Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 977–984
Marseille, 11–16 May 2020

Stylometry in a Bilingual Setup

Silvie Cinková*, Jan Rybicki
*Charles University – Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Jagiellonian University – Institute of English Studies
*Malostranské nám. 25, Praha 1, 118 00 Czech Republic
31-120 Kraków, al. Mickiewicza 9A, Poland
cinkova@ufal.mff.cuni.cz, jan.rybicki@uj.edu.pl

Abstract

The method of stylometry by most frequent words does not allow direct comparison of original texts and their translations, i.e. across languages. For instance, in a bilingual Czech-German text collection containing parallel texts (originals and translations in both directions, along with Czech and German translations from other languages), authors would not cluster across languages, since frequency word lists for any Czech texts are obviously going to be more similar to each other than to a German text, and the other way round. We have tried to come up with an interlingua that would remove the language-specific features and possibly keep the linguistically independent features of individual author signal, if they exist. We have tagged, lemmatized, and parsed each language counterpart with the corresponding language model in UDPipe, which provides a linguistic markup that is cross-lingual to a significant extent. We stripped the output of language-dependent items, but that alone did not help much. As a next step, we transformed the lemmas of both language counterparts into shared pseudolemmas based on a very crude Czech-German glossary, with a 95.6% success. We show that, for stylometric methods based on the most frequent words, we can do without translations.

Keywords: stylometry, multilinguality, Universal Dependencies; authorship attribution; translation

1. Introduction

1.1 In Search for Individual Stylistic Profiles

Computational stylistics, or stylometry, is concerned with the quantitative characteristics of individual author style and the comparison of different authors, a facet of which is authorship attribution – selecting the (most likely) author of a given document from a group of authorship candidates by comparison of that document with documents by all authorship candidates.

The assumption is that authors display their individual unconscious patterns of language use, and that these patterns remain constant, no matter the topic. These patterns manifest themselves in the most frequent words (mostly function words). Normally, the intra-authorial style variation should be smaller than the variation between different authors.

In the stylometric comparison, the texts are compared as feature vectors of the most frequent words, and their similarity is computed as the distance of each text to each other text. The classic metric of authorial difference is Burrows' Delta (Burrows 2002; Hoover 2004), defined as the Manhattan distance of z-scores of the frequencies of *n* most frequent words in the collection.

Evert et al. (Evert et al., 2017) scrutinized Burrows' Delta as well as its different modifications. They replaced the Manhattan distance in Burrows' Delta by the cosine distance as a way of vector normalization, which substantially increased its performance. They argue that vector normalization makes the metric more robust against single extreme frequency values typical of a text rather than of an author, and that the "'stylistic profile' of an author manifests itself more in the qualitative combination of word preferences, i.e. in the pattern of over- and under-utilization of vocabulary, rather than in the actual amplitude of the z-scores" (p. ii11f.).

1.2 Stylometric Research on Literary Translations

Stylometric research into literary translation has already produced some interesting insights. Burrows found that some translators who are also authors in their own right may have a stylometric signal of their own, while others' texts may differ depending on whether the authors write their own works or translate some else's (Burrows, 2002a). Rybicki showed that cluster analysis of the Delta distances for several authors' texts in translation by various translators more often than not brings together texts by the original author rather than by the translator (Rybicki, 2012). On the other hand, when dealing with translations of the same text or the same author by two different translators, their unique stylometry is easier to detect (Rybicki, 2012; Rybicki and Heydel, 2013). Similarly varying successes of stylometric translator attribution were obtained by Forsyth and Lam (2013), while Lee suggested that the translators' (in)visibility may depend on the degree of difference between languages (Lee, 2018).

Adopting a somewhat different focus, stylometry was shown to reflect the history and the aims of Biblical translation (Covington et al., 2015). Other stylometric experiments showed that, in a large collection of fiction in one language, translations may form discrete communities defined by the source language, which are very distinct from each other and from native writing in the target language (Rybicki 2017).

All of these studies share a common problem. While they shed much light on the complex issue of translated literature, the method of stylometry by most frequent words does not allow direct comparison of original texts and their translations. In other words, an attempt to hang, say, novels in English and their multiple translations from the same cluster analysis tree would produce the trivial effect of separating the texts by the two languages, and little more.

Figure 1 illustrates this problem on Cosine Delta distances between pairs of parallel Czech and German fiction texts.

Each data point is a pair of texts, sorted by the pair language(s) on the X-axis (cross-language, Czech, and German) and located on the Y-axis according to their Cosine Delta distance. All possible pairs are considered. For pairs with matching authors, the points are colored, while the transparent black points represent pairs with different authors. Most colored points are among the cross-language pairs, because most authors are represented only with one parallel document. Three authors are represented by more than one parallel document, and therefore they also occur among the monolingual pairs: "Kunde" (Kundera), "Kipli" (Kipling), and "Cap" (Capek).

The plot is divided into facets (subgraphs) according to the length of the list of most frequent words abbreviated as "MFW". The boxplots render the distribution of the pairwise distances. The title of each facet indicates the size of the MFW list, as well as the number of culled words (zero in all cases). No matter the size of the MFW list (200 to 20,000), the pairs within the same language are closer to each other than cross-language pairs. This holds, no matter whether or not the authors of the documents match.

Even a very short word list (200 MFW) reveals the smallest Cosine Delta distances between texts by the same author in the same language, while cross-language documents generally keep large distances. The size of the MFW list does not affect the result very much, except that larger MFW lists tend to decrease the Cosine Delta distances between the German pairs and to increase those between the Czech pairs.

In our figure, the extremely small distances within "Kunde" and "Kipli" document pairs prevail in both Czech and German, whereas documents by "Cap" span from extremely low values to extremely high values. This suggests that "Cap" has a less distinct style than "Kunde" or "Kipli", and it could be explained by the fact that the documents by "Cap" are partly novels and short stories, and partly journalistic interviews and columns.

This language barrier for direct stylometric comparison of literary translation could be removed if words in the texts could be replaced with representations of grammatical entities such as parts-of-speech (POS) or parts-of-sentence tagging. The problem is that differences between grammars result in incompatible tagging systems; due to divergences in the degree of inflection, the number of POS tags required in, say, English and Polish may be of at least an order of magnitude (a hundred for the former versus a thousand for the latter language).

We were looking for a way to bring two languages on a common language-agnostic denominator. In our experiment, we replaced the – obviously language-specific – words with language-agnostic strings. These strings consisted of:

1. Diverse combinations of cross-linguistically universal morphosyntactic markup (**Universal Dependencies**, (Agić et al. 2015));
2. Pseudolemmas derived from a bilingual glossary retrieved from **Treq**, a database of translation equivalents based on a large, multilingual and multi-genre parallel corpus (Rosen 2016; Škrabal and Vavřín 2017a, 2017b).

The results turned out astonishingly good. In this paper, we first describe the components of the language-agnostic strings for both languages, which are, in our case, almost randomly chosen Czech and German (cf. Section 3). Then we present the collection of texts. Eventually, we present the experiment and its results.

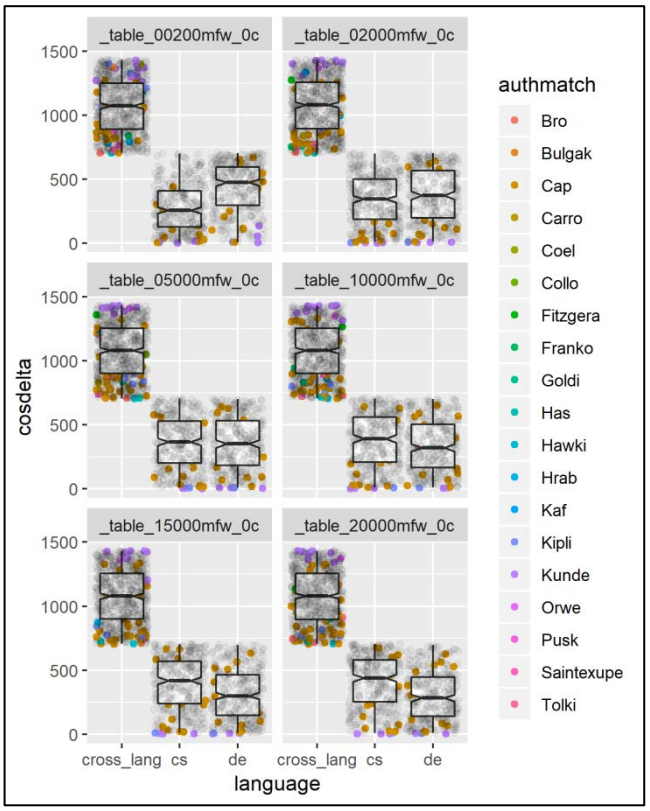

Figure 1: Cosine Delta with full bilingual texts

2. Components of a Language-Agnostic Version of our Text Collection

2.1 Universal Dependencies

We enriched all texts from our collection with **Universal Dependencies** (Agić et al. 2015, de Marneffe et al., 2014-2018). Universal Dependencies is a framework for consistent annotation of morphological categories and syntactic dependencies across different human languages. Universal Dependencies has gradually grown to a standard for syntactically annotated corpora (treebanks): the latest release from end 2019 contains more than 150 treebanks in over 90 languages.

Universal Dependencies provides two sets of morphological tags, based on Zeman, 2008: universal parts of speech (upos) and universal features. The former contains approx. 20 tags denoting exclusively parts of speech (e.g. NOUN, VERB) and diverse non-word tokens (SYM, X, PUNCT). The latter is a large pool of morphological categories associated with certain parts of speech (e.g. number or case) in form of attributes and their values (e.g. Gender= Feminine). While most languages use the full inventory of upos, the selection of feature attributes and their values is language-dependent.

For instance, nouns in several Slavic languages make use of Animacy, Nouns in North-Germanic languages do not. On the other hand, they use Definiteness (due to suffigated definite articles), which is not the case of all Slavic languages but Bulgarian. Feature attributes that nouns in both language groups have in common (case and gender) have language-specific sets of accepted values. For instance, Czech and Polish have the Vocative case, which Russian lacks. Among North-Germanic languages, Icelandic has four cases (Nominative, Genitive, Dative, and Accusative), while others only Nominative and Genitive. Nevertheless, the set of accepted attributes and values in Universal Dependencies is closed and their usage is documented in the annotation guide for each language. For instance, Swedish and English pronouns have two cases: Nominative and Accusative. Even a noun in the position of the indirect object is classified as an accusative noun, although it corresponds to dative in most languages and some grammars call the two morphologically distinct pronominal cases in these virtually non-inflective languages *subject/object case*. This is to say that Universal Dependencies abstract from language-specific traditional grammars, whenever it serves the idea of cross-lingual unification.

Universal Dependencies also have a set of syntactic dependency relations (deprels). Each label denotes the relation of the given word to its governing word in a syntactic dependendency tree. A simple visualization of a sentence tree with the deprel and upos labels is in Figure 2.

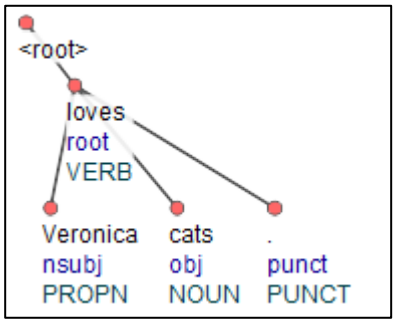

Figure 2: A syntactic dependency tree in Universal Dependencies

Most languages use the full inventory of syntactic labels.

2.2 A Bilingual Glossary from Treq

Treq (Škrabal and Vavřín 2017a) is an application to look up translation equivalents in bidirectional Czech-foreign language glossaries automatically extracted from parallel texts in the InterCorp corpus (Rosen 2016).

The InterCorp corpus contains Czech texts of diverse genres manually sentence-aligned with one or more foreign-language counterparts. More than 30 foreign languages are represented in the release currently available through the public web GUI[1] in major European languages with between 70 and 160 M running words.

The current release of Treq, which is derived from InterCorp, offers automatic word-to-word alignments for any two languages from the following: Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Japanese, Latvian, Lithuanian, Macedonian, Malay, Maltese, Norwegian, Polish, Portuguese, Romani, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, and Vietnamese. The size of each aligned word set depends, however, on the size of the parallel texts in InterCorp in the given language, on the quality of the manual sentence-to-sentence alignment, as well as on the formal equivalence (Nida, 1964) of the translations with respect to sentence splitting.

The automatic word-to-word alignment in Treq is based on the manual sentence-to-sentence alignment between Czech (the pivot language of InterCorp) and each foreign language counterpart. Only 1:1 aligned sentences of the Czech-foreign text pairs in InterCorp were used. The result is a database of cross-lingually aligned lemmas.

Treq has primarily been designed for lexicographers to manually browse through corpus-based data on translation equivalents. Therefore it is normally only available through a web GUI hosted by the Institute of the Czech National Corpus[2]. However, we obtained a tabular text file with the Czech-German alignment on request. With this bilingual glossary, we created a lexicon of cross-lingual pseudolemmas to replace the Czech and German words in our collection (see Section 4.2).

3. Our Text Collection

To find suitable parallel texts, we were searching InterCorp[3] for a subcorpus in a language pair where several authors would be represented by more than one document. The translation direction did not matter, but we preferred languages we would understand and languages strongly represented in InterCorp to make sure that the aligned word pairs from Treq would be based on large data.

From subcorpora meeting these criteria, we randomly selected the Czech-German pair from InterCorp. The Czech-German pair consists of 27 pairs of parallel texts: Czech originals and their German translations, one German original and its Czech translation, and translations of works originally written in English, French or Russian. Most of the titles was modern fiction (since 1920s), but other genres occurred as well. We list the titles here:

- Brown, *The Da Vinci Code*
- Bulgakov, *The Master and Margarita*
- Capek, *Dashenka*
- Capek, *Talks with T. G. Masaryk*
- Capek, *Krakatit*
- Capek, *The Absolute at Large*
- Capek, *War with the Newts*
- Capek, *The Gardener's Year*
- Carroll, *Alice in Wonderland*
- Coelho, *The Alchemist*
- Collodi, *The Adventures of Pinocchio*
- Fitzgerald, *The Great Gatsby*

[1] https://kontext.korpus.cz/

[2] https://korpus.cz

[3] Cf. Section 2.2, second paragraph (Rosen, 2016).

- Frank, *Diary*
- Golding, *Lord of Flies*
- Hasek, *The Good Soldier Švejk*
- Hawking, *A Brief History of Time*
- Hrabal, *I Served the King of England*
- Kafka, *The Trial*
- Kipling, *The Jungle Book*
- Kipling, *The Second Jungle Book*
- Kundera, *Immortality*
- Kundera, *The Unbearable Lightness of Being*
- Kundera, *The Joke*
- Orwell, *Nineteen Eighty-Four*
- Pushkin, *The Captain's Daughter*
- Saint-Exupéry, *The Little Prince*
- Tolkien, *Lord of the Rings* vol. 1.

So, in the Czech-German language pair, we had three authors to form each their own cluster: Capek, Kipling, and Kundera.

4. Corpus Preprocessing

4.1 Annotation with Universal Dependencies

We converted all texts to the UD annotation, using the Czech and the German language models of the UDPipe parser (Straka et al., 2016). The output from UDPipe comes in the CoNLL-U format, a derivation of CoNLL-X, which in turn is a well-established format in the NLP and CL community (Buchholz and Marsi, 2006). It is a plain text table, with one token per line and the following columns:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOS: Universal part-of-speech tag.
5. XPOS: Language-specific part-of-speech tag.
6. FEATS: Alphabetically ordered list of morphological features from the universal feature inventory or from a defined language-specific extension.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
10. MISC: Any other annotation.

To keep the annotation as language-independent as possible, we used the UPOS and DEPREL columns as they were, but we substantially reduced data in the FEATS column by removing language-specific attribute values.

4.2 Cross-lingual Pseudolemmas

As a next step, we replaced the columns FORM and LEMMA with one single PSEUDOLEMMA column. The pseudolemma strings were designed in a completely arbitrary way, starting with an "L" for "lemma" and a 5-digit sequence derived from the row ID of the tabular database output from Treq, which we had obtained from the Institute of the Czech National Corpus. Each pseudolemma stood for one Czech-German pair of aligned words.

Table 1 shows the Treq output for Czech and German. The rows are ordered according to the frequency of the alignment pairs. This illustrates very well how noisy a lexical resource Treq is: the two most frequent alignment pairs are punctuation, and the most common pair of word tokens is grossly wrong: the lemma on the Czech part means "to be", whereas the German "sie" (which could originally even have been "Sie") is a polysemous pronoun meaning "she", "they", or (when capitalized) "you" (polite form). The Czech "ten" on the fourth row is a demonstrative pronoun denoting a masculine singular of "this". It is aligned with "die", which is either a feminine singular or gender-nonspecific plural of "these", but mostly it acts as the definite article for plural and feminine singular. The first and only full match in the top sample is "že"-"dass". The others top-down make good sense, but they are mostly polysemous words.

freq	CS	DE	pseudolemma
182686	"	«	L00009
163642	?	?	L00010
139614	Být	sie	L00011
139120	Ten	die	L00012
128976	Že	dass	L00013
124208	Který	die	L00014
118355	Já	sie	L00015
112098	Na	auf	L00016
107206	S	mit	L00017
91365	!	!	L00018

Table 1: Treq glossary with generated pseudolemmas.

Since we wanted the procedure to involve no manual corrections, we left this table untouched, in spite of its evident messiness. Then we had a script go through each document in Czech, row by row in the CoNLL-U table, compare the value in the LEMMA column to the Treq table arranged according to frequency, just like in Table 1. We replaced the form and lemma in the CoNLL-U document with the first match in the CS column. That is, each occurrence of quotes, question mark, and "být" in a Czech CoNLL-U file was replaced with pseudolemma L00009, L00010, and L00011, respectively. We ran the same procedure for the German CoNLL-U files, matching lemmas to their first occurrences in the DE column. For instance, each occurrence of "auf" in the German CoNLL-U files was replaced with L00016.

By this very primitive matching, we naturally neglected polysemy, and, due to the noise in the Treq table, some equivalents were admittedly suboptimal. This occurred even in very frequent words, where we would have expected a substantial negative impact on the result.

5. Language-Agnostic Input for Stylometric Analysis

The starting format for the languages-agnostic documents was the tabulated CoNLL-U format, which we stripped off all columns we found irrelevant for the task. On the other hand, we added the pseudolemma column. Table 2 and Table 3 illustrate the format with a sentence from the beginning of *The Trial* by F. Kafka: "This had never happened before." In the actual experiment, we also stripped the token and lemma columns, and we removed language-specific attribute values from the FEATS column. In this case, it would read "Variant=Short" for the Czech reflexive pronoun "se".

Looking closer at Tables 2 and 3 and taking into account that the pseudolemmas are constructed from frequency ranks in the glossary, one might wonder about the numeric part of the pseudolemma representing the German "das", which renders a surprisingly high frequency rank, considering its role as a definite article. The explanation is simple: it is not about the frequency of a given word in its language but about the frequency of the entire alignment pair in the bilingual glossary. The most frequent usage of the German "das", the definite article, has no systematic equivalent in Czech. Apart from that, the best equivalent of the German "das" as a demonstrative pronoun, the Czech the neutral singular demonstrative pronoun "to", is lemmatized as "ten"; that is, its masculine singular form. The alignment pair "ten-das" ranks 280. The most frequent alignment to the Czech "ten" is "die", which ranks 12 (see Table 3).

token	**Lemma**	**upos**	**feats**	**deprel**	**pseudo-lemma**
Das	Das	PRON	Case=Nom|PronType=Dem	nsubj	L11737
war	Sein	VERB	Mood=Ind|Person=3|Tense=Past|VerbForm=Fin	cop	L00005
noch	Noch	ADV	NA	advmod	L00034
niemals	niemals	ADV	NA	advmod	L00542
geschehen	geschehen	VERB	VerbForm=Part	root	L00504
.	.	PUNCT	NA	punct	L00002

Table 2 : Parsed German sentence with pseudolemmas

token	**Lemma**	**upos**	**feats**	**deprel**	**pseudo-lemma**
To	ten	DET	Case=Nom|Gender=Neut|Number=Sing|PronType=Dem	nsubj	L00012
se	se	PRON	Case=Acc|PronType=Prs|Reflex=Yes|**Variant=Short**	expl	L00008
ještě	ještě	ADV	NA	advmod	L00034
nikdy	nikdy	ADV	**PronType=Neg**	advmod	L00124
nestalo	stát	VERB	Gender=Neut|Number=Sing|Polarity=Neg|Tense=Past|VerbForm=Part|Voice=Act	Root	L00110
.	.	PUNCT	NA	punct	L00002

Table 2 : Parsed Czech sentence with pseudolemmas

From the format illustrated by Tables 2 and 3, we generated files for several different experimental setups. We refer to the corresponding columns in the tables:

1. only POS (upos): Column 3;
2. POS + features: Colums 3 and 4 ;
3. only pseudolemmas: Column 6;
4. pseudolemmas + POS + syntactic dependencies (deprels): Columns 6, 4, and 5;
5. pseudolemmas + POS + features + syntactic dependencies: Columns 6, 3, 4, and 5;
6. POS + features + syntactic dependencies: Columns 3, 4, and 5;
7. pseudolemmas + POS + features: Columns 6, 3, and 4;
8. pseudolemmas + POS: Columns 6 and 3.

In each setup, each row represented one original word (token), and the row consisted of a concatenation of the values of the selected columns. This format was our language-agnostic format. For instance, the representations of the word "Das" from Table 2, would be the following:

- PRON (POS)
- PRON_Case=Nom|PronType=Dem (POS + features)
- PRON_Case=Nom|PronType=Dem_nsubj (POS + features + syntactic dependencies)
- PRON_Case=Nom|PronType=Dem_nsubj_L11737 (POS + features + syntactic dependencies + pseudolemmas)
- PRON L11737 (POS + pseudolemmas).

In the next step, we measured the distances between all texts in the collection with Cosine Delta (Smith and Aldridge 2011), which is now seen as the most reliable version (Evert et al. 2017). We performed this step separately for each experimental setup, using the R (R Core Team, 2016) library **stylo** (Eder et al., 2016).

6. Parameters of the Stylometric Analysis

We used the **classify()** function in the stylo package (Eder et al. 2016) for R to try to assess authorship attribution success, when its reference set contained texts in one language and the test set contained texts in the other. The attribution success was counted whenever the Cosine

Delta value for the pair of the translations of the same text was lowest. Each distance measurement was based on the mean of results from **classify()** runs starting with 100 most frequent string sequences and ending with 2000 most frequent string sequences, incrementing the number by 100 (19 runs), except for the "POS only" setup, where the "vocabulary" of the documents was only as large as the upos inventory (around 20 tags). For the prediction we used Support Vector Machines.

7. Results

Table 4 presents attribution success. Apart from these experiments, we performed several other analyses n-grams of the "words" resulting from the different setups, with little or no impact on the results, so we do not present them here.

Tagging	**Attribution success**
POS	3.7%
POS + features	3.7%
Pseudolemmas	3.7%
pseudolemmas + POS + syntactic dependencies	10.2 %
pseudolemmas + POS + features + syntactic dependencies	16.7%
POS + features + syntactic dependencies	20.3%
pseudolemmas + POS + features	56.7%
pseudolemmas + POS	**95.6%**

Table 3: Attribution success with different combinations of markup

When using all information we have – POS, features, syntactic dependencies, and pseudolemmas – the attribution success is 16.7%.

Without pseudolemmas, we get 20.3% attribution accuracy, which is certainly a modest result, but still far above random guess.

The winning combination is POS + pseudolemmas, at 95.6%. The success of the pseudolemmas shows that even a very crude word-to-word translation (polysemy neglected), along with the coarse part of speech tagging, helps bypass the language barrier.

8. Discussion

Figures 3 and 4 illustrate attribution success of the worst and the best feature combinations. The only cross-lingual link in Figure 3 is the obviously wrong one between the German version of Capek's *The Absolute at Large* and the Czech version of Tolkien's *The Lord of the Rings*. Interestingly, this link even prevails in the combination of POS and pseudolemmas, where it represents the only cross-lingual error. We assumed a particularly large vocabulary overlap, despite different languages. With 66 tokens represented in both language versions, this pair is at the 68th percentile of the (absolute) amount of cross-lingual homographs. Besides punctuation and numbers, we observed approximately 40 cross-lingual homographs of major word classes (e.g. *jeden, Strom, kamen, elf*), with no apparent cognitive associations. Only the homography of *elf* (German for *eleven*) in Capek and *elf* (in *Lord of the Rings*) was somewhat specific to this pair. The book pairs with the highest overlap contained the Good Soldier Svejk, with its many actual German words.

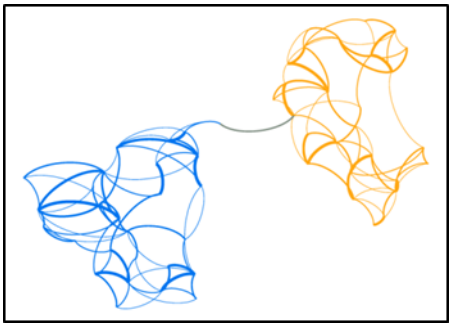

Figure 4: POS + features

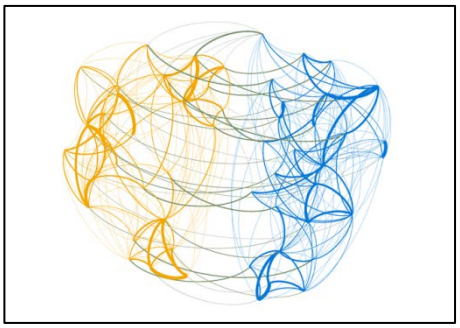

Figure 3: POS + pseudolemmas

Considering the contribution of the individual features (POS, features, syntactic dependencies, and pseudolemmas), we speculate that POS alone form too small a vocabulary to reveal style differences within one language. The features could have made it richer, but they remained too language-specific still; that is, they continued to appear in language-specific combinations with individual POS. For instance, Czech verbs in the past tense, unlike the German ones, indicate Gender. By extending the POS by the features, the vocabulary must certainly have grown, but at the same time we introduced language-specific items.

The syntactic dependencies only helped in combination with POS and features. At the moment, we have no clue how exactly the syntactic dependencies helped POS and features on the one hand, but harmed the pseudolemmas/POS combination on the other. At any rate, we realize that we have not yet exploited the potential of syntactic dependencies. In the future, we are going to explore syntactic n-grams of tokens connected by specified syntactic dependencies, with or without listing the dependency labels.

Prior to the bilingual Czech-German experiment, we experimented with a truly multilingual parallel corpus, where all documents had a Czech version and many had counterparts in several other languages. In this setup, the

language-agnostic tokens consisted only of the UD annotation.

The best-clustering titles were Hawking's *A Brief History of Time* (Figure 5) and Anne Frank's *Diary* (Figure 6). Both make sense – they are not fiction and therefore likely to be extremely different from the other titles within their language.

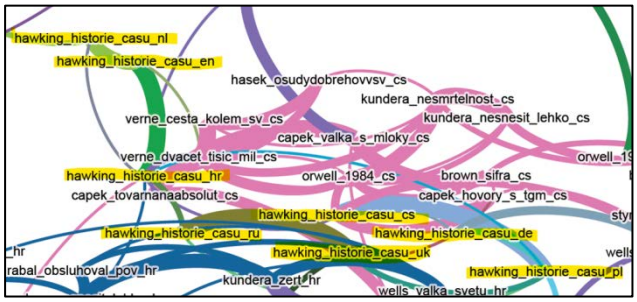

Figure 5: Multilingual clustering of Hawking's History of Time

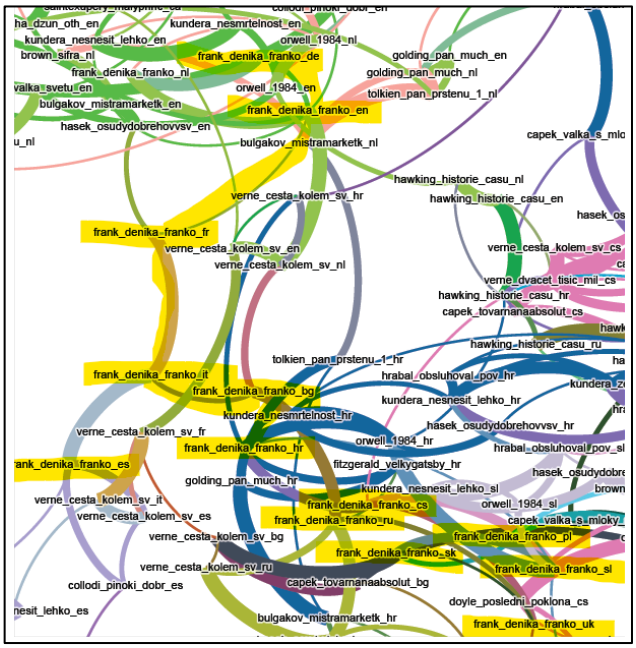

Figure 6: Multilingual Clustering of Anne Frank's Diary

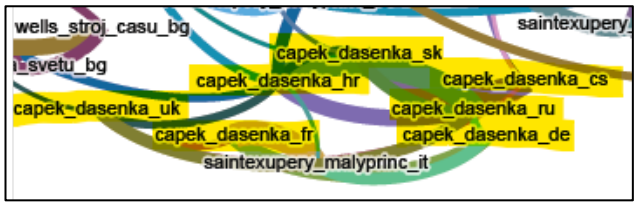

Figure 7: Multilingual clustering of children books

Another sensible (although incorrect) guess is a link between Capek's children's book *Dashenka* and Saint-Exupéry's *Little Prince* (Figure 7). In these books, both authors, who were established adult fiction authors, were formally addressing children, both doing so in a somewhat philosophical manner. This could be a style bias substantial enough to even affect the morphological and syntactic language layers in a cross-lingually uniform way – as is at least our speculation.

9. Conclusion

Although we have not succeeded to use the Universal Dependencies in a way universal enough to crack the language barrier, we have observed a few promising partial results. When we resorted to just a bilingual setup and aided the language-agnostic tokens on both language parts with a very noisy, automatically generated glossary, we saw a profound and unexpected success.

This initial study opens several interesting research questions to pursue further:

1. If even a noisy glossary works, then machine translation could work as well, perhaps even with more than just two different languages.
2. N-grams of syntactic dependencies could yield more explicit information on the syntactic structure of individual sentences than linear sequences and linear n-grams do. Intuitively, this could help abstract from language-specific differences in word order.

10. Acknowledgements

This work has been supported by LTC18020 and COST CA16204. We were using tools and infrastructure of LINDAT-CLARIAH (LM2018101; formerly LM2010013, LM2015071) and Czech National Corpus (LM2015044; 2016-2019), both fully supported by the Ministry of Education, Sports, and Youth of the Czech Republic under the "LM" programme of "Large Infrastructures".

11. Bibliographical References

Agić, Ž., Aranzabe, M., Atutxa, A., Bosco, C., Choi, J., Marneffe, M.-C. de, et al. (2015). Universal Dependencies 1.1. Praha, Czechia: LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Buchholz, S. and Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X) Pp. 149–164. New York City: Association for Computational Linguistics. http://aclweb.org/anthology/W06-2920.

Burrows, J.F. (2002). "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17: 267-287.

Burrows, J.F. (2002a). "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36(4): 677-298.

Covington, M. A., Potter, I., and Snodgrass, T. (2014). *Literary and Linguistic Computing* 30 (3): 322-325.

Eder, M., Kestemont, M., and Rybicki, J. (2016). "Stylometry with R: A package for computational text analysis." *The R Journal* 8(1): 107–121.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). *Digital Scholarship in the Humanities* 32 (sup. 2): 4-16.

Forsyth, R. S. and Phoenix W. Y. Lam. (2013). "Found in translation: To what extent is authorial discriminability

preserved by translators?" *Literary and Linguistic Computing* 29 (2): 199-217.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4), 453–475. https://doi.org/10.1093/llc/19.4.453

Lee, Ch.. (2018). "Do language combinations affect translators' stylistic visibility in translated texts?" *Digital Scholarship in the Humanities* 33 (3): 592–603.

de Marneffe, M. C. et al. (2014-2017). *Universal Dependencies*, http://universaldependencies.org, accessed 14 Nov. 2018.

Nida, E. (1964). "Principles of Correspondence." In Eugene Nida, *Toward a Science of Translating*, Leiden: E.J. Brill, 156–71.

Petrov, S., Dipanjan Das, and Ryan McDonald. 2012. "A universal part-of-speech tagset." In *Proceedings of LREC*, http://www.petrovi.de/data/universal.pdf, accessed 14 Nov. 2018.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosen, A. (2016). InterCorp – a Look behind the Façade of a Parallel Corpus. In Polskojęzyczne Korpusy Równoległe. Polish-Language Parallel Corpora. Ewa Gruszczyńska and Agnieszka Leńko-Szymańska, eds. Pp. 21–40. Warszawa: Instytut Lingwistyki Stosowanej. http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/02_Rosen.pdf.

Rybicki, J. (2012). "The great mystery of the (almost) invisible translator." In Michael P. Oakes, Meng Ji (eds), *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, Amsterdam: John Benjamins, 231-248.

Rybicki, J. (2014). "Pierwszy rzut oka na stylometryczną mapę literatury polskiej," *Teksty drugie* 2: 106-128.

Rybicki, Jan. (2016). "Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies," *Digital Scholarship in the Humanities* 31(4): 746-761.

Rybicki, J. (2017). "A Second Glance at a Stylometric Map of Polish Literature." *Forum of Poetics* 8: 6-21.

Rybicki, J. and Heydel, M. (2013). "The Stylistics and Stylometry of Collaborative Translation: Woolf's *Night and Day* in Polish." *Literary and Linguistic Computing* 28 (4): 708-717.

Smith, P. and Aldridge, W. (2011). "Improving authorship attribution: Optimizing Burrows' Delta method." *Journal of Quantitative Linguistics*, 18(1): 63–88.

Zeman, D. (2008). "Reusable Tagset Conversion Using Tagset Drivers." In *Proceedings of LREC*, http://lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf, accessed 14 Nov. 2018.

Škrabal, M., and Vavřín, M. (2017). Databáze Překladových Ekvivalentů Treq. Časopis pro Moderní Filologii 99(2): 245–260.

Škrabal, M., and Vavřín, M. (2017a). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 124–137). Presented at the Electronic lexicography in the 21st century., Leiden: Lexical Computing CZ s. r. o. https://elex.link/elex2017/proceedings-download/ . Accessed 3 March 2020

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, et al., eds. Pp. 4290–4297. Paris, France: European Language Resources Association.