Coreference Strategies in English-German Translation

Ekaterina Lapshinova-Koltunski
University of Hildesheim/
Saarland University
e.lapshinova
@mx.uni-saarland.de

Marie-Pauline Krielke
Saarland University
mariepauline.krielke
@uni-saarland.de

Christian Hardmeier
Edinburgh University/
Uppsala University
christian.hardmeier
@lingfil.uu.se

Abstract

We present a study focusing on variation of coreferential devices in English original TED talks and news texts and their German translations. Using exploratory techniques we contemplate a diverse set of coreference devices as features which we assume indicate language-specific and register-based variation as well as potential translation strategies. Our findings reflect differences on both dimensions with stronger variation along the lines of register than between languages. By exposing interactions between text type and cross-linguistic variation, they can also inform multilingual NLP applications, especially machine translation.

1 Introduction

Coreference devices and their usage vary both across and within languages depending on several factors such as register and style among others. Moreover, translation process evokes a number of translation phenomena, such as explicitation or interference (Blum-Kulka, 1986; Toury, 1995) that also have an impact on the choice of linguistic expressions used. We assume that variation in coreference devices depends on the following factors: (a) language-specific constraints, (b) functional variation across language registers as well as (spoken or written) mode and (c) effects of the translation process.

Translating between languages involves transformation of the source coreference patterns into the target ones. Analysing such patterns can give insights into translation strategies for referring expressions in texts. Variation along the above stated lines (a, b, c) causes a number of problems in multilingual coreference resolution or coreference annotation projection (Postolache et al., 2006; Ogrodniczuk, 2013; Grishina and Stede, 2015; Novák, 2018). Although several studies describe such problems (Grishina and Stede, 2015; Lapshinova-Koltunski and Hardmeier, 2017; Lapshinova-Koltunski et al., 2019b), there is still a lack of understanding as to which linguistic phenomena and concretely, which structures cause these problems. In this paper, we attempt to detect such phenomena for English-German translations in a data set containing two different text registers: TED talks, which represent spoken language, and news, a type of written discourse. Previous studies show that the choice of referring expressions depends on the mode of text production in both languages under analysis (Kunz et al., 2016; Kunz et al., 2017).

The phenomena we analyse are not restricted to expressions referring to simple, nominal antecedents, but also comprise expressions referring to events. We also include cases of comparative reference, substitution and ellipsis, which according to Kunz and Steiner (2012) trigger a type reference relation (and not the relation of identity), or "sloppy identity".

Our main goal is to shed light on cross-lingual differences in coreference expressions. Another goal is to explore possible translation strategies for the language pair under analysis and the given text types. For this, we perform an empirical, corpus-based analysis using a number of coreference features as indicators of cross-lingual variation. The features are extracted from an existing corpus annotated with coreference chains. They include morpho-syntactic and functional properties of referring expressions, as well as chain properties. Using correspondence analysis and hierarchical cluster analysis we detect specific features of coreference distinctive for the two dimensions (language contrast and register variation) under

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 139–153,
Barcelona, Spain (online), December 12, 2020.

analysis. Our results show that both language and register (or mode) give rise to differences in our data and the coreference features vary depending on the dimensions.

Knowledge on the differences in the realisation of the coreference phenomena depending on the language or register is valuable to cross-lingual coreference resolution, as it was already acknowledged in the community, e.g. the CoNLL-2012 shared task on coreference resolution included multiple languages, registers and modes within OntoNotes (Recasens and Pradhan, 2016). Knowledge of the analysed variation is even more important for contrastive linguistics and (machine) translation.

2 Related Work

Several studies in the area of translation have addressed the importance of coreference (Baker, 2011; Becher, 2011; Königs, 2011). However, these works are example-based and provide neither a comprehensive account, nor empirical evidence for their claims. There are a few corpus-based studies of coreference translation (Zinsmeister et al., 2012; Novák and Nedoluzhko, 2015; Lapshinova-Koltunski et al., 2019b), addressing mostly the challenge of translating pronouns. The awareness of this challenge has also increased in the MT community (Voita et al., 2019; Lapshinova-Koltunski et al., 2019a; Guillou et al., 2018; Bawden et al., 2018; Miculicich Werlen and Popescu-Belis, 2017; Guillou, 2016; Hardmeier and Federico, 2010), and its relevance for multilingual coreference resolution is beyond doubt (Green et al., 2011; Novák and Žabokrtský, 2014; Grishina and Stede, 2017). Still, coreference translation is affected by many factors and remains poorly understood.

Kunz et al. (2017) analyse coreference and other means of explicit discourse phenomena in English and German comparable texts. They find that English and German differ in the linguistic means available in their language systems to convey coreference. English provides less syntactic flexibility and is restricted in the distribution of referents. German has more options and tends to use more grammatical means of coreference than English (Kunz et al., 2017), which indicates that English and German differ in how coreference chains are built up in terms of form and type of referring expressions within the chains. Kunz et al. (2017) base their analyses on the assumptions within contrastive pragmatics (House, 1997) suggesting that meanings are expressed more explicitly by linguistic signals in German than in English. However, Kunz et al. (2017) claim that translation strategies cannot rely on knowledge about contrastive lexico-grammar alone – awareness of preferred patterns that distinguish the languages and registers are essential for translators. The authors state that translating coreference chains from English into German implies using a higher number of coreferring expressions, and at the same time, chains of two elements may drop out because of remetaphorisation (i.e. change in word class). The authors use comparable corpora of original texts in both languages for their analyses, which does not provide them with the insights of what is actually happening in translation in the given language pair. Their findings include recommendations for translation strategies, but not the observation of the translation behaviour.

Lapshinova-Koltunski and Martínez Martínez (2017) focus on cohesive devices in English and German original (spoken and written) texts and report a higher degree of cohesiveness in German than in English, while overall spoken texts show more cohesive devices than written texts. Lapshinova-Koltunski et al. (2019b) analyse incongruences in the annotation of nominal coreference in English-German translations. The majority of the discovered incongruences are caused by explicitation – German translations contain more explicit linguistic devices triggering coreference. They point out that these cases of explicitation do not necessarily arise from the translation process, but can also be caused by idiosyncrasies of the two languages in terms of coreference properties. However, while looking into parallel chains, the authors do not provide analyses of the differences in the type of referring expressions in the source and the target texts. Moreover, their analysis is restricted to nominal coreference only.

Based on the previous literature, we can assume that the German translations contain more referring expressions and more chains than the English source texts. Conversely, shorter chains should be more common in the English originals than in the German translations. In terms of the structure and function of coreference chain members, there should be more explicit linguistic devices expressing coreference in the translations. In particular, translations would prefer demonstrative forms instead of personal reference, and entity reference instead of events (with entity reference being more concrete). In addition to such differences between originals and translations, we also expect variation in terms of register or mode, which may occasionally be even more prominent than that between languages (Kunz et al., 2016).

3 Analysed Features

In our analyses, we use a number of coreference features that are related to the form, functional and structural properties of chain members, as well as chain properties. First of all, we are driven by the structures available in the annotated corpus at hand. The morpho-syntactic and functional subtypes of anaphors and cataphors (which we collectively refer to as referring expressions) are motivated by the analyses by Becher (2011), who grades various types of referring expressions according to their degree of explicitness. This is important for our analyses, as our data contains translations, and explicitation – a higher explicitness of linguistic means in translated texts – is a well-known effect of the translation process. The levels of explicitness of referring expressions are related to Ariel (1990)'s concept of Accessibility. Morpho-syntactic types of referring expressions are related not only to accessibility, but also to the givenness or salience of a referent in the recipient's mind (Prince, 1981; Grosz et al., 1995; Gundel et al., 2003).

In studies involving register or genre variation, the distribution of morpho-syntactic types of mentions, such as the prevalence of pronouns vs. nouns, also plays an important role (Fox, 1987; Biber et al., 1999; Amoia et al., 2012; Kunz et al., 2016). Morpho-syntactic subtypes of referring expressions, substitution and ellipsis, as well as the scope of antecedents were analysed by Kunz et al. (2017) and Lapshinova-Koltunski and Martínez Martínez (2017) to reveal differences between registers and between the languages English and German. The scope of coreference is reflected in the differentiation between reference to entities vs. events, and in the form of the antecedent (nominal, verbal, clausal). As referring expressions can have more than one antecedent, the distinction between split and simple antecedent is also important. The resolution of anaphors with multiple antecedents differs in its processing from the resolution of single anaphors (Eschenbach et al., 1989).

The accessibility of a referent is also related to certain chain properties (Eckert and Strube, 2000): a high degree of accessibility is related to low distance between anaphors in long coreference chains and a low overall number of different coreference chains. The distance between anaphors and their antecedents is an important metric in many coreference resolution systems. Distance can be measured in different ways. The distance measure used in this work is the number of intervening sentence boundaries. This metric was also used by Nguy et al. (2011) for coreference resolution in Czech and by Amoia et al. (2012) for the analysis of variation in spoken and written texts.

In this study we include the following categories to analyse variation in the English-German coreference chains (a more detailed description is contained in Table 2 in the Appendix):

1. **morpho-syntactic types of all mentions** (including antecedents and referring expressions): pronoun (pp.m), noun phrases (np.m), verbal phrases (vp.m), clause (clause.m);

2. **types of reference**: pronouns functioning as anaphors (pp.anap), as cataphors (pp.cat), expressing substitution (pp.subs), comparative reference (pp.cmp), extratextual (extrtxt.ref) and pleonastic pronouns (pp.pleon);
   nominal phrases used as apposition (np.app), as comparative reference (np.cmp), as referring expression (np.ref),

3. **morpho-syntactic types of anaphoric expressions**:
   - pronouns: personal (pers.pp), possessive (poss.pp), demonstrative (dem.pp), reflexive (refl.pp), relative (rel.pp);
   - noun phrases sorted by their modifiers: possessive (poss.np), demonstrative (dem.np), def-article (def.np), indefinite (indef.np), bare noun phrases (bare.np);
   - comparative reference: particular and general (np.cmp.part, pp.cmp.part, np.cmp.gen and pp.cmp.gen);
   - substitution: nominal (np.subs) and verbal (vp.subs);

- ellipsis: nominal (np.ell) and verbal (vp.ell);

4. **types of referring expressions** measured by types of antecedents they refer to: referring expressions to entities (entity.ant.ref), referring expressions to events (event.ant.ref) and referring expressions to generics (gen.ant.ref); simple antecedent (simple.ant.ref), split reference (split.ant.ref), no explicit antecedent (noexpl.ant.ref);

5. **types of antecedents** by their form: pronoun (pp.ant), nominal phrase (np.ant), verbal phrase (vp.ant), clauses (clause.ant);

6. **chain properties**:
   - number of chains: total number (nr.chain)
   - chain length: mean chain length (mn.chain.lngth), median chain length (mdn.chain.lngth), standard deviation of chain length (stddv.chain.lngth), longest chain (lngst.chain), number of shortest, i.e. two-member chains (m2.chain), three-member chain (m3.chain), four-member chain (m4.chain) and five and more member chain (m5.chain)
   - distance between chain members measured in sentences (chain.dist).

4 Data and Methods

4.1 Data

For our analyses, we use ParCorFull (Lapshinova-Koltunski et al., 2018), a parallel corpus of English-German translations that is manually annotated for full coreference chains[1]. Coreference chains in this corpus consist of (mostly) chain-initial antecedents and anaphoric expressions that include pronouns, nouns, nominal phrases. Verbal phrases and clauses are also included as antecedents of event anaphors. The authors annotated elliptical constructions and cases of substitution, see details described by Lapshinova-Koltunski and Hardmeier (2017) and Lapshinova-Koltunski et al. (2018). The corpus contains transcribed TED talks and news texts in English (EN) and their corresponding German translations (DE). The summary statistics for the corpus data are given in Table 1.

subcorpus	texts	tokens	sentences	mentions	chains
EN-TED	20	70,736	3,379	5,970	2,121
DE-TED	20	66,783	3,555	5,911	2,206
EN-news	19	10,798	543	684	213
DE-news	19	10,602	543	576	269
EN-TOTAL	39	81,534	3,922	6,654	2,334
DE-TOTAL	39	77,385	4,098	6,487	2,475

Table 1: Corpus statistics

The corpus contains 39 parallel texts varying in their size from 368 tokens (the shortest news text) to 6,128 tokens (the longest TED talk). Although the number of texts in the news and in the TED part are similar, the size of the news portion in terms of tokens is much smaller. The news texts and part of the TED talks are aligned on the sentence level.

4.2 Methods

We extract frequency distributions of the features defined in Section 3 above from the corpus, save them in a contingency table with subcorpora or texts in rows and features in columns. Then, we use two explorative multivariate techniques to analyse the data: Correspondence Analysis (CA) and Hierarchical Cluster Analysis (HCA). Both analyses are performed in R environment (R Core Team, 2017, R version 3.6.1): we use the package `ca` to perform correspondence analysis and `pvclust` and `pvrect` to perform hierarchical cluster analysis.

[1]The corpus is available from the LINDAT repository at `http://hdl.handle.net/11372/LRT-2614`.

Correspondence analysis (Greenacre, 2007) is an extension of principal component analysis and fits good to explore relations between variables in a data set, as it summarises and visualises data in a two-dimensional plot. CA allows to study both sets of variables – those constituting the rows and those in columns of the contingency table[2]. We use CA to see which variables, in our case subcorpora, have similarities and how these subcorpora correlate with the coreference features contributing to the similarities. Weighted Euclidean distances, termed the χ^2 distances are measured on the basis of the feature distributions across the subcorpora. The data in the contingency table is scaled so that rows (subcorpora) and columns (features) are treated in an identical manner and so, the row and column projections in the new space may both be plotted on the same graph. The larger the differences between the subcorpora, the further apart they are on the map. Likewise, dissimilar categories of coreference features are further apart. Proximity between subcorpora and coreference features in the merged map is an approximation of the correlation between them. CA transforms the correlations between subcorpora and features in our table into a set of uncorrelated variables – principal axes or dimensions. These dimensions are computed in such a way that any subset of *k* dimensions accounts for as much variation as possible in one dimension, the first two principal axes account for as much variation as possible in two dimensions, and so on. Like this, we can identify new meaningful underlying variables, which should ideally correlate with such variables as language or register, indicating the reasons for the similarities or differences between the subcorpora. The position of the dots (subcorpora) and triangles (coreference features) indicates the relative importance of a feature for a subcorpus (see Figure 1). Moreover, the angle formed by the lines connecting the subcorpus or feature labels to the origin must be taken into account. Small angles indicate association, 90 degrees angle means no relation and angles up to 180 degrees mean negative association. The length of the lines also indicates association between subcorpora and features: the longer the line, the stronger is the association.

We use HCA (Everitt et al., 2011; Hothorn and Everitt, 2014), an unsupervised technique derived from exploratory data mining, that allows us to identify groups in the data which were not previously known. We do not prescribe what the groupings could be, since we want the algorithm to work on its own to discover all kinds of unknown patterns in the data. Specifically, we aim to see how our coreference features naturally group without applying prior knowledge of what the output groups should be. The core idea of HCA is that objects, in our case coreference features, are more related to nearby objects than to objects farther away. Coreference features are connected to form clusters based on their Euclidean distance measured here on the basis of the feature distributions (as also in the case of CA). The results are represented graphically in a dendrogram, a branching diagram that shows the relationships of similarity among a group of entities. The arrangement of the branches tells us which features are most similar to each other. We apply a technique based on bootstrap resampling, with the help of which we are able to produce *p-value*-based clusters, i.e. the ones that are highly supported by the data will have large *p-values*. For the sake of visibility our output dendrogram demonstrates AU (Approximately Unbiased) *p-value* only, which is computed by multi-scale bootstrap resampling and is a better approximation to unbiased *p-value* – indicated with red colour in Figure 2 below. The red numbers indicate the support for the split into clusters using an unbiased estimate. We draw rectangles around significant clusters (with the threshold value for p-values of 0.95). The resulting significant clusters demonstrate groups of features that are observed in our data.

5 Results

We first perform a correspondence analysis with all the features described in Section 3 above to get a general overview of how features are distributed along the lines of text types or languages. The resulting two-dimensional graph is shown in Figure 1. A detailed output of all the three CA analyses with the information on feature weights, contribution to eigenvalues, their distance to the centroid, etc. is given in Tables 3, 4, 5 in Appendix. The most obvious information we can obtain from this is that variance is most strongly pronounced between the two registers, while language contrast only marginally seems to play a role. The registers vary along dimension 1 (x-axis) explaining a very high portion (84.6%) of

[2]In PCA, either the rows or the columns would be considered.

the variation, while languages vary along dimension 2 (y-axis) explaining only 14% of variance. The outcome is not surprising, since on the one hand, feature distribution is strongly connected to register types, on the other hand, language contrast is expected to be weaker in translations due to shining through effects (i.e original text structures are kept in the target texts).

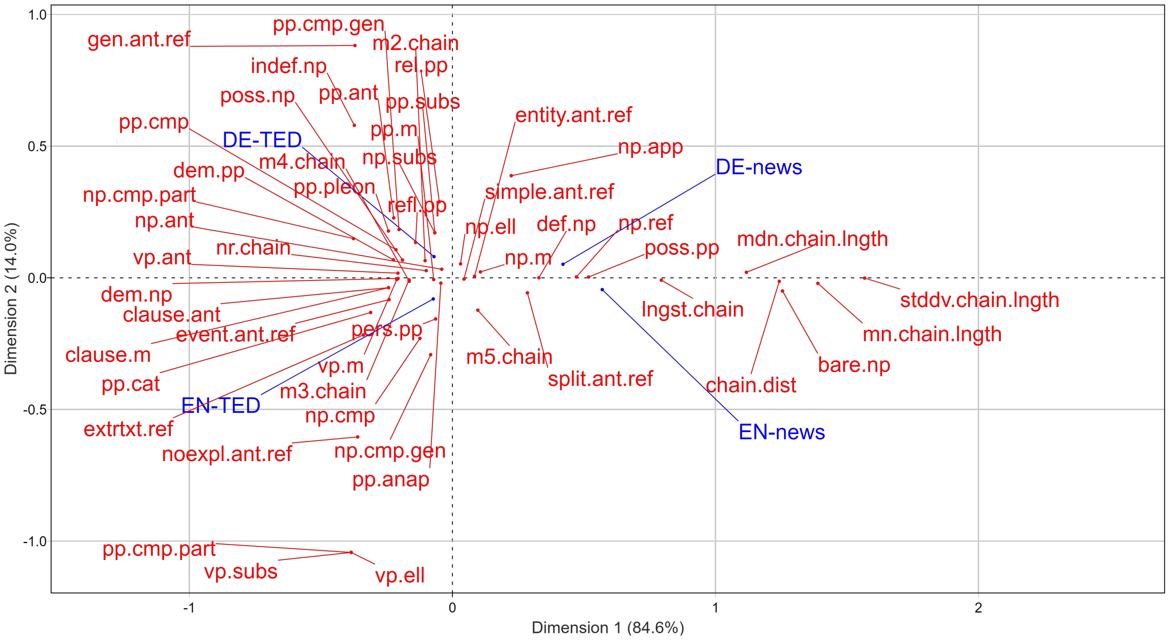

Figure 1: CA for all subcorpora with all features

In the next step, we perform clustering of features – an unsupervised analysis to explore groups of features that arise in our data. We hope that the resulting feature groups will help us to find specific features responsible for the observed register or language contrasts.

The cluster analysis (Figure 2) shows that there are two clear cut groups of features. A closer look at the distributions of the features reveals that the smaller group of features consists of the nine more general and also most common features in all subcorpora, while the other group is much more diverse including more fine-grained features, which are also less common.

Next, we perform a correspondence analysis on the two groups of features that resulted from HCA, starting with the smaller group (Figure 3). We find that the biggest part of the variation is explained by dimension 1 (x-axis, 71.3%) indicating a strong variance between the registers. The language contrast, i.e. the difference between originals and translations is seen in the second dimension (y-axis). Although its contribution is small (explaining 28.5% variation in the data), it is twice as big as the contribution to dimension 2 in Figure 1 above, showing that this first feature group reveals more differences between originals and translations than the whole group of features. The features that especially contribute to language contrast are two-member chains and personal pronouns.

Looking at the first dimension, EN news is the most distinct subcorpus, least associated with the selected features with only very weak associations with simple and entity antecedents. In the DE news texts we see a moderate association with NPs as referring expressions. This is most likely due to the high number of nominal references to a human antecedent (i.e. *Simone Biles, die Turnerin, die 19-jährige amerikanische Turnerin, die Amerikanerin, Biles, dieses Mädchen, der 19-Jährigen.*). Short, two-member chains are most strongly associated with the translated TED talks (DE-TED), as well as NPs as antecedents. These two features suggest that in the DE-TED translations, two member chains with an NP antecedent are relatively common. It is possible that some of these short chains are instances where in English there is no chain at all, for instance due to a reduced relative clause as in example (1), or other implicit references in the source text made explicit in the target text.

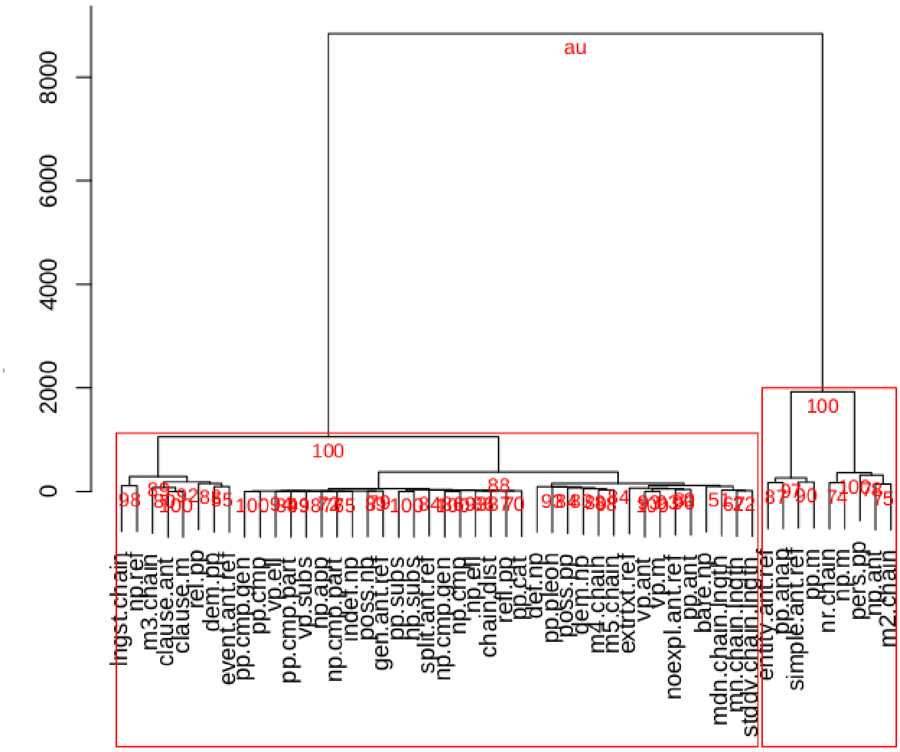

Figure 2: Clustering of all features under analysis

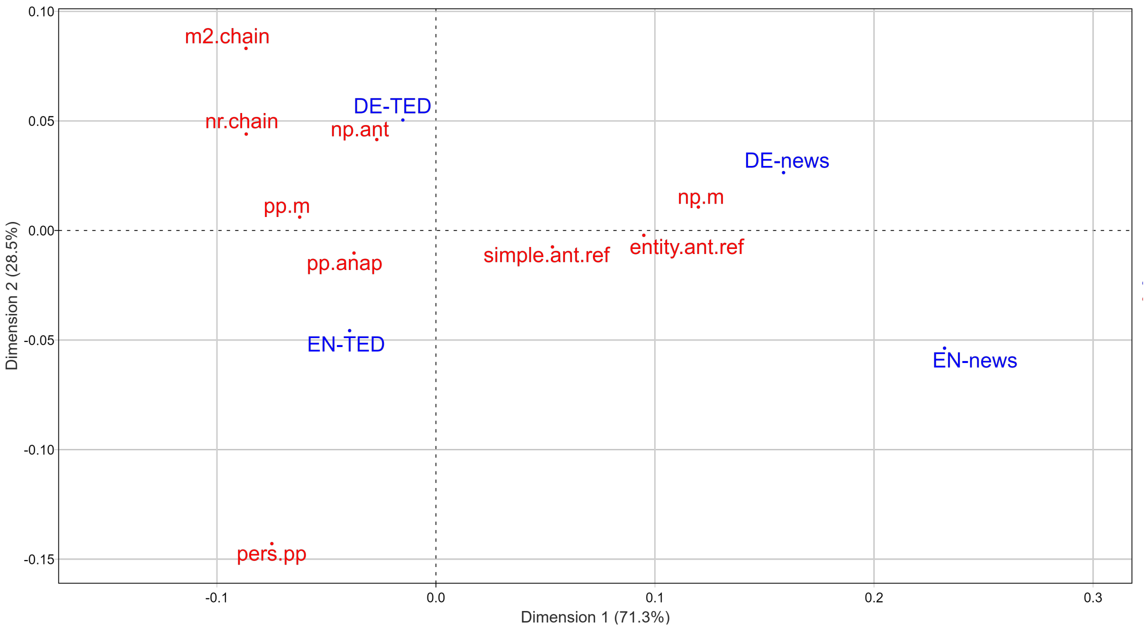

Figure 3: CA for all subcorpora with the first group of features resulting from HCA

(1) a. *Everybody talks about happiness these days. I had somebody count the number of books with "happiness" in the title published in the last five years [...].*

b. *Jeder spricht heutzutage über das Glück. Ich habe einige Leute die Anzahl [der Bücher] zählen lassen, [die] mit "Happiness" im Titel in den letzten fünf Jahren veröffentlicht wurden [...].*

Two-member chains might therefore not indicate a general tendency towards shorter chains in the German translations but rather be an indicator of explicitations of cases where the original keeps the reference implicit. The EN-TED talks show a strong association with personal pronouns, indicating frequent pronominal reference. Pronominal reference hints at a relatively low level of formality (Lapshinova-Koltunski and Martínez Martínez, 2017). The frequent use of personal pronouns in TED talks points to recurring reference to persons as the topic of the talks, see example (2) for illustration.

(2) *Now, I should mention that [Nathaniel] refuses treatment because when [he] was treated it was with shock therapy and Thorazine and handcuffs, and that scar has stayed with [him] for [his] entire life. But, as a result now, [he] is prone to these schizophrenic episodes. The worst of which can manifest themselves as [him] exploding, and then disappearing for days, wandering the streets of Skid Row, exposed to its horrors, with the torment of [his] own mind unleashed upon [him].*

The second group of features shows an even stronger variation along dimension 1 (*x*-axes), explaining 86.2% of the variation between the two registers (see Figure 4).

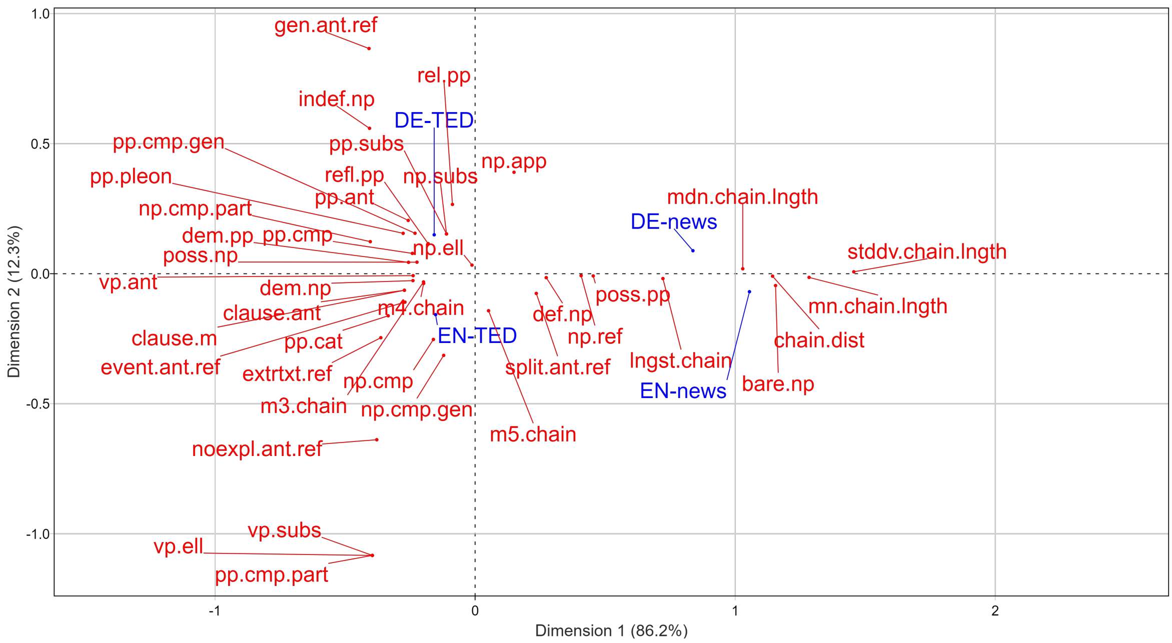

Figure 4: CA for all subcorpora with the second group of features resulting from HCA

Here, the language contrast, i.e. the difference between originals and translations is much less prominent. This indicates that the groups we detected by clustering vary in their explaining power for language contrast. While the first group explained 28% of variation between the languages, the second group explains only half of that (12.3%). The second group rather seems to represent distinctiveness between the registers. News texts are highly distinct from TED talks on this graph. Responsible features for their distinctiveness are *mdn.chain.lngth* and *stddv.chain.lngth* indicating a strong variation in chain length. Also the feature *lngst.chain* is associated with news in both languages, reflecting the fact that news texts often deal with one topic that is focused throughout the whole text. Appositions are associated with DE-news as well as DE-TED Talks. Appositions can represent explicitations, as in example (3), further defining a referent.

(3) a. *Boundless Informant is a program that the NSA hid from Congress.*
 b. *[“Boundless Informant”] [“Informant ohne Grenzen”] ist ein Programm, das die NSA vor dem Kongress verborgen hielt.*

The EN-news texts are most strongly associated with bare NPs and features related to chain length and distance. Bare NPs in English are, besides plural indefinites, mostly names. Their strong association reflects the fact that the EN-news texts tend to reiterate the names while German texts tend to reformulate and substitute proper names. In news texts the distance between mentions in a chain seems to be very long, which is especially not problematic for cohesiveness if the referent is reiterated.

The EN-TED talks show the strongest association with verbal ellipsis (see Example (4)) and substitution. These are cohesive devices which represent implicit and "sloppy" ways of creating coreference and are therefore susceptible to explicitation by a translator.

(4) a. *Nobody wants to change [how they live] just because [it] 's good for the world, or because we are [supposed to].*
b. *Niemand möchte [sein Leben ändern], weil [es] gut für die Umwelt ist oder weil wir [es] sollten.*

The same holds true for comparative reference (*np.cm.part, pp.cmp.gen*) as well associated with EN-TED talks. For extratextual reference we find a moderate association with EN-TED talks. Plausibly so, since TED talks are video talks where speakers often point at visualization material (presentations, pictures, videos or even other people present on the stage). Extratextual reference is often not retained in the target texts, since deictic reference is especially hard to match in subtitles, see example (5). The same mechanism seems to be at work with the feature "no explicit antecedent", even more strongly associated with the EN-TED talks. In cases, where a cue to a possible extratextual referent cannot be found in the text material, the annotator has the option of labelling a referring expression "no explicit". The German translators frequently avoid reference to undefined antecedents leaving them out altogether as illustrated in example (6).

(5) a. *[These] are ancient dice, made out of sheep's knuckles. Right?*
b. *[Es] gibt diese antiken Würfel, aus Schafsknöcheln. Wissen Sie?*

(6) a. *There was a case study done in 1960 's Britain, when [they] were moving from grammar schools to comprehensive schools.*
b. *In den 60er Jahren wurde in Großbritannien eine Fallstudie durchgeführt. Damals wurden Gymnasien in Gesamtschulen umgewandelt.*

The DE-TED talk translations are most strongly associated with generic and indefinite NPs. On the German translation side these distinctive features may be results of explicitation attempts, inserting a generic noun where in the English source texts there is no explicit antecedent. Also relative clauses distinctive for DE-TED texts are typical cases of explicitation. While English offers the option of a reduced variant (in object relative clauses), as well as participle (*-ing* and *-ed* clauses), in German, these options do not exist. Since TED talks often deal with more complex, scientific topics where a sound understanding by the listener/reader is essential, the feature is found more strongly related to the TED talks than to the news texts. Regarding this second group of features, the TED talks differ more from each other than the news texts. One plausible explanation might be the fact that the originally spoken texts through translation are turned into more written-like texts, which is reflected by the distinctive features respectively in the two languages.

6 Conclusion

In the present paper, we explored a number of coreference features in English-German translations that contain texts belonging to two registers. The results show that there is more variation in terms of registers or modes (spoken vs. written) than of languages, which means that English originals and their German translations differ to a lesser extent than news texts and TED talks. This confirms findings of other studies, e.g. Kunz et al. (2016) who show that variation along the dimension of mode is more prominent than that along the dimension of language. The authors show this using a set of comparable texts, whereas our data contains originals and their translations, i.e. the same texts in two languages.

By using cluster analysis we found two clusters of the features at hand, showing that the more general features represent language contrast better, while a more fine-grained classification of feature categories reflects the register or mode variation. This finding indicates that more general features reveal the differences between originals and translations, whereas a more fine-grained classification of feature categories reflects the register or mode variation. However, we were not able to discover features that would strongly indicate concrete differences between the original and the translated texts, which was one of our original goals.

In our future work, we plan to integrate further techniques, such as feature selection technique, e.g. information gain, as used by Lapshinova-Koltunski and Martínez Martínez (2017) to see which features are more informative to predict the two languages in the analysed data. Furthermore, it would be interesting to extend our analyses to the other translation direction and see if we observe the same translationese phenomena for the German-English translations. Another extension of the work would be adding translations of the same texts into further languages.

Acknowledgements

Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930. The ParCorFull annotation work was funded by the European Association for Machine Translation. We would also like to thank our reviewers for their useful comments and suggestions.

References

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. Coreference in Spoken vs. Written texts: a Corpus-based Analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 158–164, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Mira Ariel. 1990. *Accessing Noun-phrase Antecedents*. Routledge, London.

Mona Baker. 2011. *In Other Words: A Coursebook on Translation*. Taylor & Francis.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Viktor Becher. 2011. *Explicitation and Implicitation in Translation. A Corpus-based Study of English-German and German-English Translations of Business Texts*. Ph.D. thesis, Universität Hamburg.

Douglas Biber, Edward Finegan, Stig Johansson, Susan Conrad, and Geoffrey Leech. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow, 1 edition.

Shoshana Blum-Kulka. 1986. Shifts of Cohesion and Coherence in Translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication*, pages 17–35. Gunter Narr, Tübingen.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. 1989. Remarks on Plural Anaphora. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, pages 161–167. Association for Computational Linguistics.

Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

Barbara A. Fox. 1987. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge University Press, Cambridge.

Spence Green, Nicholas Andrews, Matthew R Gormley, Mark Dredze, and Christopher D Manning. 2011. Cross-lingual Coreference Resolution: A New Task for Multilingual Comparable Corpora. Technical Report 6, HLT-COE, Johns Hopkins University.

Michael J. Greenacre. 2007. *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean Projection of Coreference Chains across Languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, page 14, Beijing, China.

Yulia Grishina and Manfred Stede. 2017. Multi-source Annotation Projection of Coreference Chains: Assessing Strategies and Testing Opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 41–50. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels, October. Association for Computational Linguistics.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis, School of Informatics. University of Edinburgh.

Jeanette Gundel, Michael Hegarty, and Kaja Borthen. 2003. Cognitive Status, Information Structure, and Pronominal Reference to Clausally Introduced Entities. *Journal of Logic, Language and Information*, 12(3):281 – 299.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the IWSLT (International Workshop on Spoken Language Translation)*, pages 283–289, Paris, France.

Torsten Hothorn and Brian S. Everitt. 2014. *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC Press, 3rd edition.

Juliane House. 1997. *Translation Quality Assessment: A Model Revisited*. Tübinger Beiträge zur Linguistik. G. Narr.

Karin Königs. 2011. *Übersetzen Englisch - Deutsch. Lernen mit System*. Oldenbourg Verlag, Oldenbourg, 3 edition. vollstäädig überarbeitete Auflage.

Kerstin Kunz and Erich Steiner. 2012. Towards a Comparison of Cohesive Reference in English and German: System and Text. In M. Taboada, S. Doval Suárez, and E. González Álvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. Equinox, London.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and José Manuel Martínez Martínez. 2016. Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 23–31, San Diego, California, June. Association for Computational Linguistics.

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel, and Erich Steiner. 2017. Gecco – an Empirically-based Comparison of English-German Cohesion. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, volume 300 of *TILSM series*, pages 265–312. Mouton de Gruyter. TILSM series.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier, 2017. *Coreference Corpus Annotation Guidelines*, December.

Ekaterina Lapshinova-Koltunski and José Manuel Martínez Martínez. 2017. Statistical Insights into Cohesion: Contrasting English and German across Modes. In Markéta Janebová, Ekaterina Lapshinova-Koltunski, and Michaela Martinková, editors, *Contrasting English and Other Languages*, pages 130–163. Cambridge Scholars Publishing, Newcastle upon Tyne.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a Parallel Corpus Annotated with Full Coreference. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki, Japan, may. European Language Resources Association (ELRA).

Ekaterina Lapshinova-Koltunski, Cristina España-Bonet, and Josef van Genabith. 2019a. Analysing Coreference in Transformer Outputs. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 1–12, Hong Kong, China, November. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019b. Cross-lingual Incongruences in the Annotation of Coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 26–34, Minneapolis, USA, June. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September. Association for Computational Linguistics.

Giang Linh Nguy, Michal Novák, and Anna Nedoluzhko. 2011. Coreference Resolution in the Prague Dependency Treebank. Technical report, UFAL, Prague. Technical report.

Michael Novák and Anna Nedoluzhko. 2015. Correspondences between Czech and English Coreferential Expressions. *Discours*, 16.

Michal Novák and Zdeněk Žabokrtský. 2014. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland.

Michal Novák. 2018. A fine-grained large-scale analysis of coreference projection. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 77–86, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maciej Ogrodniczuk. 2013. Translation- and Projection-based Unsupervised Coreference Resolution for Polish. *Language Processing and Intelligent Information Systems, IIS 2013*, 7912.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Tranferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Ellen F. Prince. 1981. Towards a Taxonomy of Given-new Information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Marta Recasens and Sameer Pradhan. 2016. Evaluation Campaigns. In *Anaphora Resolution – Algorithms, Resources, and Applications*, pages 165–208.

Gideon Toury. 1995. *Descriptive Translation Studies - and Beyond*. John Benjamins Publishing Company, Benjamins edition.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. 2012. Abstract Pronominal Anaphors and Label Nouns in German and English: Selected Case Studies and Quantitative Investigations. *Translation: Computation, Corpora, Cognition*, 2(1).

A Appendix

feature	description	example
	Count features	
np.cmp.gen	general comparison NP	*a different person*
np.cmp.part	particular comparison NP	*a taller person*
pp.cmp.gen	general comparison pronoun	*other one*
pp.cmp.part	particular comparison pronoun	*bigger one*
np.ell	NP ellipsis	*I count the [neighboring balls].., the answer's always twelve [].*
vp.ell	verb ellipsis	*They knew about this. I did not [].*
def.np	NP with def. article	*the person*
dem.np	demonstrative NP	*this person*
indef.np	indefinite NP	*any person*
bare.np	NP without modifier	*person*
poss.np	possesive NP	*my sister*
dem.pp	demonstrative pronoun	*these*
pers.pp	personal pronoun	*he, she it, etc.*
poss.pp	possessive pronoun	*hers, his, its, etc.*
refl.pp	reflexive pronoun	*herself, himself*
rel.pp	relative pronoun	the person *who*
pp.subs	any pronoun expressing substitution	
vp.subs	verb substitution	*Has the plane landed? – Yes, it has [done].*
np.subs	nominal substitution	*He wants a green apple, but she wants the red [one].*
noexpl.ant.ref	reference to a non identifiable antecedent	*see example (6)*
simple.ant.ref	reference to one single referent	*He eats [peas]. They are green.*
split.ant.ref	two or more antecedents	*[Tim] likes [Tom]. They are happy.*
clause.ant	clausal antecedent	*[Tim hates Tom]. This is sad.*
np.ant	antecedent is an NP	*Tim likes [cats]. They are soft.*
pp.ant	pronominal antecedent	*Tim likes [them]. They are soft.*
vp.ant	verbal antecedent	*Tim [writes]. It is his hobby.*
entity.ant.ref	reference to an entity	*Tim likes [Tom].He is blond.*
event.ant.ref	reference to an event	*Tim loves Tom.[This] is nice.*
gen.ant.ref	reference to generic antecedent	*Pigs are clever.[They] can read.*
np.app	nominal apposition	*A friend, [Marco], got married.*
np.cmp	all cases of comparative nominal phrases	
np.ref	any reference to an NP	
pp.anap	any anaphoric pronoun	
pp.cat	cataphoric pronoun	*[She] is strong. Her name is Uma.*
pp.comp	any pronoun expressing comparative reference	
extrtxt.ref	reference to an extratextual referent	*see example (5)*
pp.pleon	pleonastic *it*	*[It]'s raining.*
clause.m	all clausal mentions	
np.m	all nominal mentions	
pp.m	all pronominal mentions	
vp.m	all verbal mentions	
nr.chain	total number of coreference chains	
m2.chain	chain with two members	
m3.chain	chain with three members	
m4.chain	chain with four members	
m5.chain	chain with five or more members	
	Other features	
lngst.chain	maximum chain length	
mn.chain.lngth	mean chain length	
mdn.chain.lngth	median chain length	
stddv.chain.lngth	standard deviation of chain length	
chain.dist	mean distance in sentences between mentions in same chain	

Table 2: Overview of all features under analysis with definitions and examples

feature	Mass	ChiDist	Inertia	Dim. 1	Dim. 2
np.cmp.gen	0.000663	0.316624	0.000066	-0.436676	-3.789908
np.cmp.part	0.000108	0.414074	0.000018	-1.985124	1.930274
pp.cmp.gen	0.000231	0.297027	0.000020	-1.072092	2.377709
pp.cmp.part	0.000015	1.127224	0.000020	-2.032161	-13.533658
np.ell	0.000678	0.064981	0.000003	0.162947	0.686875
vp.ell	0.000062	1.127224	0.000078	-2.032161	-13.533658
def.np	0.005950	0.329001	0.000644	1.734675	0.000025
dem.np	0.007630	0.211763	0.000342	-1.089966	-0.031847
indef.np	0.000139	0.713970	0.000071	-1.968138	7.514472
bare.np	0.008709	1.256376	0.013746	6.624325	-0.650892
poss.np	0.000493	0.207501	0.000021	-1.002061	0.877135
dem.pp	0.020130	0.236394	0.001125	-1.186002	0.887559
pers.pp	0.059773	0.169113	0.001709	-0.337756	-2.030630
poss.pp	0.009294	0.518437	0.002498	2.733614	0.039440
refl.pp	0.000832	0.255761	0.000054	-0.738817	1.742443
rel.pp	0.020084	0.301273	0.001823	-0.262351	3.750173
pp.subs	0.001233	0.226172	0.000063	-0.353595	2.217421
vp.subs	0.000031	1.127224	0.000039	-2.032161	-13.533658
noexpl.ant.ref	0.002435	0.707096	0.001218	-1.901654	-7.843646
simple.ant.ref	0.133726	0.044202	0.000261	0.231679	-0.062243
split.ant.ref	0.001988	0.307337	0.000188	1.505294	-0.740406
clause.ant	0.009988	0.245664	0.000603	-1.281402	-0.487098
np.ant	0.057446	0.052750	0.000160	-0.214535	0.421710
pp.ant	0.002682	0.326889	0.000287	-1.177903	2.949368
vp.ant	0.003961	0.208049	0.000171	-1.093846	0.229017
entity.ant.ref	0.117434	0.084428	0.000837	0.444521	0.077095
event.ant.ref	0.015567	0.254615	0.001009	-1.270361	-1.075932
gen.ant.ref	0.000200	0.990971	0.000197	-1.956178	11.446540
nr.chain	0.070686	0.103485	0.000757	-0.526150	0.353133
lngst.chain	0.003160	0.807960	0.002063	4.192428	-0.117235
mn.chain.lngth	0.000215	1.389670	0.000415	7.334646	-0.275031
mdn.chain.lngth	0.000123	1.122915	0.000155	5.899956	0.270630
stddv.chain.lngth	0.000248	1.568046	0.000610	8.271408	-0.013635
m2.chain	0.045546	0.124456	0.000705	-0.548020	0.845941
m3.chain	0.012762	0.166895	0.000355	-0.869407	-0.101058
m4.chain	0.005379	0.165987	0.000148	-0.867163	-0.172977
m5.chain	0.006982	0.157495	0.000173	0.509968	-1.595385
chain.dist	0.000074	1.242745	0.000115	6.558996	-0.176087
np.app	0.000262	0.739706	0.000143	1.180567	5.032233
np.cmp	0.000771	0.276492	0.000059	-0.653459	-2.989083
np.ref	0.023351	0.472276	0.005208	2.487352	0.042190
pp.anap	0.110683	0.049181	0.000268	-0.232044	-0.262816
pp.cat	0.001187	0.342668	0.000139	-1.641004	-1.711176
pp.cmp	0.000247	0.267869	0.000018	-1.132096	1.383249
extrtxt.ref	0.003591	0.402488	0.000582	-1.787352	-2.817923
np.subs	0.001233	0.226172	0.000063	-0.353595	2.217421
pp.pleon	0.007737	0.315157	0.000769	-1.284026	2.307029
clause.m	0.009988	0.245664	0.000603	-1.281402	-0.487098
np.m	0.082631	0.109066	0.000983	0.562322	0.291150
pp.m	0.127607	0.071472	0.000652	-0.375570	-0.088586
vp.m	0.004054	0.211373	0.000181	-1.115253	-0.084960

Table 3: Output of CA with all features: masses or weights of features (Mass), chi-squared distances of feature points to the centroid, i.e. their average (ChiDist), feature contribution to principal inertias or eigenvalues (Inertia) and standard coordinates in Dimension 1 (Dim1) and Dimension 2 (Dim2))

feature	Mass	ChiDist	Inertia	Dim. 1	Dim. 2
pers.pp	0.074203	0.161340	0.001932	-0.999866	-3.018902
simple.ant.ref	0.166010	0.053775	0.000480	0.710814	-0.158291
np.ant	0.071314	0.049546	0.000175	-0.360819	0.875915
entity.ant.ref	0.145785	0.094975	0.001315	1.267290	-0.047300
nr.chain	0.087750	0.097184	0.000829	-1.155579	0.929273
m2.chain	0.056542	0.120464	0.000821	-1.157407	1.756098
pp.anap	0.137404	0.039039	0.000209	-0.498657	-0.217014
np.m	0.102579	0.120281	0.001484	1.599027	0.226883
pp.m	0.158413	0.062801	0.000625	-0.830071	0.128954

Table 4: Output of CA with the 1st group of features: masses or weights of features (Mass), chi-squared distances of feature points to the centroid, i.e. their average (ChiDist), feature contribution to principal inertias or eigenvalues (Inertia) and standard coordinates in Dimension 1 (Dim1) and Dimension 2 (Dim2))

feature	Mass	ChiDist	Inertia	Dim. 1	Dim. 2
np.cmp.gen	0.003481	0.345312	0.000415	-0.314363	-2.160139
np.cmp.part	0.000567	0.427257	0.000103	-1.048137	0.847752
pp.cmp.gen	0.001214	0.304792	0.000113	-0.600798	1.069651
pp.cmp.part	0.000081	1.167677	0.000110	-1.027334	-7.448581
np.ell	0.003562	0.036934	0.000005	-0.031473	0.227877
vp.ell	0.000324	1.167677	0.000441	-1.027334	-7.448581
def.np	0.031245	0.274728	0.002358	0.711096	-0.104608
dem.np	0.040068	0.242720	0.002361	-0.621538	-0.182838
indef.np	0.000729	0.709433	0.000367	-1.055649	3.843650
bare.np	0.045735	1.156825	0.061204	3.003787	-0.313568
poss.np	0.002590	0.230630	0.000138	-0.581516	0.304382
dem.pp	0.105716	0.260101	0.007152	-0.665836	0.300682
poss.pp	0.048811	0.455132	0.010111	1.181082	-0.059239
refl.pp	0.004371	0.259922	0.000295	-0.469710	0.795526
rel.pp	0.105473	0.290837	0.008922	-0.226358	1.831489
pp.subs	0.006476	0.220449	0.000315	-0.286092	1.051881
vp.subs	0.000162	1.167677	0.000221	-1.027334	-7.448581
noexpl.ant.ref	0.012790	0.745846	0.007115	-0.983296	-4.389810
split.ant.ref	0.010442	0.267015	0.000744	0.611908	-0.521618
clause.ant	0.052453	0.279416	0.004095	-0.706994	-0.438857
pp.ant	0.014085	0.333571	0.001567	-0.668673	1.408843
vp.ant	0.020803	0.238604	0.001184	-0.619869	-0.051979
event.ant.ref	0.081756	0.291118	0.006929	-0.701308	-0.752017
gen.ant.ref	0.001052	0.983372	0.001018	-1.060938	5.953188
lngst.chain	0.016594	0.736247	0.008995	1.877954	-0.129035
mn.chain.lngth	0.001128	1.284371	0.001861	3.338840	-0.098603
mdn.chain.lngth	0.000648	1.033924	0.000692	2.675730	0.130440
stddv.chain.lngth	0.001303	1.456807	0.002766	3.785737	0.049886
m3.chain	0.067024	0.201356	0.002717	-0.515901	-0.218129
m4.chain	0.028250	0.201405	0.001146	-0.514197	-0.257608
m5.chain	0.036669	0.151949	0.000847	0.134401	-0.979939
chain.dist	0.000391	1.144067	0.000512	2.973885	-0.065053
np.app	0.001376	0.692390	0.000660	0.389001	2.682152
np.cmp	0.004047	0.308970	0.000386	-0.417091	-1.739034
np.ref	0.122634	0.408645	0.020479	1.060135	-0.052987
pp.cat	0.006233	0.377442	0.000888	-0.868716	-1.112941
pp.cmp	0.001295	0.283857	0.000104	-0.627456	0.537262
extrtxt.ref	0.018860	0.438271	0.003623	-0.942805	-1.692524
np.subs	0.006476	0.220449	0.000315	-0.286092	1.051881
pp.pleon	0.040635	0.325911	0.004316	-0.719559	1.067513
clause.m	0.052453	0.279416	0.004095	-0.706994	-0.438857

Table 5: Output of CA with the 2nd group of features: masses or weights of features (Mass), chi-squared distances of feature points to the centroid, i.e. their average (ChiDist), feature contribution to principal inertias or eigenvalues (Inertia) and standard coordinates in Dimension 1 (Dim1) and Dimension 2 (Dim2))