Text Complexity in the Digital Humanities — A Case Study on 18th Century Periodicals

Han Tran, Bernhard C. Geiger
Know-Center GmbH
Graz, Austria
geiger@ieee.org

Yvonne Völkl, Christina Glatz
Institute for Romance Studies, University of Graz
Graz, Austria

Martina Scholger, Sanja Saric
Centre for Information Modelling, University of Graz
Graz, Austria

Philipp Koncar, Roman Kern
Institute of Interactive Systems and Data Science, Graz University of Technology
Graz, Austria

ABSTRACT

Historical texts are an important source to understand historical events and societal developments. In traditional literary studies, such documents are analyzed using close reading, which is tedious and time-consuming for large text corpora. Thus, computational literary scholars have adopted and developed automated and quantitative methods that complement and contribute to insights of close reading. In this case study, we show that text complexity measures are meaningful additions to these methods. Specifically, we apply interpretable measures of reading ease and of syntactic and lexical richness to historical texts and show that the obtained quantitative results are consistent with findings from close reading.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Applied computing** → **Arts and humanities**.

KEYWORDS

text complexity, digital literary studies

ACM Reference Format:
Han Tran, Bernhard C. Geiger, Yvonne Völkl, Christina Glatz, Martina Scholger, Sanja Saric, and Philipp Koncar, Roman Kern. 2022. Text Complexity in the Digital Humanities — A Case Study on 18th Century Periodicals. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital humanists use a set of methodologies for distant reading and use them to complement traditional literary studies of individual texts with investigations of entire corpora. Typical methodologies are topic modeling [3, 16], sentiment analysis [19], stylometry [1], and network analysis [14, 21]. The selection of methods and statistical quantities is guided by the nature of the historical text corpus under investigation, which is often multi-lingual and contains inconsistent or outdated orthography. Thus, the methods and quantities of choice are characterized by robustness and interpretability, rather than by their performance on (modern) text benchmarks.

In this case study, we investigate the analysis of text complexity as yet another possible methodology for distant reading. Text complexity measures have been used to date works [20], for authorship attribution [17], and for the analysis of pulp fiction [10, 11]. We selected and investigated three statistical quantities that are simple to compute and interpret: (1) the Flesch reading ease; (2) syntactic complexity, measured by the normalized perplexity of the sequence of part-of-speech (PoS) tags; and (3) lexical richness, measured by a normalized type-token-ratio (TTR). We apply these quantities to a multi-lingual corpus of periodicals of the journalistic genre of the "Spectators" [5] and test hypotheses regarding the complexity of translations, the impact of topical focus, and how text complexity of individual authors changed over time. These hypotheses are motivated from literary studies and targeted at individual periodicals; as such, one cannot expect that the observed phenomena (e.g., that some author's texts become more complex over time) generalize to other authors or periodicals, let alone to other journalistic genres. Within the narrow scope of the hypotheses, however, the obtained quantitative results are in line with insights from literary studies. This suggests that the considered quantities are valid additions to the panoply of digital humanities methodologies.

2 CORPUS DESCRIPTION

Table 1: Summary of the used subset of the corpus.

Language	Periodicals	Issues	Words	Unique Words
French	25	1,735	4,052,733	87,686
Italian	13	1,348	3,254,453	125,807
Spanish	23	688	2,037,764	89,782

The "Spectators" originated in 18th century England and quickly spread and carried ideas of the Enlightenment all over Europe. These essayistic periodicals feature a variety of narrative forms such as letters, tales, examples, metaphorical frames, etc. [4, §20] and ask questions concerning morale, virtue, as well as changes in society. They comment on literature, art, culture, history, economics, religion, and other aspects of private and public life.

Our data is part of a digital edition [5] containing roughly 4,000 TEI-annotated[1] issues in six languages. We focus on the corpora

Conference'17, July 2017, Washington, DC, USA
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/P5/

of French, Italian, and Spanish texts, which contain 61 periodicals, more than 3,500 issues, and 9 million words (cf. Table 1). Our hypotheses in Section 4 focus on selected periodicals of these corpora.

3 COMPUTATION OF TEXT COMPLEXITY

Flesch Reading Ease. The Flesch reading ease [8] is computed from a document's average sentence length and the average length of words in syllables, i.e.,

$$\mathsf{FRE} = C_1 - C_2 \frac{N}{S} - C_3 \frac{Syl}{N} \tag{1}$$

where S, N, and Syl are the numbers of sentences, words, and syllables in the document, and where C_1 through C_3 are language-dependent constants. High values result from short sentences and/or short words, and thus indicate higher readability.

We computed the Flesch reading ease using the `textstat` library, which implements the formula with constants for French [12], Spanish [6], and Italian [9]. We removed all sentences with a length less than four words using `nltk`, as these most probably account for sequences of abbreviations or sentence fragments in our corpus.

PoS Information Rate. [23] proposed the PoS information rate (PoS-IR) as a measure of syntactic complexity. If p_i is the PoS tag of the i-th token in a document with N^+ tokens, then the PoS-IR is computed as the ratio between the PoS trigram perplexity

$$2^{-\frac{1}{N^+-2}\sum_{i=3}^{N^+}\log_2 P(p_i|p_{i-1},p_{i-2})} \tag{2a}$$

and the PoS unigram perplexity

$$2^{-\frac{1}{N^+}\sum_{i=1}^{N^+}\log_2 P(p_i)} \tag{2b}$$

where $P(p_i)$ and $P(p_i|p_{i-1},p_{i-2})$ are the empirical probabilities (within the considered periodical) of PoS tag p_i and of PoS tag p_i following a sequence of PoS tags p_{i-1} and p_{i-2}. This ratio can be shown to be an estimate of $2^{-I(P_i;P_{i-1},P_{i-2})}$. Easily predictable sequences of PoS tags have high $I(P_i;P_{i-1},P_{i-2})$ and low PoS-IR. Thus, high PoS-IR indicates a rich syntactic structure.

We used `TreeTagger`[2] to obtain sequences of PoS tags. To mitigate finite sample effects, we reduced the number of PoS tags via grouping as indicted in Table 2 (e.g., we grouped all declensions of verbs). Instead of 33, 39, and 67 tags we thus only considered 15, 18, and 20 tag groups for French, Italian, and Spanish, respectively.

Normalized Type-Token-Ratio ΔTTR.. The TTR of a document with N words is the ratio N_u/N, where N_u is the number of unique words in the document. Heaps' law states that the TTR depends on the document length via $N_u \approx kN^\alpha$, where k and $\alpha < 1$ depend on the corpus. Several variations have been proposed in the literature that remove the dependence on N, e.g., by computing the TTR on text segments of fixed length [15]. We normalized the TTR computed on a document by the TTR expected for a document of this size, where this expectation is computed using Heaps' law. Positive values indicate a small vocabulary compared to the corpus.

Specifically, we computed the expected TTR of an issue of length N as follows: In a first step, we determined the parameters $k^\bullet$ and $\alpha^\bullet$ by minimizing the sum of $|N_u - kN^\alpha|$, summed over all issues of a given language. Then, for a concrete issue of length N, the normalized TTR is defined as $\Delta\mathrm{TTR} = (k^\bullet N^{\alpha^\bullet} - N_u)/N$. This definition has some advantage over approaches working with test segments of fixed length, as they underestimate the lexical richness if different segments have different vocabulary (e.g., due to a topic shift). Our normalized TTR considers the document as a whole, but requires a larger corpus for computation.

[2] https://www.cis.lmu.de/~schmid/tools/TreeTagger/

Table 2: Groups of PoS tags considered for the computation of the PoS information rate.

Language	Grouped	Original
French	DET	Tags begin with DET
	VER	Tags begin with VER
	PRO	Tags begin with PRO
	PRP	Tags begin with PRP
	PUN	Tags begin with PUN
Italian	DET	Tags begin with DET
	PRO	Tags begin with PRO
	PRE	Tags begin with PRE
	VER	Tags begin with VER
Spanish	CONJ	'CC', 'CCAD', 'CCNEG', 'CQUE', 'CSUBF', 'CSUBI', 'CSUBX'
	NOUN	'NC', 'NMEA', 'NP', 'NMON'
	PRO	'DM', 'INT', 'PPC', 'PPO', 'PPX', 'REL'
	PUN	'CM', 'COLON', 'DOTS', 'SEMICOLON'
	PAL	'PAL', 'PDEL'
	SYM	'QT', 'BACKSLASH', 'SLASH' 'DASH', 'ALFS', 'CODE', 'FO', 'NEG', 'PERCT', 'UMMX', 'RP', 'LP'
	VER	Tags begin with V

4 HYPOTHESES

Our first hypothesis was generated during exploratory data analysis of the Italian corpus. It appeared that *Gazzetta urbana veneta* (Antonio Piazza, 1787-1798) has a comparatively large vocabulary and, at the same time, short sentences. In contrast, we noticed longer sentences in *La Spettatrice* (1752), the anonymous Italian translation of Eliza Haywood's *The Female Spectator* (1744-1746). These two periodicals thus appear as *outliers*, and we want to determine whether the observed differences are statistically significant.

Second, we perform a longitudinal study of three French periodicals of Jean-François de Bastide, which were published subsequently between 1758-1761. Our aim is to determine whether the text complexity of this author changed over time.

Third, many of the periodicals in our corpus are translations and adaptations of previously published periodicals. Of particular interest is thus the scenario where an author of periodicals also acted as translator of a different periodical, and the question whether the text complexities of translations and original works. To this end, we investigate the French periodicals of Justus van Effen. Besides publishing three periodicals between 1711 and 1725, he also translated Addison and Steele's *The Guardian* (1713) from English to French in 1723. Our third hypothesis is that this translation has significantly different text complexity than van Effen's original works.

Finally, we hypothesize that periodicals with a strong topical focus on science and (male) education exhibit higher text complexities than periodicals focusing on images of women and men, deemed as examples worthy of imitation, as these latter periodicals should be accessible to women and men alike. To test this last hypothesis, we consider two Spanish periodicals, the anonymously published *El*

Regañón general (1803-04) for the former and Beatriz Cienfuegos' *La Pensadora Gaditana* (1763-64) for the latter category.

5 RESULTS AND INTERPRETATION

To assess the text complexity of a given periodical, we computed the quantities described in Section 3 for each issue.[3] The means and standard deviations of these quantities are provided in Table 3. If we make statements regarding whether a periodical is more or less complex than another w.r.t. a text complexity measure, we accompany these statements with the p-values of a one-tailed, two-sample Mann-Whitney-U test performed using `scipy`. We call a difference significant if the p-value is below 0.05.

Gazzetta urbana veneta vs. La Spettatrice. As expected, *Gazzetta urbana veneta* has a higher Flesch reading ease ($p \approx 5 \cdot 10^{-6}$) than *La Spettatrice*. Flesch reading ease depends strongly on sentence lengths, cf. (1), and close reading showed that *La Spettatrice* contains long sentences that are often segmented using semi-colons. In contrast, *Gazzetta* has a lively style characterized by short sentences and dialogues. Furthermore, the final sections in issues of *Gazzetta* contain notes on theatre plays and other announcements, which are often formulated as short sentences or sentence fragments. Finally, in *Gazzetta*, names are often abbreviated with a period, which may induce the sentence splitter to create shorter sentences. Surprisingly, *La Gazzetta* appears to have a more complicated grammatical structure, as suggested by a larger PoS-IR ($p \approx 0.00065$). We currently do not have an explanation for this, and future work will be needed to investigate this apparent conflict.

Our statistical analysis confirmed that *Gazzetta* has indeed a larger vocabulary ($p \approx 3 \cdot 10^{-6}$). *Gazzetta*'s broader topic coverage contributes to this rich vocabulary as much as its topical focus on theatre, literature, and arts, which introduces many named entities that are interpreted as unique words. In contrast, *La Spettatrice* has a more narrow focus and contains only few named entities.

Longitudinal Study: de Bastide. In the longitudinal study of de Bastide's periodicals, Table 3 suggests that text complexity increased along all dimensions: By using longer sentences, a larger vocabulary, and a richer grammatical structure, de Bastide trained his readers in developing a more sophisticated style and in expressing themselves more eloquently.

However, only the changes in the Flesch reading ease are significant, with the decrease between *Le Nouveau Spectateur* and *Le Monde comme il est* being smaller ($p \approx 0.014$) than the subsequent decrease to *Le Monde* ($p \approx 4 \cdot 10^{-5}$). Increases in vocabulary size ($p > 0.1$) were not significant. Regarding grammatical richness *Le Monde comme il est* has a significantly higher PoS-IR than *Le Nouveau Spectateur* at $p \approx 0.0016$.

We confirmed using both close reading and LDA-based topic modeling that the significant change in the Flesch reading ease is not confounded by a change in topical focus. Indeed, the manual annotation reveals a focus on the idea of man, the image of women, and love. Similarly, topic modeling discovers a strong focus on romantic feelings and relationships in all three periodicals. This coincides with the fact that de Bastide contributed to the propagation of the concept of tender love, which surfaced at the beginning of the second half of the 18th century [7].

Complexity of Translations: van Effen. To investigate whether the complexity of a translation differs from the complexity of an original contribution, we compared the three French periodicals of Justus van Effen (*Le Misantrope*, *La Bagatelle*, *Le Nouveau Spectateur français*) with van Effen's translation of *The Guardian*, entitled *Le Mentor moderne*. Our findings indicate that this translation has indeed higher complexity w.r.t. Flesch reading ease ($p < 3 \cdot 10^{-12}$) and lexical richness ($p \approx 0.07$ for *La Bagatelle*, $p \approx 0.004$ for *Le Misantrope*, and $p \approx 4 \cdot 10^{-9}$ for *Le Nouveau Spectateur français*). Results regarding grammatical structure are inconsistent, with *Le Mentor moderne* being syntactically more complex than *Le Nouveau Spectateur français* and *Le Misantrope* ($p < 0.035$), but less complex than *La Bagatelle* ($p \approx 2.5 \cdot 10^{-6}$).

These results are in line with previous findings. Translated texts were found to have longer sentences [22] and more subordinate clauses [13], thus potentially a more complicated grammatical structure. Furthermore, text complexity was shown to change only little during translations [2], indicating that the complexity of *The Guardian* may have had a strong impact. Finally, we assume that van Effen used a distinct vocabulary to translate aspects of the English reality of life that differ from that of the European mainland, thus contributing to *Le Mentor moderne*'s lexical richness.

Science vs. Image of (Wo)Men. We finally consider the text complexities of *El Regañón general* and *La Pensadora Gaditana*. The topical foci of these two periodicals are vastly different: While *El Regañón* mainly covers topics such as theatre, literature, arts, manners and customs, science, and education, *La Pensadora* focuses on the image of women and men and on morale. As with other periodicals, the issues depicting images of women and men were intended to provide role models for the public. Since there is such a high number of them in *La Pensadora*, we assume that this periodical has lower text complexity than *El Regañón*.

Our results reveal that, contrary to expectation, the texts in *La Pensadora* have lower Flesch reading ease ($p < 3 \cdot 10^{-11}$) and a richer vocabulary ($p \approx 0.00015$), while only their grammatical structure is simpler ($p \approx 7 \cdot 10^{-8}$). One may argue that topics of high semantic complexity were transported with texts of low complexity. Preliminary analyses towards this explanation, however, do not confirm this assumption. An explanation may lie in the writing styles of *El Regañón* and *La Pensadora*. Indeed, the former prefers a direct communication and lacks text elements such as narratives, which are more common in *La Pensadora*. In any case, the statistically significant differences between these periodicals and their nearly orthogonal coverage of topics merit closer investigation.

6 DISCUSSION AND LIMITATIONS

We investigated a set of hypotheses about 18th century Spectator periodicals using three quantities that measure orthogonal dimensions of text complexity: the Flesch reading ease as an indicator of the lengths of sentences and words; the PoS-IR as a measure of the grammatical complexity of the text; and the normalized TTR as a document length-independent measure for lexical richness. We selected these three quantities for their key strengths, namely i)

[3] Program code for these experiments is available at https://github.com/distantspectators/DiSpecs/tree/master/text_complexity.

Table 3: Text complexity measure of considered periodicals, as means and standard deviations over all issues of a periodical. Arrows indicate in which direction complexity increases.

Periodical	Author	Date	no. issues	avg. length (words)	Flesch ↓	ΔTTR ↓	PoS-IR ↑
La Spettatrice [Translation]	Anon.	1752	7	12276	40.3 (6.2)	0.038 (0.018)	0.138 (0.033)
Gazzetta urbana veneta	Piazza	1787-98	245	3888	64.7 (7.4)	-0.057 (0.021)	0.147 (0.005)
Le Nouveau Spectateur	de Bastide	1758-60	135	3354	67.2 (23.6)	0.013 (0.033)	0.156 (0.022)
Le Monde comme il est	de Bastide	1760	61	1763	67.1 (7.0)	0.006 (0.02)	0.162 (0.008)
Le Monde	de Bastide	1760-61	34	3126	58.8 (9.8)	0.005 (0.035)	0.162 (0.026)
Le Misantrope	van Effen	1711-12	95	2472	69.7 (6.5)	0.001 (0.024)	0.159 (0.012)
La Bagatelle	van Effen	1718-19	103	1444	70.0 (7.0)	-0.003 (0.021)	0.168 (0.007)
Le Nouveau Spectateur français	van Effen	1723-25	56	2278	73.6 (10.3)	0.044 (0.055)	0.16 (0.016)
Le Mentor moderne [Translation]	van Effen	1723	146	1741	62.0 (7.4)	-0.006 (0.02)	0.164 (0.008)
La Pensadora Gaditana	Cienfuegos	1763-64	52	3155	25.8 (16.4)	-0.005 (0.013)	0.148 (0.003)
El Regañón general	Anon.	1803-04	130	2993	43.8 (14.3)	0.007 (0.024)	0.152 (0.005)

their applicability and transferability to multiple languages, ii) their ease of implementation without the need of dictionaries, and iii) the fact that they are interpretable.

Our results show that significant differences can occur between consecutively published periodicals of the same author, which is in line with the findings of [20]. Further, we have evidence that the complexity measures are influenced by the topical focus, cf. [10], and the usage of certain narrative forms (short dialogues, fables, etc.). We finally found that a translated text significantly differed from other, original texts of the translator, cf. [13, 22].

Considered separately, each complexity measure has its limitations. Shallow measures, such as the average number of words per sentence, have been criticized as poor measures of readability [18]. The PoS-IR depends on the performance of the tree tagger and, due to finite sample effects, on the size of the tag sets. Finally, the normalized TTR requires the documents under consideration to be part of a larger corpus to fit Heaps' law. The computation of all complexity measures is further complicated when considering historical texts, as they often have inconsistent orthography and outdated grammar, which affects vocabulary size or confuses sentence and syllable splitters. Finally, the considered measures can only account for a small subset of text properties that affect how complex a text is perceived by the reader.

Considered as a whole, however, we have seen that these text complexity measures often paint a consistent picture, thus reinforce each other and partly mitigate their individual weaknesses. We have shown that their application to historical texts is robust due to their simplicity, and that the obtained quantitative results are in line or can be explained with existing knowledge in literary studies (with the usual risks of post-hoc theorizing). Furthermore, investigating "outliers" in terms of these text complexity measures, such as *Gazzetta urbana veneta* and *La Spettatrice*, can yield novel insights that may have so far been overlooked with close reading. Thus, we believe that analyzing text complexity, instantiated by the considered quantities, is a valid methodology for exploration and hypothesis testing in the digital literary studies.

ACKNOWLEDGMENTS

This work was supported by the Austrian Academy of Sciences within the go!digital Next Generation project "DiSpecs" (GDNG_2018-046_DiSpecs). The work by B. C. Geiger & R. Kern was additionally funded by grant P-32700-N from the Austrian Science Fund.

REFERENCES

[1] J. Burrows. 2002. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17, 3 (Sept. 2002), 267–287. https://doi.org/10.1093/llc/17.3.267

[2] A. M. Ciobanu, L. P. Dinu, and F. I. Pepelea. 2015. Readability Assessment of Translated Texts. In *Proc. Recent Advances in Natural Language Processing*. Hissar, Bulgaria, 97–103.

[3] K. Du. 2019. A Survey On LDA Topic Modeling In Digital Humanities. In *Proc. Digital Humanities (DH)*. Utrecht, Netherlands. https://dev.clariah.nl/files/dh2019/boa/0326.html

[4] K.-D. Ertler. 2012. *Moralische Wochenschriften*. Europäische Geschichte Online (EGO), http://www.ieg-ego.eu/ertlerk-2012-de (Accessed: 2 August 2021).

[5] K.-D. Ertler, A. Fuchs, M. Fischer, E. Hobisch, M. Scholger, and Y. Völkl. 2011-2021. *The 'Spectators' in the international context*. https://gams.uni-graz.at/spectators (Accessed: 2 August 2021).

[6] J. Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna* 214 (1959), 29–32.

[7] M. Fischer-Pernkopf, V. Mussner, and K.-D. Ertler. 2018. *Die «Spectators» in Frankreich. «Le Nouveau Spectateur» und «Le Monde comme il est» von Jean-François de Bastide*. Peter Lang, Frankfurt a. M., Germany.

[8] R. Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[9] V. Franchina and R. Vacca. 1986. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi* 3 (1986), 47–49.

[10] F. Jannidis, L. Konle, and P. Leinen. 2019. Markoanalytische Untersuchung von Heftromanen. In *Proc. Digital Humanities im deutschsprachigen Raum (DHd)*. Frankfurt/Mainz, Germany, 167–173. https://doi.org/10.5281/zenodo.2596094

[11] F. Jannidis, L. Konle, and P. Leinen. 2019. Thematic Complexity. In *Proc. Digital Humanities (DH)*. Utrecht, Netherlands. https://dev.clariah.nl/files/dh2019/boa/0504.html

[12] L. Kandel and A. Moles. 1958. Application de l'indice de Flesch à la langue française. *Cahiers Etudes de Radio-Télévision* 19, 1958 (1958), 253–274.

[13] A. Karakanta, H. Przybyl, and E. Teich. 2018. Exploring Variation in Translation with Relative Entropy. In *Proc. Int. Symposium on Parallel Corpora (ECETT/PaCor)*. Madrid.

[14] A. Klimashevskaia, B. C. Geiger, M. Hagmüller, D. Helic, and F. Fischer. 2020. 'To be or not to be central' – On the stability of network centrality measures in Shakespeare's 'Hamlet'. In *Proc. Digital Humanities (DH)*. online. https://doi.org/10.5281/zenodo.3989291

[15] M. Kubát and J. Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20, 4 (2013), 339–349. https://doi.org/10.1080/09296174.2013.830552

[16] E. Meeks and S. B. Weingart. 2012. The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities* 2, 1 (2012). http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/

[17] S. Pielström. 2015. Complexity Measures as Style Markers. In *Complexity Measures in Stylometry*. Würzburg, Germany. https://dhd-blog.org/?p=6471

[18] E. Pitler and A. Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Honolulu, 186–195.

[19] T. Schmidt, M. Burghardt, K. Dennerlein, and C. Wolff. 2019. Sentiment Annotation for Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts. In *Poster Session of Conf. on Language, Data and Knowledge (LDK)*. Leipzig, Germany, 45–50.

[20] J. A. Smith and C. Kelly. 2002. Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities* 36 (2002), 411–430.

[21] P. Trilcke, F. Fischer, M. Göbel, D. Kampkaspar, and C. Kittel. 2016. Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730-1930. In *Proc. Digital Humanities (DH)*. Krakow, Poland, 385–387.

[22] V. Volansky, N. Ordan, and S. Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities* 30, 1 (07 2013), 98–118. https://doi.org/10.1093/llc/fqt031

[23] K. von Prince and V. Demberg. 2018. POS tag perplexity as a measure of syntactic complexity. In *Proc. First Shared Task on Measuring Language Complexity*. Toruń, Poland, 20–25.