooihuizen, Remco and Oscar Strik. 2014. Relative productivity potentials of dutch verbal inflection patterns. *Folia Linguistica*, 48(Historica-vol-35):173–200.

Krielke, Marie-Pauline. 2021. Relativizers as markers of grammatical complexity: A diachronic, cross-register study of english and german. *Bergen Language and Linguistics Studies*, 11(1):91–120.

Krielke, Marie-Pauline. 2023. Optimizing scientific communication: the role of relative clauses as markers of complexity in english and german scientific writing between 1650 and 1900.

Krielke, Marie-Pauline, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. Tracing syntactic change in the scientific genre: Two Universal Dependency-parsed diachronic corpora of scientific English and German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4808–4816, European Language Resources Association, Marseille, France.

Kroch, Anthony. 2001. Syntactic Change. In Mark Baltin and Chris Collins, editors, *The Handbook of Contemporary Syntactic Theory*. Blackwell, Malden, pages 629–739.

Lei, Lei and Matthew L Jockers. 2020. Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.

Lei, Lei and Ju Wen. 2020. Is dependency distance experiencing a process of minimization? a diachronic study based on the state of the union addresses. *Lingua*, 239:102762.

Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716.

Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics*, 15:1–12.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based

on dependency treebanks. *Lingua*, 120(6):1567–1578. Contrast as an information-structural notion in grammar.

Liu, Haitao, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Liu, Xueying, Haoran Zhu, and Lei Lei. 2022. Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types. *Humanities and Social Sciences Communications*, 9(1):1–9.

Ma, Xianghe, Michael Strube, and Wei Zhao. 2024. Graph-based clustering for detecting semantic change across time and languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics.

Ma, Xuezhe, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Association for Computational Linguistics, Melbourne, Australia.

Mann, Henry B. 1945. Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, pages 245–259.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Association for Computational Linguistics, Baltimore, Maryland.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Association for Computational Linguistics, Baltimore, Maryland.

Nübling, Damaris. 1998. Warum werden bestimmte Verben regelmäßig unregelmäßig? Prinzipien und Funktionen der Irregularisierung. *Germanisten*, 3:28–37.

Moscoso del Prado Martín, Fermín and Christian Brendel. 2016. Case and cause in Icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2430, Association for Computational Linguistics, Berlin, Germany.

Qi, Peng, 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, Association for Computational Linguistics, Online.

Rodda, Martina A, Marco SG Senaldi, and Alessandro Lenci. 2017. Panta rei: Tracking semantic change with distributional semantics in ancient greek. *IJCoL. Italian Journal of Computational Linguistics*, 3(3-1):11–24.

Rother, David, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, International Committee for Computational Linguistics, Barcelona (online).

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, 73:161–183.

Simpson, Edward H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Association for Computational Linguistics, Brussels, Belgium.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Temperley, David. 2007. Minimization of dependency length in written english. *Cognition*, 105(2):300–333.

Thai, Katherine, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Conference on Empirical Methods in Natural Language Processing*.

Tily, Harry. 2010. *The role of processing complexity in word order variation and change*. Stanford University.

Walter, T., C. Kirschner, S. Eger, G. Glavas, A. Lauscher, and S. Ponzetto. 2021. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–60, IEEE Computer Society, Los Alamitos, CA, USA.

Weerman, Fred and Petra De Wit. 1999. The decline of the genitive in dutch. *Linguistics*, 37(6):1155–1192.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Association for Computational Linguistics, Brussels, Belgium.

Zhang, Kaizhong and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18:1245–1262.

Zhang, Ran, Jihed Ouni, and Steffen Eger. 2023. Cross-lingual cross-temporal summarization: Dataset, models, evaluation. *ArXiv*, abs/2306.12916.

Zhang, Ruoyang and Guijun Zhou. 2023. An investigation of the diachronic trend of dependency distance minimization in magazines and news. *Plos one*, 18(1):e0279836.

Zhang, Yu, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Association for Computational Linguistics, Online.

Zhu, Haoran and Lei Lei. 2018. Is modern english becoming less inflectionally diversified? evidence from entropy-based algorithm. *Lingua*, 216:10–27.

Zhu, Haoran, Xueying Liu, and Nana Pang. 2022. Investigating diachronic change in dependency distance of modern english: A genre-specific perspective. *Lingua*, 272:103307.

9. Appendix

9.1 Parser Configurations

We train model checkpoints for StackPointer, Biaffine, and CRF2O on the concatenated train set of the treebanks in UD2.12, separately for each language.[17] The authors of TowerParse released the checkpoints trained on individual treebanks in UD2.5; as they did not provide the training script, we directly take the models trained on the largest treebank for each language, i.e., EWT for English and HDT for German.[18] For CoreNLP and Stanza, we use them with the out-of-the-box settings.[19] Note that the default models in Stanza 1.5.1 were also trained using UD2.12, however, the default German model is only trained on the GSD treebank, and the default English model is trained on the combination of 5 out of 7 train sets in the treebanks.

Except for Stackpointer, which does not have a built-in tokenizer, and CoreNLP, which has its own tokenizer, the other parsers implement the tokenizer from Stanza by default. Thus, for ease of use, we implement the Stanza tokenizer for every parser consistently, disabling sentence segmentation, as we will apply them to the curated sentences from our target corpora, for which further sentence splitting is unexpected.

9.2 Comparison to Zeman et al. (2018)

As Figure 20 shows, the great majority of the parsers obtain higher LAS than the average LAS over parsers in Zeman et al. (2018), except for CRF2O on German, whose LAS is slightly lower than the average. Stanza, Towerparse, and Biaffine consistently outperform the best systems across different languages. Thus, we confirm the legitimacy of the used parsers.

9.3 Sensitivity of metrics to data noise

While our results in §4 indicate that the parsers perform decently on the target corpora, it remains uncertain whether the metrics derived from their parsing results are affected by data noise and to what extent. Since we will inspect the trends of the metrics over time, it is important to ensure that the data noise does not disrupt the rankings of those metrics. Hence, we compute Spearman's ρ correlations between our 15 metrics on the parses of the original and human corrected sentences (without spelling and OCR errors, etc.; collected in Section 3); higher correlations indicate the rankings are less disrupted and vice versa.

Overall, as Figure 5 illustrates, the metrics for the original and corrected sentences strongly correlate with each other, shown with an overall correlation of 0.926. This overall correlation is computed by first averaging over parsers for each metric and then averaging over all metrics. The lowest average correlation per metric is observed with #crossing, 0.869, which, however, still implies a strong correlation. The least sensitive metric is the head-final ratio, shown with an average correlation of 0.954 over parsers. Among the parsers, Stanza is least sensitive (0.983), followed by Biaffine (0.979), whereas TowerParse and StackPointer are the most sensitive ones (0.857 and 0.879 respectively). For this reason, we conclude that our metrics are not much sensitive to data noise in the target corpora across different parsers.

17 We use the implementation of Biaffine and CRF2O from https://github.com/yzhangcs/parser, StackPointer from https://github.com/XuezheMax/NeuroNLP2.

18 https://github.com/codogogo/towerparse

19 We use Stanza of version 1.5.1 and CoreNLP via the client API bundled in Stanza.

Metrics	Stanza	CoreNLP	StackPointer	Biaffine	TowerParse	AVG
mDD	0.976	0.905	0.823	0.966	0.888	0.912
nDD	0.995	0.892	0.862	0.985	0.960	0.939
Height	0.983	0.915	0.862	0.986	0.820	0.913
$Ratio_{head\text{-}final}$	0.975	0.954	0.959	0.972	0.909	**0.954**
treeDegree	0.968	0.916	0.870	0.973	0.815	0.908
#Leaves	0.998	0.999	0.908	0.997	0.832	0.947
degreeVar	0.966	0.935	0.923	0.976	0.834	0.927
degreeMean	0.998	0.988	0.902	0.998	0.832	0.944
depthVar	0.982	0.910	0.861	0.976	0.827	0.911
depthMean	0.987	0.918	0.874	0.980	0.818	0.915
$d_{head\text{-}final}$	0.989	0.967	0.891	0.990	0.841	0.936
#Crossing	0.942	-	0.815	0.938	0.782	<u>0.869</u>
$Height_{dependency}$	0.995	0.935	0.874	0.994	0.884	0.936
d_{root}	0.999	0.870	0.873	0.965	0.976	0.937
$d_{randomTree}$	0.991	0.991	0.894	0.986	0.830	0.938
AVG	**0.983**	0.935	0.879	0.979	<u>0.857</u>	0.926

Table 5: Spearman's ρ between the metrics on the original sentences and that on the human-corrected sentences (collected in Section 3). The "AVG" values in the rows refer to the average correlation over different parsers for a specific metric; those in the columns denote the average correlation over different metrics for a specific parser. We bold the highest and underline the lowest average correlations.

1. "is_sent": is the text a complete sentence or not? Please enter/select either "TRUE" or "FALSE" in the corresponding cell. If the answer is "FALSE", skip the following steps.

2. "has_errors": does the sentence have any issues/errors that make it differ from a fully correct and modern English sentence? Please enter/select either "TRUE" or "FALSE" in the corresponding cell. If the answer is "FALSE", skip the following steps.

3. "errors": which errors/issues does the sentence have? You can choose multiple errors from the provided selections and can also choose a specific error multiple times. The selections are:
   a. ***Spelling***: incorrect or historical/out-dated spelling of words
   b. ***Space***: incorrect or historical/out-dated use of spaces.
   c. ***Punctuation***: incorrect or historical/out-dated use of punctuations.
   d. ***Symbol***: incorrect or historical/out-dated use of symbols. For example, "welche für den Fall der Annahme des **Z** 33b in demselben die Worte", where **"Z"** should be **"§"**.
   e. ***Extra material***: any parts that are not grammatically connected to other parts of a sentence or additional parts from the OCR errors. E.g.,
      i. Ob der Weg, den der **(v)** Herr Abgeordnete Möller eben angedeutet hat, der richtige ist, ist mir allerdings zweifelhaft.
      ii. Um endlich eine Klarheit zu gewinnen, wie lange denn das Vorrücken der deutschen Heere nach Ratifikation des Friedens von Brest-Litowsk noch vor sich gehen solle, **(Zuruf bei den Unabhängigen Sozialdemokraten) (L)** hat die Sowjet-Regierung durch Funkspruch in dieser Weise beim Auswärtigen Amt angefragt.
      iii. Und wenige Tage später lesen wir im Heeresbericht: Die deutschen Truppen sind in die Krim einmarschiert, **(hört! hört! bei den Unabhängigen Sozialdemokraten)** und abermals nach wenigen Tagen: Die deutschen Truppen haben Sebastopol besetzt, und nach einigen Wochen hüll der Kaiser in Aachen vor den dortigen Stadtverordneten eine Rede, in der sich wörtlich folgende Wendung fand: In der Krim geht es gut vorwärts.
   f. ***Missing material:*** any parts that are missing compared to the original PDF file. E.g., "...auch in diesem Hohen Hause solche **»Pensionsgewinnlerantreffen** zu können.", where the "«" after "Pensionsgewinnlerantreffen" was missing due to the OCR error.
   g. ***Other***: If the issue is not listed above. Please explain it in the "comments" column.

4. "origin_of_errors": Where are the errors coming from? You should choose an origin for every item listed in the "error" column. The selections are:
   a. ***OCR***
   b. ***Historic***: historical/out-dated spelling of words, usage of punctuation/symbols…
   c. ***Genre***: The issues exist due to the genre of the texts. For example, political texts may switch between written texts and speeches, so there might be some interjections in a sentence.
   d. ***Unknown***: please select this if the origin is not explicit to you.
   e. ***Other***: if it is not listed above. Please explain it in the "comments" column.

5. "Correction": correct the sentence if any issues exist.

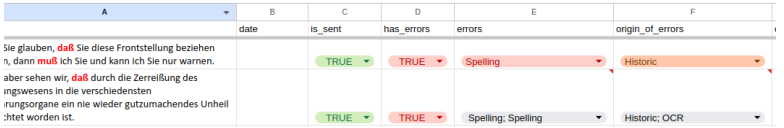
6. NOTE:

a. Please select a specific error **only once** if its origins are the same, even when it occurs multiple times. E.g., for sentence: "Wenn *Sie glauben, **daß** Sie diese Frontstellung beziehen müssen, dann **muß** ich Sie und kann ich Sie nur warnen.*", you should choose "spelling" only once.

b. Please select a specific error **multiple times** when its origins are NOT the same. E.g., for sentence "***Hellte** aber sehen wir, **daß** durch die Zerreißung des Wohnungswesens in die verschiedensten Ausführungsorgane ein nie wieder gutzumachendes Unheil angerichtet worden ist.*", you should choose "spelling" twice.

c. The order of the errors in the column "errors" can be random — it can be irrelevant to the order of the errors appearing in the sentence.

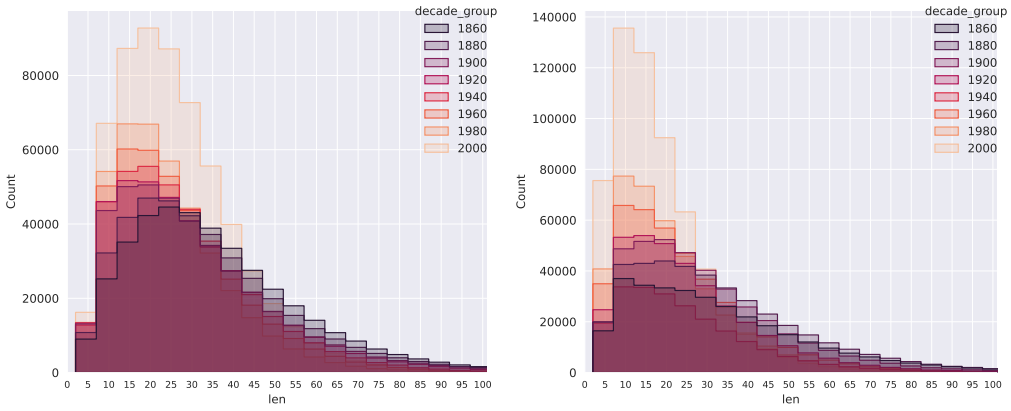
d. **BUT,** the items listed in the 'origin_of_errors' column should correspond to the items in the 'errors' column in a **one-to-one manner**, maintaining the exact same order; the number of items in the two columns should be the same. E.g.:



A	B	C	D	E	F
	date	is_sent	has_errors	errors	origin_of_errors
Sie glauben, daß Sie diese Frontstellung beziehen n, dann muß ich Sie und kann ich Sie nur warnen.		TRUE	TRUE	Spelling	Historic
aber sehen wir, daß durch die Zerreißung des ungswesens in die verschiedensten rungsorgane ein nie wieder gutzumachendes Unheil chtet worden ist.		TRUE	TRUE	Spelling: Spelling	Historic; OCR

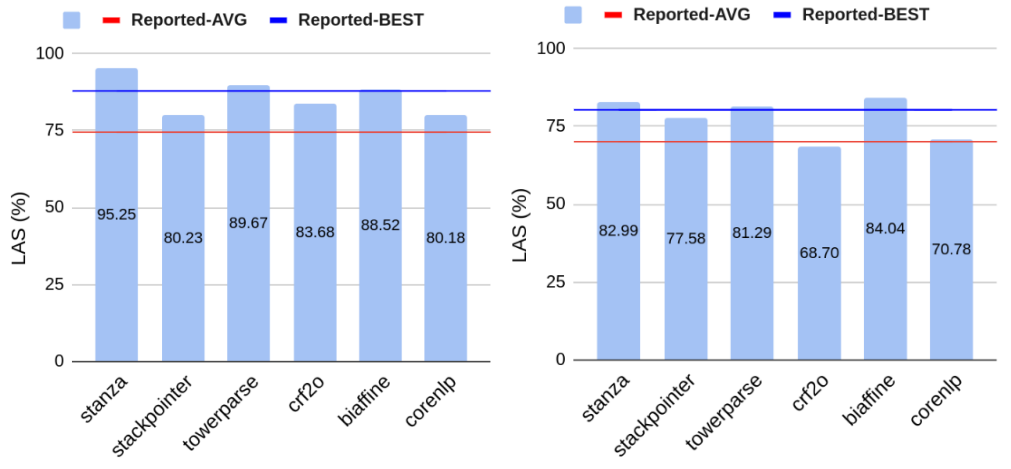
e. There might be a slight delay after you choose the selections, as additional script was added to the spreadsheet to enable multiple choices :). You can also type them yourself in the form of "***[item1]" + ";" + [space] + [item2]+...***". Please ignore the warning after the multiple choices, which raises because google spreadsheet originally doesn't allow that.

Figure 18: Annotation Guidelines for Hansard.



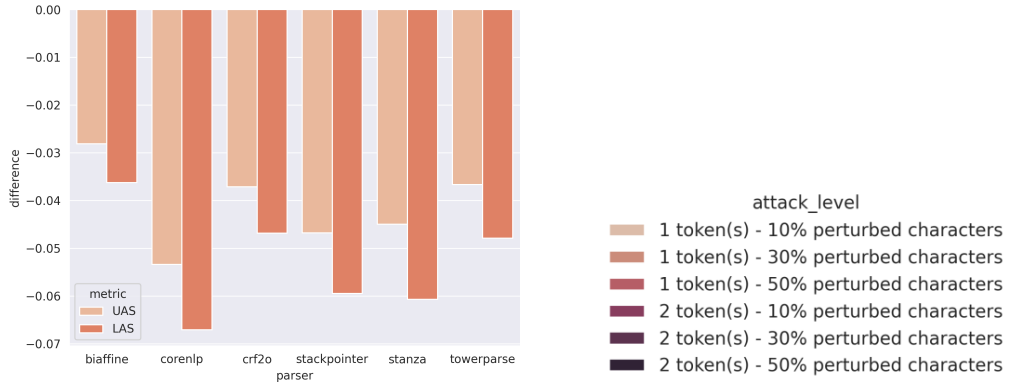
(a) Hansard (b) DeuParl

Figure 19: Histogram of sentence length distribution with bin size of 5 for the randomly sampled sentences having at most 100 tokens (which compose 99% of the data). Darker colors indicate older periods and vice versa.

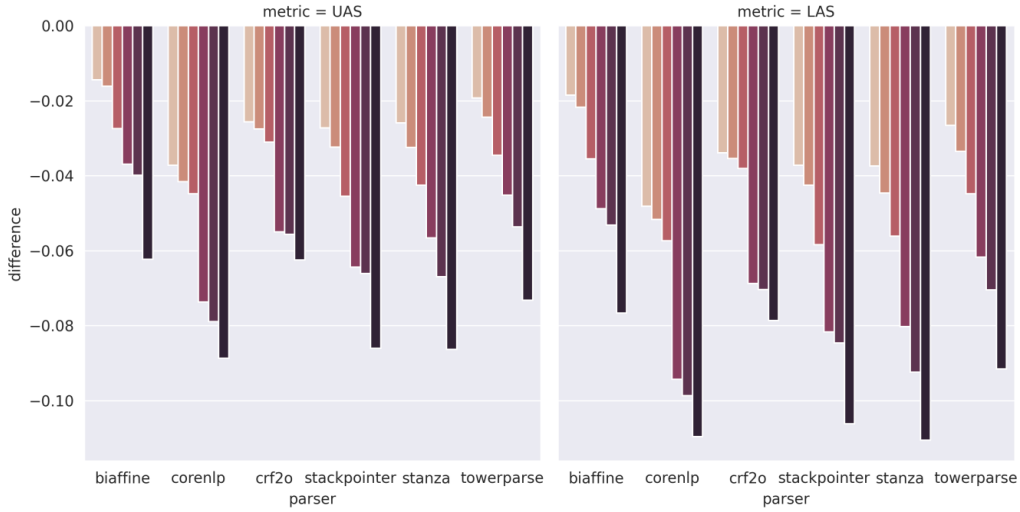


(a) EN-PUD (b) DE-GSD

Figure 20: Barplot of LAS of the parsers in our evaluation in comparison to the average and the best LAS over various parsers on the English PUD and German GSD test sets reported in Table 15 of Zeman et al. (2018).



(a) Historical Spelling Attack



(b) OCR Spelling Attack

Figure 21: Absolute difference in UAS and LAS between the sentences before and after the attack.

Metric	5-7	10-12	15-17	20-22	30-32	40-42	50-52	60-62	70-72
mDD	-	-	-	-	-	-	-	-	-3
nDD	+4	-	-	-	+4	-	-	-	-
d_{root}	+3	-	-	-	-	-	-	-	-
Height	-5	-	-	+3	-	+4	-	+5	+5
depthVar	-5	-	-	+3	-	+4	+3	+5	+5
depthMean	-5	-3	-	-	-	+3	+3	+5	+5
treeDegree	+5	-	-	-	-5	-3	-5	-5	-4
degreeVar	+5	-	+5	-	-	-3	-3	-4	-4
degreeMean	+5	-	-	-	-	-	-	-	+5
$Ratio_{head\text{-}final}$	+5	-	-	-3	-3	-	-4	-4	-5
$d_{head\text{-}final}$	-5	-	-3	-	-	-	-	-	+3
#Leaves	+5	-	-	-	-	-	-	-	-
#Crossings	-	-	-	-	-	-	-	-	-
$Height_{dependency}$	-	-3	-	-	-	-	-	-	-
$d_{randomTree}$	-	-	-3	-	-	-	-	-	-

(a) English

Measure	5-7	10-12	15-17	20-22	30-32	40-42	50-52	60-62	70-72
mDD	+5	+5	+5	+4	-	+5	+5	+4	-
nDD	-5	-5	-	-3	-5	-	-	-	-
d_{root}	-	-	-	-	-4	-	-	-	-
Height	-5	-	-	-	+4	-	-	+3	+3
depthVar	-5	-	-	-	+3	-	-	+3	+3
depthMean	-5	-3	-	-	+4	+5	+4	+3	+3
treeDegree	+5	+5	-	-4	-4	-5	-5	-5	-5
degreeVar	+5	+5	-	-4	-4	-5	-5	-5	-5
degreeMean	+4	+5	-	-	-	-3	-3	-4	-4
$Ratio_{head\text{-}final}$	-	-	-	-	-4	-5	-5	-5	-5
$d_{head\text{-}final}$	-5	-5	-3	-	+4	+4	+5	+5	+4
#Leaves	+5	+5	-	-	-4	-5	-5	-5	-5
#Crossings	-	-3	-3	-4	-	-	-	-	-
$Height_{dependency}$	+5	+5	+3	-	+4	+5	+5	+4	+3
$d_{randomTree}$	-	-	-	-	-	-	-	-	-

(b) German

Table 6: Metrics that have a stable trend supported by at least 3 parsers (except for *#Crossings*, for which we relax the minimal requirement to 2 parsers, as CoreNLP is unable to predict crossing edges.). "+" and "−" denote the increasing and decreasing trend respectively; the values refer to the number of parsers that support this trend.