Exploring Coreference Features in Heterogeneous Data

Ekaterina Lapshinova-Koltunski
Saarland University and
University of Hildesheim
Saarbrücken Campus A2.2 Germany
e.lapshinova@mx.uni-saarland.de

Kerstin Anna Kunz
Heidelberg University
Plöck 57a, 69117 Heidelberg
kerstin.kunz@
iued.uni-heidelberg.de

Abstract

The present paper focuses on variation phenomena in coreference chains. We address the hypothesis that the degree of structural variation between chain elements depends on language-specific constraints and preferences and, even more, on the communicative situation of language production. We define coreference features that also include reference to abstract entities and events. These features are inspired through several sources – cognitive parameters, pragmatic factors and typological status. We pay attention to the distributions of these features in a dataset containing English and German texts of spoken and written discourse mode, which can be classified into seven different registers. We apply text classification and feature selection to find out how these variational dimensions (language, mode and register) impact on coreference features. Knowledge on the variation under analysis is valuable for contrastive linguistics, translation studies and multilingual natural language processing (NLP), e.g. machine translation or cross-lingual coreference resolution.

1 Introduction

The way in which coreference is realised in texts is governed by the mode of production, by typical contexts of situation and by language peculiarities. In this study, we are particularly concerned with coreference variation as a result of these three influencing factors. The production and reception of referring expressions in naturally occurring discourse is a reflection of discourse mode (spoken vs. written discourse, see Kibrik, 2011, 11). Another influence is exerted by discourse genres or registers[1] that correspond to standard configurations of communicative topics, goals and speaker interaction, typical of particular discourse communities. We know from register and genre studies (for instance Biber, 2012, 33) that register differences can be observed at all linguistic levels and be deduced from lexico-grammatical features. The production and reception of referring expressions is governed by language-specific factors, as coreference relations in different languages vary considerably in the range of linguistic means triggering these relations (Kunz and Steiner, 2012; Kunz and Lapshinova-Koltunski, 2015; Novák and Nedoluzhko, 2015). Moreover, there are language-specific preferences for using particular means over others.

Variational dimensions such as mode, register and language influence the choice and the frequency of referential expressions in language use. We therefore need to know how this influence is reflected in the kinds of coreference phenomena, their internal organisation (structure) and in their interplay with other related phenomena. Apart from answering linguistically motivated contrastive questions, this knowledge is also beneficial to the area of natural language processing, i.e. when designing features for coreference resolution tasks in multilingual heterogeneous data. The importance of the information on this variation is known, as for instance, the CoNLL-2012 shared task on coreference resolution included multiple languages, modes and registers within OntoNotes (Recasens and Pradhan, 2016). Information on language-driven variational mechanisms in coreference is valuable for multilingual coreference resolution systems (Rahman and Ng, 2012; Pradhan et al., 2012; Recasens et al., 2010; Harabagiu and Maiorano, 2000). Kübler and Zhekova (2016) describe difficulties and challenges of this task showing that many issues remain unsolved in multilingual coreference resolution. In coreference projection, when the annotation of coreference chains in a source lan-

[1] We prefer to use the term 'register' instead of 'genre', as register reflects functional variation of a language, whereas 'genre' rather refers to the cultural belonging of a text.

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 53–64

https://doi.org/10.18653/v1/P17

guage is projected onto a target language (Novák et al., 2017; Grishina and Stede, 2015; Yarowsky et al., 2001), non-equivalences resulting from language contrasts cause numerous errors. Knowledge on the register and on mode differences is also useful for coreference resolution that requires domain adaptation (Rösiger and Teufel, 2014; Uryupina and Poesio, 2012; Yang et al., 2012; Apostolova et al., 2012). There are studies showing that register and mode impact on anaphora prediction models (see e.g. Zeldes, 2018).

In this paper, we define a number of coreference features that are inspired through several sources – cognitive parameters, pragmatic factors and typological status. We pay attention to the distributions of these features in a dataset containing English and German texts that belong to two different discourse modes (spoken and written) and can be classified into seven different registers (academic speeches, political essays, general interviews, literature, technical manuals, popular science and texts from company websites). As our main goal is to find out how these variational dimensions (language, mode and register) impact on coreference features, we apply data mining techniques focusing on the following research questions (RQs):

RQ1 Which coreference features are most informative in the three prediction tasks: (a) language, (b) mode, (3) register?

RQ2 Which parameters are distinctive for the languages, modes and registers under analysis?

2 Theoretical Background

In our study, coreference includes cohesive relations of identity, i.e. relations between coreferring expressions in a text pointing to the same extra-linguistic referent. This is illustrated in example (1).

(1) *... what relativity is really about, is the question of what two different people, in motion with respect to another, relative to one another, when they look at something happening, or they measure something, the distance between two points or the time between two events, the question is what do these two guys get, if they're in relative motion...*

The first referring expression in the text is the `antecedent` (*two different people*) introducing the referent into the textual world. We account for antecedents referring to referents such as persons, objects, times and locations, and also to more complex semantic concepts such as actions or processes, as in example (2), or facts and events. Concepts relating to persons or objects are often expressed by simpler linguistic structures such as noun phrases, while complex concepts are typically reflected by less condensed structures such as sentences and even larger stretches of text and may therefore also function as antecedents in coreference chains. These are also included in our analysis.

All other subsequent expressions, referring to the same referent are `anaphors` – explicit linguistic triggers indicating an anaphoric relation to another stretch of text. They include personal and demonstrative pronouns (*it* and *this*), cohesive adverbs of place and time (e.g. *here*, *then*) and pronominal adverbs (e.g. *herewith* in English or *damit* in German), which are especially frequent in German. They all function as anaphoric heads. Moreover we include possessive and demonstrative determiners (*these* in *these two guys* in example (1)) and the definite article, functioning as modifiers within the anaphor. The antecedent and all subsequent anaphors pointing to the same referent occur in a `coreference chain`.

In our study, we account for variation of form and structure of coreferring expressions, their grammatical function and syntactic position, as well as variation with respect to the chain relation. Most studies on (automatic) anaphora resolution are based on the assumption that the reasons for differences in form, grammatical function and position of coreferring expressions in one and the same chain are related to differences in the degree of accessibility, givenness or salience that a referent has in the recipient's mind at a given point (Ariel, 2001; Prince, 1981; Gundel et al., 2003; Grosz et al., 1995; Eckert and Strube, 2000, among others). For instance, coreferring expressions that are realised as pronouns and occur as subjects in sentence-initial position typically signal a high degree of accessibility, whereas full lexical phrases that are non-subjects at sentence-final position typically reflect a lower degree of accessibility. Furthermore, the accessibility of a referent is related to chain features (e.g. Eckert and Strube, 2000): low distance in long coreference chains together with a low general number of different coreference chains is related to a high degree of accessibility.

Our main interest is functional variation of coreference that mainly stems from three variables of language use – mode of production, register variation and language contrast. We are aware of the fact that these variables interact with the general principles of cognitive processing mentioned above. However, the reflection of the cognitive status in coreference variation itself is not the focus of the current paper.

As mentioned above, the range of available and preferred linguistic structures for realising coreference chains differs across languages (Kunz and Steiner, 2012; Kunz and Lapshinova-Koltunski, 2015; Novák and Nedoluzhko, 2015). The two languages under analysis differ in the linguistic forms available to signal coreference: German has more fine-grained options for differentiating degrees of accessibility, such as pronominal adverbs or demonstrative articles, whereas the English language system provides less syntactic flexibility and is more restricted than German in the distribution of accessible or less accessible referents. Besides that, English prefers more lexical means for establishing cohesion, whereas German tends to use more grammatical means of coreference (Kunz et al., 2017). German also seems to tend towards explicitating coreference relations, especially by using more demonstratives than English. We therefore argue that English and German differ in how correferring expressions vary in their form, syntactic function and position if looking inside coreference chains. For instance, frequent alternations in the use of demonstrative and personal pronouns are common in German, whereas in English, the form of the anaphor generally does not often change.This is illustrated in examples (2) and (3).

(2) *We work for prosperity and opportunity because they're right. It's the right thing to do.*

(3) *Wir arbeiten für Wohlstand und Chancen, weil das richtig ist. Wir tun damit das Richtige. ("We work for prosperity and opportunity because that is right. We do thereby the right").*

In the English example, the personal pronoun *they* refers to the entities *prosperity and opportunity*, and the personal pronoun *It* – to the event *working for prosperity and opportunity*. In the translation into German, the demonstrative *das* and the pronominal adverbial *damit* refer to the event *working for prosperity and opportunity* in both cases. In the second case, an additional logico-semantic relation of instrument is encoded implying a change in terms of form, grammatical function as well as position of the anaphor.

High or low variation in the use of different coreference expressions in texts may not only by subject to language contrasts but may also be a reflection of register or/ and the type of language production:

(4) *I live in a town called called Reigate. It's between London and the countryside which is quite nice. It takes us about 25 minutes to get to London on the train. It's I say it's a town, it's more of a village. It's quite small. It's very nice actually, it's a nice place to live.*

Example (4) is an extract from our spoken register INTERVIEW. It not only shows no variation at all in terms of the form of anaphors used in one coreference chain but concerning their syntactic function and position. Moreover, high thematic continuity is reflected by a long coreference chain with small distance between all elements in the coreference chain. These features used in combination typically reflect spontaneous spoken language involving dialogue between at least to speech participants. Much more variation can be expected in particular written language registers of our corpus.

3 Feature Categories under Analysis

Our coreference features can be classified into several groups[2]. The first group (features 1-24) includes features that are related to the form, to functional and structural properties of coreferring expressions – categories motivated by various pragmatic factors.

1-5. Subtypes of antecedents: nominal phrases (**ant-np**), pronouns (**ant-pron**), fact sentences (**ant-fact-s**), verbal phrases representing events (**ant-event-vp**) and other structurally more complex segments such as complex sentences or paragraphs(**other**). This classification is based on the scope of the coreference relation: the distinction between entities and events / states is reflected in the distinction between nominal and verbal expressions (Kibrik, 2011, 7). Since languages, modes and registers show variation in terms of nominal

[2]Note that we count the total number of items per category instead of a boolean feature normally used in a coreference resolution system

and verbal expressions, we also expect that the scope of the coreference relation may vary depending on contextual influence.

6. ante-ttr The feature reflecting antecedent variability – 'type-token-ratio' of antecedents per text. We measure variability of antecedents – their structural complexity, i.e. pronouns, nominal phrases, event verbal phrases, fact sentences or bigger elements occurring as antecedents per text.

7-13. Morpho-syntactic subtypes of anaphors: personal pronoun *it* (**ana-pers-it**)[3], other third person personal pronouns (**ana-pers-head**), possessive pronouns triggering cohesiveness of the whole nominal phrase (**ana-pers-mod**), demonstrative pronouns such as *this* and *that* used as nominal heads (**ana-dem-head**), demonstrative pronominal adverbs, such as *hereby*, *herewith*(**ana-dem-pronadv**), definite articles triggering cohesiveness of the whole nominal phrase (**ana-dem-art**), demonstrative modifying pronouns triggering cohesiveness of the whole nominal phrase (e.g. *this* and *these*, as in *this project/ these projects* (**ana-dem-mod**). This classification is based on a two-fold motivation: On the one hand, it partly reflects the *Givenness Hierarchy* (Gundel et al., 1993, 275). On the other hand, this is related to the levels of *explicitness* of coreferential expressions proposed by Becher (2011) who distinguishes three degrees (low, medium and high) of explicitness that rise with the information provided by the referring element. This also goes along with the concept of Accessibility of cohesive referents by Ariel (1990) – a suitable means to measure coreferential explicitness, although Ariel (1990) does not make use of the term explicitness in her work.

14-15. Subtypes of anaphors referring to location (**ana-dem-local**) and time (**ana-dem-temp**). This is motivated by the fact that time and location are often conceptualised as referents in human languages (Kibrik, 2011). This kind of referent is captured by our classification of anaphor forms only.

16-17. Subtypes of comparative reference indicating the level of their specificity: general (**ana-comp-general**) and particular (**ana-comp-partic**). We here follow Halliday and Hasan (1976, 78) who argue that comparison (in terms of likeness or unlikeness) is a form of reference as likeness is referential property. General comparison refers to general likeness, expressed by adjectives such as *same, similar, other*. Particular comparison concerns comparability between discourse units in terms of quantity (e.g. *more, fewer*) or quality (expressed by comparative adjectives and adverbs).

18-21. Grammatical functions of antecedents and anaphors: antecedent as a subject (**ant-subj**), antecedent as an object (**ant-obj**), anaphor as a subject (**ana-subj**), anaphor as an object (**ana-obj**). Grammatical functions were often used as a parameter of discourse salience in coreference resolution systems (Lappin and Leas, 1994; Mitkov et al., 2002; Klenner and Tuggener, 2011).

22-24. Total number of mentions: **mention** includes the total number of anaphors and antecedents, **anaphor** accounts for the total number of anaphors and **antecedent** for the total number of antecedents, respectively.

The other feature group is related to the properties of chains and includes the following categories.

25. Length of coreference chains measured in the number of chain elements within one chain (**length**), reflecting how coreference chains contribute to thematic continuity in a text - the longer the chain, the more continuity is explicitly expressed by cohesion.

26. Total number of coreference chains (**chain**). Higher frequencies of different chains per text reflect thematic progression or thematic variation in a text as opposed to continuity.

27. Distance between chain members within a coreference chain measured by tokens (**dist-t**). A similar feature was used by Aone and Bennett (1996) who included distance between anaphor and antecedent into their feature set[4].

28. Number of anaphors per chain that occur at sentence-initial position (**ana-p-start**).

29. Number of anaphors per chain that have a subject function (differs from ana-subj which indicates total number of subjects as anaphors) (**ana-is-subj**).

We also include features reflecting structural variation in chains measured by switch rates, as

[3] We include the pronoun *it/es* in a separate category, as it is ambiguous and semantically very vague, both in English and in German.

[4] Here, there is again a difference to features normally used by coreference systems, where distance is computed for a given mention pair.

well as variation in terms of parallel constructions and structural complexity of antecedents. The switch rates are calculated for the members of a chain in linear order. In example (1), there is a coreference chain of five members (*two different people – they – they – these two guys – they*). If the corresponding property of the first anaphor is the same as the second (in both cases *they*), there is no switch. If the property is different, as between the second and the third anaphor (*they* vs. *these two guys* in terms of form – personal pronoun vs. nominal phrase modified by a demonstrative), we observe a switch. The switch rate ($srate$) is calculated using Formula (1) where N_s is the number of switches and N_e the total number of elements in a chain.

$$srate = \frac{N_s}{N_e} \quad (1)$$

30. Variation in the sentence position of the coreferring expression (**srate1**): sentence-initial vs. other positions – e.g. $srate1$ for the chain in example (4) equals 0, as we have no chain members at sentence start[5].

31. Variation in grammatical function (**srate2**): subjects vs. other functions. Both $srate1$ and $srate2$ are supposed to reflect variation in the degree of accessibility of coreferring expressions – the higher the observed values, the more variation and less standardisation we observe.

32. Variation in the form of anaphors (**srate3**): $srate2$ amounts to 0.5 in example (1), as there are two switches between *they* and *these two guys*, and *these two guys* and *they*.

33-34. Parallelism (**srate4.1** and **srate4.2**): Based on (Mitkov et al., 2002, 4), who used parallelism in the syntactic role of the nominal verb complements. If the property of all the mentions in a chain is the same, the chain is considered to be parallel and $srate$ equals 0. Any difference in the properties of a chain member make a chain non-parallel. The proportion of chains being parallel and non-parallel is calculated for $srate1$ – chains in which all mentions occur at sentence-initial vs. non-initial position ($srate4.1$), and for $srate2$ – all mentions in a corresponding chain function as subjects vs. non-subjects in a sentence ($srate4.2$). This is related to the general principles of priming and information distribution.

[5]Please note that this feature is calculated for coreference relations beyond sentence borders

4 Data and Methods

4.1 Data

Since our main goal does not include automatic coreference resolution and we are interested in exploring different coreference preferences in heterogeneous data, we decided for a manually-annotated corpus of English and German comparable texts (EO and GO) that represent a variety of different registers representing both spoken and written discourse. We therefore use the corpus GECCo annotated for various cohesive devices, including coreference chains. The texts in the data represent seven different registers: five written and two spoken, see Table 1,. The written part was extracted from the corpus described by Hansen-Schirra et al. (2012) and contains popular-scientific articles (POPSCI), political essays (ESSAY), technical manuals (INSTR), texts from company websites (WEB) and fictional texts (FICTION). The latter register is considered to be at the borderline between written and spoken discourse, as it contains dialogues. The spoken part was extracted from the corpus described by Lapshinova-Koltunski et al. (2012) and includes academic speeches (ACADEMIC) and transcribed interviews on general topics (INTERVIEW)[6].

	EO		GO	
register	text	token	text	token
ACADEMIC	10	40,559	10	43,703
ESSAY	29	34,998	23	35,668
FICTION	10	36,996	10	36,778
INSTR	10	36,167	14	36,880
INTERVIEW	12	37,898	14	40,198
POPSCI	11	35,148	10	36,177
WEB	12	36,119	13	35,779
TOTAL	94	257,885	94	265,183

Table 1: Information on the corpus size.

The corpus contains annotations of various categories of textual cohesion elaborated for a multilingual dataset. They provide uniform coreference annotations capturing different types and subtypes of coreferring expressions existing in English and German. We select those corresponding to our features 7–15 described in Section 3 above. Besides that, the corpus contains various categories of antecedents reflecting their structural complexity that

[6]More information about the corpus and how to gain access to it can be found at http://fedora.clarin-d.uni-saarland.de/gecco.

correspond to our features 1–5 described above. An overview of the anaphor and antecedent types annotated in GECCo are provided along with language illustrations in both languages in the Appendix.

4.2 Methods

For RQ1, we use a feature selection procedure, which is normally applied to automatically select attributes relevant to the predictive modeling problem (prediction of a class membership). We use Information Gain (IG) to reduce the number of the analysed coreference features to those relevant for a concrete prediction task – to see which coreference features are especially informative if we deal with the data that is influenced by different variation dimensions: (I) languages, (II) modes, (III) registers. IG measures the expected reduction in entropy – uncertainty associated with a random feature (Roobaert et al., 2006, 464–465), or in other words, the feature's contribution to reduce the entropy.

To answer the second research question, we apply text classification with Support Vector Machines (SVM, cf. Vapnik and Chervonenkis, 1974; Joachims, 1998) with a linear kernel to answer the the second research question. We label our data with the information on classes represented in our case by (I) languages, (II) modes and (III) registers, collect the information on the frequencies of cohesive categories from our corpus, and see if our corpus data support these classes. We apply separate binary classification tasks for languages, modes and registers. In case of both languages and modes, we have two classes only: English vs. German and spoken vs. written. However, we have a multi-class task in case of registers, as our dataset contains seven different registers. For this, we use a pairwise classification, i.e. one-versus-one classifiers are built for register distinction: ESSAY vs. FICTION, ESSAY vs. INSTR, etc. The performance scores of classifiers are judged in terms of precision, recall and f-measure. They are class-specific and indicate the results of automatic assignment of class labels to certain texts. Afterwards, we inspect in detail the whole range of features that make the pre-defined classes distinct from one another. For this, the SVM weights (representing the hyperplane and corresponding to the support vectors) are judged – the magnitude of the weights provides the information on the importance of each feature: the higher the weight of a feature, the more distinctive it is for a particular class in the respective classification task.

5 Analyses and Results

5.1 RQ1: Distinctive feature selection

We use 188 instances (text-based) and start with 34 attributes. Our prediction tasks with IG depend on the dimension of variation under analysis as described above. In the task for language prediction (I), where we need to select the features from our dataset that are most informative in the distinction between English and German texts – the algorithm delivers 10 attributes. In the mode prediction task (II), the algorithm delivers 21 attributes features that are most informative in the distinction between spoken vs. written texts. And in the register prediction task (III), we receive a list of 24 features that are most informative in the distinction between several classes of registers: academic speeches vs fiction or essays, etc. In this last prediction task, almost all the features have higher scores if we compare them to the output of the two previous scenarios. We provide the resulting lists of features in Appendix, where the selected features are ranked according to their IG score.

Interestingly, four of the 34 features (**ana-pers-it**, **ana-dem-local**, **ant-fact-s**, **ana-obj**) are informative in all the three prediction tasks – the first one is related to the anaphor form and thus givenness/salience/accessibility parameters, the second and the third describe the nature of the referent (fact and location) and the last one is also associated with givenness/salience/accessibility. Apart from these, language and mode prediction scenarios share one feature only that reflects the nature of the referent represented here by time (**ana-dem-temporal**). Language and register do not share any features, whereas mode and register prediction task share 14 features, which is more than a half of the selected features in both tasks. This is not surprising as both mode and register are related to contextual, i.e. functional variation.

The features that are informative for the language prediction only include reference via pronominal adverbs (**ana-dem-pronadv**), both comparative subtypes (**ana-comp-partic and ana-comp-general**), parallelism in grammatical function (**sr4.2**) and the number of anaphors per chain that have a subject function (**ana-is-subj**). They are attributed to existing language contrasts (extensive use of pronominal adverbs in German, less

flexible word order in English and others) and can be used for a language prediction task regardless of the mode and the register the texts belong to. In a multilingual coreference resolution task, the features reflecting language contrast should be used with caution, as they might be confounding if used for both languages involved.

The features that are informative for the distinction of modes only are related to the position in the sentence (**sr1** and **ana-p-start**). They can be attributed to the specific speech conditions such as constraints on working memory capacity, spontaneous and partially unreflected text production. We assume that such features should be excluded from a feature set, when a coreference system is trained on a dataset containing both spoken and written texts.

Grammatical function of antecedent (**ant-subj** and **ant-obj**), anaphors in the subject role (**ana-subj**), demonstrative modifiers triggering coreference (**ana-dem-mod**), pronominal antecedents (**ant-pron**) and the distance between the chain members (**dist_t**) are informative if we are predicting the register a text belongs to. They are related to the contextual parameters that may vary across registers, e.g. thematic progression or degree of accessibility of referents, and may also have something to do with textual functions. This kind of features could be confounding, when a coreference resolution system is trained on a dataset of texts from different registers.

5.2 RQ2: Automatic classification

(I) Language We start with the prediction for languages between two classes – English and German. As the size of our dataset is small, we evaluate the performance of the classifier in a 10-fold cross-validation step. We judge the performance scores in terms of precision, recall and f-measure. These scores are class (in our case, language) -specific and indicate the results of automatic assignment of language labels to certain texts in our data. The results of the classification performance are presented in Table 2.

	Precision	Recall	F
EO	88.7	100.0	94.0
GO	100.0	87.2	93.2
Weight.av.	94.3	93.6	93.6

Table 2: Classification results for language distinction.

Overall, we achieve a good classification result (93.62% of accuracy with an f-measure of 93.6%) predicting between English and German texts on the basis of coreference features. This confirms that coreference phenomena have language-specific properties. All the English texts in our data were assigned with the correct labels which consequently contributes to 100% of recall for EO and 100% of precision for GO. The confusion matrix reveals that 12 German texts were erroneously classified as being English.

EO	GO
ana-pers-it	ana-dem-pronadv
ana-comp-general	ana-dem-local
ana-comp-partic	ant-fact-s
ana-is-subj	ana-dem-temp

Table 3: Class-specific features for languages.

In Table 3, we list the top four distinctive features for English and German. The most prominent feature in English is coreference via the personal pronoun *it*, whereas pronominal adverbs are the most distinctive features in German. German pronominal adverbs can function as referring expressions or establish a conjunctive relation. Interestingly, the antecedent-related features such as extended referents (clauses or sentences) and reference to location and timeare distinctive for German only.

(II) Mode The same analysis steps are performed for the differentiation between spoken and written modes. The dataset is bilingual – we do not separate them according to their languages, as our task is to predict modes regardless of the language (language-independent classification)[7]. The results for the mode prediction (presented in Table 4) are better than those for the language prediction, as we achieve 96.81% of accuracy here (with an f-measure of 96.8%). Overall, mode prediction works better for the written texts. However, spoken texts are classified with better precision (97.6% vs. 96.6%).

In Table 5, we list the top four distinctive features for the prediction of spoken vs. written mode. Both lists contain features related to the sentence-initial position of chain members (**ana-p-start** and **srate1**). The first position in the list of distinctive spoken features is occupied by demonstrative func-

[7]This means that texts labelled as 'spoken' are in both English and German

	Precision	Recall	F
spoken	97.6	89.1	93.2
written	96.6	99.3	97.9
Weight.av.	96.8	96.8	96.8

Table 4: Classification results for mode distinction.

Spoken	Written
ana-dem-head	ana-obj
ana-pers-it	ana-pers-mod
ana-dem-local	srate1
ana-p-start	antecedent

Table 5: Class-specific features for modes.

tioning as heads in texts, e.g. *dies/this*, followed by *es/it* in the same function.

(III) Register Here, we also perform classification on the bilingual dataset, as we did in the previous task. The results of the classification performance are presented in Table 6. This prediction task delivers the least satisfactory results, which is not unexpected, since we have here a multi-class task with a smaller number of items. However, the overall result is 90.88% of accuracy (with an f-measure of 66.1%). The best result was achieved for fictional texts and academic speeches, whereas the lowest scores were observed for websites and technical manuals.

	Precision	Recall	F
ESSAY	59.5	96.2	73.5
FICTION	85.0	85.0	85.0
INSTR	69.2	37.5	48.6
POPSCI	61.1	52.4	56.4
WEB	53.8	28.0	36.8
ACADEMIC	92.9	65.0	76.5
INTERVIEW	80.8	80.8	80.8
Weight.av.	69.4	68.1	66.1

Table 6: Classification results for register distinction.

We suggest that the registers whose texts are not misclassified possess very strong coreference features that distinguish them from other texts. This means that when building systems for coreference resolution, register adaptation for these registers is essential.

Most misclassified texts of various registers were labelled as ESSAY (34)[8]. Most distinctive features of this register seem to be shared by other registers resulting in the high number of noisy texts in the ESSAY class. While erroneous assignment of 'foreign' classes is typical for ESSAY, ACADEMIC seems to be very different from all other register classes, with one exception of an interview text. Therefore, we decide to analyse the top distinctive features of these two registers in detail.

In Table 7, we summarise the top five features distinctive for ESSAY and for ACADEMIC, if classified against the other six registers. As seen from the table, the features of ESSAY are related to the distance between chain elements and the properties of antecedents: variation in the scope of relation and the subject/object function, which is a salience feature. In the prediction between FICTION and ESSAY, only two features turned out to be distinctive. The longest list was observed in the prediction between ESSAY and INSTR.

Most features in ACADEMIC are related either to the form of anaphors or the antecedent types Here, we observe a preference for events or states. Overall, the ESSAY features are more diverse in their categories. Moreover, the ACADEMIC lists are longer: the longest one contains 19 members (in the prediction between ACADEMIC and ESSAY).

6 Conclusion and Discussion

We used a set of coreference features of cognitive, pragmatic and typological nature to analyse their variation in heterogeneous data – texts that belong to two different languages, spoken and written modes classified into seven different registers. We used different methods to find out in which way the three variational dimensions that are present in our dataset (language, mode and register) influence the constellation of features.

The results show that depending on the variational dimension, we can have different sets of coreference features. The information on the nature of features derived from our analyses can be useful for studies that use heterogeneous datasets for automatic coreference resolution tasks and multilingual coreference projection. Depending on the dataset at hand, a feature adaptation is recommended.

Information on the features that are distinctive for certain classes included into our analysis provide us with patterns of systematic contrasts. The differences in position, grammatical function and forms of coreferring expressions in a source and a target text belonging to the same register cause

[8] We provide the confusion matrix in Appendix.

			Features distinctive for ESSAY		
FICTION	dist-t	ant-ttr	NA	NA	NA
INSTR	dist-t	ant-subj	srate3	ant-other	ana-pers-mod
POPSCI	dist-t	ant-ttr	ana-dem-local	ant-other	length
WEB	dist-t	ant-ttr	ant-subj	srate4.1	ant-event-vp
ACADEMIC	dist-t	ant-subj	ana-pers-mod	ant-ttr	ana-obj
INTERVIEW	ant-subj	ant-obj	dist-t	ana-pers-mod	ana-obj
			Features distinctive for ACADEMIC		
ESSAY	ana-dem-mod	ana-dem-local	ana-dem-head	ana-pers-it	srate4.1
FICTION	ana-dem-head	srate4.1	ana-dem-mod	ana-pers-it	ant-fact-s
INSTR	ana-dem-head	ana-dem-local	srate3	ana-pers-it	srate4.1
POPSCI	ana-dem-head	ana-dem-local	ant-other	ana-pers-it	srate4.1
WEB	ana-dem-head	srate4.1	ana-dem-local	ana-pers-it	ant-event-vp
INTERVIEW	ana-dem-mod	antecedent	chain	ant-obj	ant-subj

Table 7: Features distinctive for ESSAY in different classification tasks.

frequent problems in automatic alignment of the members of coreference chains which may result in erroneous coreference projection. The knowledge on systematic error sources can be used for an automatic improvement of alignment. In coreference resolution, the information on registerial differences may be helpful for domain adaptation. Political essays turn out to have the smallest number of prominent coreference features, which means that working with texts of this register does not require any domain adaptation. There is an opposite tendency for academic speeches – these texts differ strongly from other text types, and thus, domain adaptation is necessary. The knowledge on language, mode and register contrasts is also important for contrastive linguistics and translation studies. In the future, it would be interesting to test whether our assumptions about the correlation of specific types of features and variational dimensions may influence the performance of automatic coreference resolution systems and multilingual coreference projection tasks.

Acknowledgments

The present work was done within the GECCo project funded through the German Research Foundation (DFG). We would like to thank José Manuel Martínez Martínez for contributing to the extraction and calculation of a number of features used in the analyses. We would also like to thank our reviewers for their useful comments and suggestions.

References

Chinatsu Aone and Scott William Bennett. 1996. Applying machine learning to anaphora resolution. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 302–314. Springer, Berlin, Heidelberg.

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat, and Dina Demner-Fushman. 2012. Domain adaptation of coreference resolution for radiology reports. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 118–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mira Ariel. 1990. *Accessing noun-phrase antecedents*. Routledge, London.

Mira Ariel. 2001. Accessibility theory: an overview. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation*, pages 29–88. John Benjamins, Amsterdam/Philadelphia.

Victor 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. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.

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

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*, Beijing, China. 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.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

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.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London, New York.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 142–149, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142, London, UK. Springer.

Andrej A. Kibrik. 2011. *Reference in Discourse*. Oxford University Press.

Manfred Klenner and Don Tuggener. 2011. An incremental model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, pages 81–85.

Sandra Kübler and Desislava Zhekova. 2016. Multilingual coreference resolution. *Language and Linguistics Compass*, 10(11):614–631.

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.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Special Issue of Nordic Journal of English Studies*, 14(1):258–288.

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.

Shalom Lappin and Herbert J. Leas. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, pages 535–561.

Ekaterina Lapshinova-Koltunski, Kerstin Kunz, and Marilisa Amoia. 2012. Compiling a multilingual spoken corpus. In *Proceedings of the VIIth GSCP International Conference: Speech and corpora*, pages 79–84, Firenze. Firenze University Press.

Ruslan Mitkov, Richard Evans, and Constantin Orsan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings*, pages 168–186.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2017. Projection-based coreference resolution using deep syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Valencia, Spain. Association for Computational Linguistics.

Michal Novák and Anna Nedoluzhko. 2015. Correspondences between Czech and English coreferential expressions. *Discours: Revue de linguistique, psycholinguistique et informatique*, 16.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

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

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 720–730, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marta Recasens, Luís Màrquez, Emili Sapena, Toni Martí, Mariona Taulé, Veronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

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

Danny Roobaert, Grigoris Karakoulas, and Nitesh V. Chawla. 2006. Information Gain, Correlation and Support Vector Machines. In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh, editors, *Feature Extraction: Foundations and Applications*, pages 463–470. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ina Rösiger and Simone Teufel. 2014. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the EACL 2014 Student Research Workshop*, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 187–191, Istanbul, Turkey. European Language Resources Association (ELRA).

Vladimir Naumovich Vapnik and Alexey Yakovlevich Chervonenkis. 1974. *Theory of Pattern Recognition*. Nauka, Moscow.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor Wai-Hung Tsang, Kian Ming Adam Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 744–753, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amir Zeldes. 2018. A predictive model for notional anaphora in English. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

type & form	**examples**
pers. head	*he/er, she/sie, they/sie*
pers. modifier	*her/ihr, his/sein, their/ihr*
it-endophoric	*it/es*
demonstr. head	*this/dies/das, that/jenes*
demonstr. modifier	*this/diese(r/s), that/jene(r/s)*
local	*here/hier, there/da*
temporal	*now/jetzt, then/dann*
pronadv	*herewith/hiermit, dagegen, damit*
comparat. particular	*bigger/grösser, better/besser*
comparat. general	*other/andere, such/solche*

Table 8: Anaphors and their subtypes annotated in GECCo.

type	**example**
pronoun	he wrote back saying if <this> is what i think it is...
np	*This set of euro coins will cost <20 marks>. For this, you get 20 coins...*
event-vp	<calculating the number of cannonballs in piles> for him, but this sparked...
fact-s	*<At the same time, we need to double the current level of prosperity>...this is the most urgent moral challenge we face.*
other	longer segments

Table 9: Types of antecedents annotated in GECCo.

feature	score
ana-dem-pronadv	0.8024
ana-pers-it	0.2351
ana-dem-local	0.1641
ana-comp-partic	0.1031
ana-comp-general	0.1031
ana-dem-temp	0.0833
ana-is-subj	0.0696
ant-fact-s	0.0677
srate4.2	0.0662
ana-obj	0.0496

Table 10: Features selected for language prediction with their IG scores.

feature	score
ana-dem-head	0.4672
ana-dem-local	0.2472
ana-pers-it	0.1337
ant-event-vp	0.1272
srate1	0.1247
srate4.1	0.1237
ana-pers-head	0.1194
ana-p-start	0.1181
ant-other	0.1166
ant-fact-s	0.1072
mention	0.1064
chain	0.1045
ana-pers-mod	0.0997
anaphor	0.0915
srate3	0.0825
ana-obj	0.0794
antecedent	0.0743
ant-np	0.0643
length	0.0602
ana-dem-temp	0.0580
ant-ttr	0.0577

Table 11: Features selected for mode prediction with their IG scores.

feature	score
ana-subj	0.557
ana-pers-head	0.547
ant-np	0.538
ana-dem-head	0.524
ana-pers-mod	0.497
anaphor	0.438
mention	0.430
chain	0.412
antecedent	0.408
ana-obj	0.368
length	0.321
ana-dem-mod	0.316
ant-ttr	0.306
ant-obj	0.297
ant-subj	0.294
ana-dem-local	0.248
ana-pers-it	0.230
ant-pronoun	0.225
ant-event-vp	0.216
ant-other	0.207
srate3	0.187
srate4.1	0.181
dist-t	0.166
ant-fact-s	0.162

Table 12: Features selected for register prediction with their IG scores.

a	b	c	d	e	f	g	⇐ classified as
50	0	0	0	2	0	0	**a = ESSAY**
0	**17**	0	3	0	0	0	**b = FICTION**
12	1	**9**	2	0	0	0	**c = INSTR**
6	0	1	**11**	3	0	0	**d = POPSCI**
11	2	3	2	**7**	0	0	**e = WEB**
	0	0	0	1	**13**	5	**f = ACADEMIC**
4	0	0	0	0	1	**21**	**g = INTERVIEW**

Table 13: Confusion matrix for the SVM register classification.