Inter- and Intralingual Variation in a Multilingual Context: Dimensions, Interactions, and their Implications

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
A 2.2 Universität Campus
66123 Saarbrücken
Germany
e.lapshinova@mx.uni-saarland.de
## Contents

### INTRODUCTION
0.1 Aims and Motivation .................................................. 2
0.2 Main Concepts .......................................................... 6
  0.2.1 Language Contrast ............................................... 6
  0.2.2 Register/Genre Variation ...................................... 9
  0.2.3 Text Production Type .......................................... 11
  0.2.4 Variation within Translation .................................. 13
0.3 Methods and Data ...................................................... 19
  0.3.1 Building Resources ............................................. 19
  0.3.2 Feature Identification ......................................... 24
  0.3.3 Feature Exploration ........................................... 28
0.4 Findings ................................................................. 31
  0.4.1 Language Contrast ............................................... 31
  0.4.2 Register Variation ............................................... 34
  0.4.3 Text Production Type .......................................... 37
0.5 Final Discussion and Outlook ...................................... 39
0.6 Authors’s Contribution to the Jointly-Authored Publications ... 44

### References

### COLLECTION OF ARTICLES

Building Multilingual Resources

### Article 1


1 Introduction ............................................................... 59
  1.1 Aims ................................................................. 59
  1.2 Motivation .......................................................... 59
2 Theoretical Background ............................................... 59
3 Corpus Compilation ..................................................... 60
  3.1 Data Collection .................................................... 60
  3.2 Problems in Spoken Data Compilation .......................... 60
4 Corpus Annotation ........................................ 61
5 Corpus Querying .......................................... 61
6 Conclusion and Future Work ............................... 62

Article 2 64

1 Introduction: Aims and Motivation ....................... 64
2 Theoretical Background and Resource Requirements .... 65
   2.1 Translation analysis and translationese .............. 65
   2.2 Language variation .................................. 65
   2.3 Machine translation .................................. 66
3 Methodology .............................................. 66
4 Corpus Resources ......................................... 66
   4.1 Corpus data collection ................................ 66
   4.2 Corpus annotation .................................... 67
   4.3 Corpus querying ...................................... 67
5 Preliminary Analyses ...................................... 68
   5.1 Profile of VARTRA-SMALL in terms of shallow features ... 68
   5.2 Interpretation of results ................................ 69
   5.3 First statistical experiments ......................... 70
6 Conclusion and Future Work ............................... 71

Article 3 74

1 Aims and Motivation ...................................... 74
2 Theoretical Background .................................... 75
   2.1 Frameworks for the analysis of English and German .... 75
   2.2 Framework for the analysis of Czech ................. 75
3 Data and Experiment ....................................... 76
   3.1 GECCo - German and English corpora ............... 76
   3.2 Prague Dependency Treebanks ....................... 76
   3.3 Experiment settings .................................. 77
Article 6


1 Introduction
2 Motivation and Theoretical Background
2 Corpus Resources
4 Annotation of Cohesion
  4.1 Categories to Annotate
  4.2 Annotation of Cohesive Devices
  4.3 Annotation of Coreference
  4.4 Annotation Availability
5 Conclusion and Future Work

Article 7


1 Introduction
2 Theoretical Background
  2.1 Cohesive conjunctions vs. other cohesive devices
  2.2 Conceptualisation and classes
3 Resources and tools to analyse conjunctive relations
  3.1 Corpus resources
  3.2 Annotation and extraction tools
  3.3 Annotation procedures
4 Querying and analysing conjunctive relations
5 Conclusion

Intralingual Variation

Article 8

1 Introduction: Aims and Motivation .................................. 160
2 Theoretical Background .................................................. 161
   2.1 Related Feature Work ........................................... 161
   2.2 Translation Features and their Operationalisation .......... 163
   2.3 Hypotheses ......................................................... 164
3 Resources, Methods and Tools .......................................... 165
   3.1 Corpus Resources ................................................ 165
   3.2 Feature Extraction ................................................ 166
4 Results and their Interpretation ....................................... 167
   4.1 Simplification ....................................................... 167
   4.2 Explicitation ....................................................... 169
   4.2.1 Normalisation and “Shining through” .................... 173
   4.3 Convergence ....................................................... 175
   4.4 Summary ........................................................... 176
5 Conclusion and Future Work ......................................... 176

Article 9


1 Introduction .............................................................. 181
2 Related Work ............................................................ 181
   2.1 Discourse properties in English and German ............... 181
   2.2 Discourse properties in human and machine translation .. 182
3 Methodology ............................................................ 182
   3.1 Data ................................................................. 182
   3.2 Feature selection .................................................. 183
   3.3 Methods ............................................................ 183
4 Analyses ................................................................. 184
   4.1 Discourse properties in English and German ............... 184
   4.2 Originals and translations ..................................... 186
   4.3 Human and machine translations ............................. 187
5 Conclusion and Discussion ........................................... 189
Article 10


1 Introduction ............................................................................. 192
2 Related Work and Theoretical Background ................................. 193
3 Methods .................................................................................. 194
   3.1 Data .................................................................................. 194
   3.2 Algorithms ......................................................................... 194
4 Results .................................................................................... 194
   4.1 Genres and Methods .......................................................... 195
   4.2 Translation Methods ........................................................... 195
   4.3 Different Genres: Different Language? ................................ 196
   4.4 Human vs. Machine ......................................................... 197
   4.5 Feature Analysis ............................................................... 197
5 Conclusion and Outlook .......................................................... 199

Interaction of Dimensions and their Impact ................................. 202

Article 11


1 Research goals and motivation ............................................... 203
2 Related work ......................................................................... 205
   2.1 Language variation and translation ................................... 205
   2.2 Register variation and translation ...................................... 207
   2.3 Variation and translation method ...................................... 208
3 Linguistic features, data and methodology ................................ 209
   3.1 Features under analysis .................................................... 209
   3.2 Corpus data ...................................................................... 211
   3.3 Features extraction .......................................................... 212
   3.4 Statistical analysis ........................................................... 214
4 Analyses, results and interpretation ........................................ 215
4.1 Intra-dimensional variation ........................................ 216
4.2 Inter-dimensional variation in translation varieties .......... 219
4.3 Features involved in the cluster formation ....................... 223
5 Conclusion and future work ......................................... 229

Article 12

Rubino, Raphaël, Ekaterina Lapshinova-Koltunski and Josef van Gen-
abith (2016). Information Density and Quality Estimation Features
as Translationese Indicators for Human Translation Classification. In
Proceedings of NAACL HLT 2016, San Diego, California. 237

1 Introduction .......................................................... 237
2 Related Work ....................................................... 238
  2.1 Translationese .................................................. 238
  2.2 Information Density .......................................... 239
  2.3 Quality Estimation ............................................ 239
  2.4 Translator Experience ....................................... 239
3 Experimental Setup .................................................. 240
  3.1 Supervised Classification .................................... 240
  3.2 Datasets ........................................................ 240
  3.3 Feature Sets ................................................... 241
  3.4 Preprocessing and Tools .................................... 241
4 Results and Analyses ................................................ 242
  4.1 Original vs Translated Texts ................................ 242
  4.2 Translation Expertise ........................................ 242
  4.3 3-way Classification ........................................ 243
  4.4 Feature Performance ........................................ 243
5 Conclusion .......................................................... 244

Article 13

‘registerness’ in human and machine translation: A text classification
approach. In Proceedings of EMNLP 2015 Workshop on Discourse in
Machine Translation, Lisbon, September 17. 247

1 Introduction: Motivation and Goals ............................... 247
2 Related Work ....................................................... 247
  2.1 Main notions within register theory ......................... 247
  2.2 Register in translation ...................................... 248
  2.3 The impact of target and source texts in translation quality .. 248
3 Methodology and Resources ........................................ 249
3.1 Research questions .......................... 249
3.2 Feature selection ......................... 249
3.3 Corpus resources .......................... 250
3.4 Classification methods .................... 250

4 Classification Analysis ...................... 252
4.1 Classifier performance ..................... 252
4.2 Question 1: Translations and register .... 252
4.3 Question 2: The best performance ........ 252
4.4 Question 3: Human vs. machine .......... 253

5 Discussion and Outlook ..................... 253

Article 14


1 Introduction: Motivation and Goals .................. 257
2 Related Work .................................. 258
2.1 Machine Translation Evaluation ................. 258
2.2 Main Notions within Register Theory .......... 258
2.1 Register in Translation ........................ 259

3 Methodology and Resources ..................... 259
3.1 Research Questions .......................... 259
3.2 Feature Selection ............................ 260
3.3 Data ....................................... 261
3.4 Classification Approaches .................... 261
3.4.1 K-Nearest Neighbors (KNN) .............. 262
3.4.2 Support Vector Machines (SVM) .......... 263
3.5 Experimental Setup .......................... 263

4 Classification Results .......................... 263
4.1 Setting 1: Originals vs. Machine Translations .... 263
4.2 Setting 2: Reference Translations vs. Machine Translations ... 265

5 Conclusion .................................. 266
INTRODUCTION
0.1 Aims and Motivation: Why Analyse Variation?

Languages vary due to places, over time, as a result of social settings, contextual situation, in which they are used and other dimensions. These dimensions of language variation have an impact on the language production in the way that they are characterised by certain specific linguistic features diversifying the resulting language varieties. Numerous studies, varying in approaches and the levels of description, are concerned with the phenomena of language variation: diachronic language analysis (Mair, 2006, Leech et al., 2009), dialectology (Kortmann, 2004), sociolinguistics (Agha, 2007) genre/register theory (Halliday and Hasan, 1989, Biber, 1995) and others.

We are interested in the language variation in multilingual contexts which involve more than one language and result in the interaction between several dimensions impacting language production. These contexts may emerge in different multilingual situations promoting either the adoption of lingua franca or multilingual communication in its various forms, see (House and Rehbein, 2004, p. 1) for details. According to the authors, due to the situation of contact between different communication systems represented by different languages, languages mutually influence one another resulting in multilingual communication systems. House and Rehbein (2004, p. 2) characterise the relationship between the languages in this situation according to various factors, e.g. the constellation of the languages involved, the types of text and discourse, the types of media used, the type of social institution and the relative status of participants. In fact, these factors mostly coincide with the dimensions of language variation defined above.

Our focus lies on the multilingual communication that involves German and English, languages that are historically closely related, but that reveal a range of contrasts on different levels of description, see (Hawkins, 1986, König and Gast, 2012). We are especially interested in the contrasts that may have impact on the products of the resulting multilingual communication, e.g. on translation. The language variation under analysis can be subdivided into interlingual – the discrepancy observed between English, German (and in some studies Czech), and intralingual – the divergence across different registers and genres (including written and spoken modes) and also the differences between text production types (translated vs. untranslated).

Translation, being a multi-faceted phenomenon, can also be influenced by different dimensions of language variation. We know that translations are influenced by both language systems of the source and the target languages. The context of situation, i.e. the register a text instantiates, has also an impact on translated texts (Neumann, 2013). We believe that there are more dimensions in play than has been assumed so far. Due to recent developments in natural language processing (NLP), more and more translations are being produced not only by human
0.1 **Aims and Motivation**

translators, but also by machine translation (MT) systems. Furthermore, there are mixed forms of translations, e.g. computer-aided human translation (CAT) or post-edited machine translation (PET). Within these varieties, one can observe further subtypes, e.g. human translations assisted by different systems, or machine translations produced with rule-based or statistical systems, etc. Moreover, both human and machine translation can vary as well, e.g. in the amount of experience or knowledge involved. Human translation can be produced by experienced (e.g. professionals) and inexperienced novice translators (e.g. students). Machine translations produced with statistical MT systems can be trained with large and small data sets. Although statistical machine translation (SMT) is used in most areas nowadays, there exist also other subtypes distinguished by the method or model involved in the training process, e.g. example-based, word-/phrase-/tree-based SMT or even hybrid MT which also contains borrowings from different approaches, e.g. SMT and rule-based MT.

Our conception of language variation in multilingual context include the following dimensions:

1. **Language** reflected in language contrasts, linguistic properties derived from the language systems (mostly English and German in our analyses).

2. **Register** reflected in contextual variation of languages, depending on the context of use, e.g. in political essays or speeches, fictional texts, instruction manuals, etc.

3. **Text Production Type** reflected in the contrasts between the texts originally authored in a language and those translated into this language.

Dimensions for variation that we analyse for translation only:

4. **Method** reflects how a translation was produced: manually (human translation) or automatically (machine translation) or with mixed methods (CAT or PET).

5. **Experience** or data involved in translation depends on the level of proficiency of translators in human translation and on the amount of data used for training in SMT.

In fact, non-translated texts can also vary along the fourth and the fifth dimensions. The successful development of the NLP applications for automatic

---

1Please note that this notion of ‘experience’ here is different from ‘domain experience’ in register theory. We mean here the level of expertise which is also related to discourse in terms of register theory.
text generation allow production of machine-generated texts, which differ from human-produced ones. At the same time, the texts produced by professional writers, e.g. journalists, differ from those produced by novice writers, e.g. school children or students. This dimension is analysed, for instance, in studies on English academic writing, see (Haase and Schmied, 2013, Flowerdew, 2014, Hyland, 2015).

We are also aware of the existence of further dimensions of language variation which are relevant for multilingual communication, e.g. author/translator style, geographical location, time and others. However, their consideration goes beyond the scope of this study.

We call the types of translations distinguished by the different dimensions translation varieties as defined in Article 11 of this volume. The resulting complex interaction of these dimensions is reflected in the translation product, more precisely in its linguistic features, e.g. active vs. passive verb constructions, preferences for certain functional verb classes, modality meanings, proportion of nominal vs. verbal phrases, etc. These features allow us to analyse and model the interplay of the different dimensions of translation. One of our main aims includes the exploration of these features with a corpus-based approach applying various explorative techniques. Methodologically, we focus on quantitative distributions of linguistic features reflected in the lexico-grammar of texts. These distributions result in certain feature sets or profiles reflecting the dimensions of variation involved. Differentiation of the features into dimensional profiles is one of our major tasks, which requires application of suitable statistical methods. The detected feature profiles (each per dimension) form a basis for a model of translation variation.

Our insights from the analyses performed so far show that if we are able to recognise and extract linguistic features from corpora along with their distributional information, we should be able to statistically model language variation. For instance, within the DFG-funded project “Registers in Contact (RegiCo)” at Saarland University, we have done extensive work on discovery of features responsible for linguistic variation along context parameters, which also indicate recent language change. For example, in (Degaetano-Ortlieb et al., 2012), descriptive and explorative statistical techniques were applied to find register features relevant for a diachronic analysis. In a later work (Teich et al., 2015), methods derived from automatic text classification were applied to analyse diversification of scientific registers and to measure the contribution of the linguistic features involved. Within another project, called “German-English Contrasts in Cohesion (GECCo)”, feature work has been restricted to textual properties expressed in cohesion. This work involves contrastive analysis of English and German and identification of contrasts in the textual realisations of cohesion triggering coherence in a discourse. We believe that the knowledge on the language resources used in
both languages and their mechanisms of creating cohesive relations in naturally occurring texts is highly relevant for the analysis of multilingual communication too. For instance, Kunz et al. describe the variation in different registers of the two languages in terms of cohesive devices. With the help of various statistical techniques, they show that it is possible to model English and German registers on the basis of the distributional information extracted from a linguistic corpus. We apply similar techniques for our analysis of the dimensions of language variation.

We assume that with the help of linguistic knowledge about these dimensions, which are intrinsic parameters of translation variation, it will be possible to judge translation quality, and consequently to improve it. This knowledge will be valuable for translation quality assessment and improvement. To our knowledge, none of the existing approaches to translation quality has considered all the dimensions in one complex model; see e.g. works by Toury (2012), House (2014), etc.

In quality assessment, this knowledge will allow us to judge translation quality without interfering factors. For example, in measuring translation quality in terms of registers, we will be able to exclude features which are responsible for other types of variation (e.g. language, method or experience). We believe that in only this case is it possible to claim that a translation meets the register standards of the target language. Our assumption here is that the presence of features from other dimensions may be confounding, and as a consequence, judgement on translation quality might be compromised.

In our future work, we plan to design and test the model experimentally on human and machine translation. In human translation, it will be applied in translation analysis and translator training (e.g. in the form of a book with a systematic description of differences between translation varieties, and guidelines for teaching and evaluation). In machine translation, we intend to use the model for quality estimation, quality evaluation, and improvement of statistical machine translation systems. The latter might be possible by incorporation of the resulting model into the system training data, or in tuning language models.

The remainder of this introduction is structured as follows: Section 0.2 provides a brief clarification of main concepts Section 0.3 deals with the procedures and methodologies for corpus-based analyses, i.e. building and annotating linguistic corpora and preparing them for the required exploitation. Section 0.4 summarises the main findings with respect to the dimensions of language variation under analysis. Section 0.6 contains information about the author’s contribution to the sections contained in each of the fourteen articles presented below.
0.2 Main Concepts of Multilingual Linguistic Variation

In the following section, we will concentrate on the main concepts within our conception of language variation, providing an overview of the related work and the theoretical background used in our analyses.

0.2.1 Language Contrast

House and Rehbein (2004) claim that the processing of knowledge and the very structure of knowledge strongly depend on language or may even be impossible without language. Therefore, one of the most important methods for investigating multilingualism is to contrast languages.

There are a number of studies contrasting English and German that have gained substantial insights in this area, mostly in a system-based, descriptive approaches, where text excerpts and examples are used to illustrate assumptions and claims. For instance, Hawkins (1986) and König and Gast (2012) comprehensively describe contrasts on the lexical and grammatical levels. The difference between these two approaches lies in the nature of described differences. In the comparative typology of Hawkins (1986), there is a general difference in the way English and German map semantics and syntax onto each other, whereas König and Gast (2012) do not give any generalisations of this kind. The authors provide a detailed description the contrasts between the two language systems, starting from the phoneme inventory to a wide range of aspects of sentence grammar.

There are also some example-based studies on the contrasts on the textual level, e.g. by Doherty (2006) and Fabricius-Hansen et al. (2005). The description on the level of the language system is very valuable for contrastive linguistics. However, it does not cover all the cases in the usage of natural language, since there are differences between the potential provided by language systems and the choices made by speakers of a given language (Neumann, 2012, p. 193). The latter cases can be analysed with empirical research strategies involving corpus data that allow interpretations against the background of the system-based comparison.

In the existing corpus-based approaches, many contrastive studies of English and German concentrate on individual phenomena, e.g. (Borthen et al., 2014) on nominal phrases serving as referring expressions, (Johansson, 2007) on generic person and the verb spend [time]. On the level of text/discourse, there are works that examine textual instantiations of various phenomena, e.g. abstract anaphors in (Zinsmeister et al., 2012) cohesive conjunctions in (Becher, 2011, Kranich et al., 2011, House, 2011). Most of them are limited to the investigation of individual devices and consider certain contexts (registers) only. Teich (2003) analyses English-German contrasts in a number of phenomena for verbal and nominal categories, such as modification, voice, transitivity, and the categories of Theme,
agency and grammatical metaphorisation. Some of the examples of more comprehensive corpus-based analyses of English and German that cover several register include (Hansen-Schirra et al., 2012, Neumann, 2013) and (Kunz et al., ress).

Corpus-based studies perform analyses on the basis of information extracted from text corpora. What is commonly being extracted are text instances of a particular lexico-grammatical pattern: particular tokens and sequences of tokens or part-of-speech tags. However, it is difficult to compare languages on the level of tokens, even on the level of parts-of-speech. Due to the differences in the grammatical categories available part-of-speech taggers mostly operate with different tagsets in different languages. This is also related to such closely related languages as English and German, compare tags in the two commonly used tagsets Penn Treebank Tagset (Santorini, 1990) and STTS (Schiller et al., 1999). However, the patterns mentioned are typically realisations of more abstract linguistic categories, such as voice, modality, co-reference, cohesion.

For example, if analysing English and German, we are interested in the contrasts in terms of voice, we would compare the distribution of verbal groups expressing active and passive voice in these two languages. For the passive voice, Teich (2003, p. 181) states that German uses more passive alternatives than English and the difference in their distributions is significant. However, German has several constructions for passive meaning at its disposal. For instance, those with the reflexive verb lassen sich as in example (1-a), the constructions with man as in (1-b) or infinitives as in (1-c).

(1) a. Struktur und Regeln von Systemen lassen sich häufig erst erfassen, wenn sie gestört und zerstört werden. [It often happens that the structures and rules of a system can be comprehended only after they have been disturbed and destroyed].
   b. Doch man kann es eben auch genau anders herum sehen... [Yet this argument can be turned around...].
   c. Mit den Unternehmen, und nicht gegen sie, sind weltweite Regeln und Normen durchzusetzen. [Global rules and standards can only be implemented with business, and not against it].

So, we need to define all linguistic means expressing passive-like constructions in the both languages. However, what is being compared in our analysis, is not the distribution of the concrete sequences of tokens or parts-of-speech, but the distribution of the category passive-like in English and German, which is a subcategory of voice in language grammar.

This is also relevant for the phenomena on the level of discourse. For instance, German dispose of more various patterns expressing coreference, i.e. different forms of personal and demonstrative pronouns and pronominal adverbs, e.g. das,
**0.2 Main Concepts**

*damit* and *diese* in example (2). For a comprehensive comparison of coreference in English and German, we need to define all linguistic means that the two languages have at their disposal to express coreference. Then, we compare the distributions of such categories as **personal** (*They/they, it, sie*) and **demonstrative reference** (*das, damit, diese*) and their structural and functional properties, such as demonstrative with a head function, like *this* in *this is interesting* or *das, damit, diese* in example (2), or with a modifier function, like *this in this book is interesting*, instead of comparing the distributions of the concrete parts-of-speech, i.e. pronouns in this case.

(2) a. *Wir arbeiten für Wohlstand und Chancen, weil das richtig ist. Wir tun damit das Richtige.* [*We work for prosperity and opportunity because they’re right. It’s the right thing to do*].

b. *Für Schleswig-Holstein und Mecklenburg-Vorpommern informieren zwei spezielle Broschüren über Angebote rund ums Radfahren – diese können kostenlos angefordert werden können.* [*There are informative brochures available, explaining everything about cycling through Schleswig-Holstein and Western-Pomerania. They can be ordered free of charge*].

c. *Die TPA ist mehr als ein Werkzeug; sie ist eine Partnerschaft mit den gewählten Vertretern des amerikanischen Volks.* [*TPA is more than a tool; it is a partnership with the elected representatives of the American people*].

In Section 0.3.2, we describe automatic procedures which can be used to annotate such abstract linguistic categories.

This level of abstraction is even more important, if we compare typologically different languages, e.g. German/English and Czech. Here, we would observe even more differences in the lexi-co-grammar, and thus stronger differences in the abstract categories too. For instance, to express relations of identity in a text (co-reference), Czech does not necessarily need a cohesive device, i.e. relation trigger. In cases where English and German operate with personal pronouns, we might have a zero anaphor (pro-drop) in Czech. In this way, the comparison of the pronoun distribution is not sufficient for the analysis of contrasts in the distribution of co-reference relations in German, English and Czech. We need to operate here with more abstract categories, such as coreference or identity relations, head/modifier/temporal/local function, etc. It is at this level of abstraction that grammatical paradigms are most conveniently described in descriptive grammar, see (Quirk et al., 1985) or (Biber et al., 1999), and it is this level of linguistic organisation grammarians or lexicologists are typically interested in.

The main focus in our studies on English-German contrasts lies on the dis-
course level. We describe cross-lingual variation in the two languages in terms of cohesive conjunctions in Article 7 of the present volume and in terms of different discourse phenomena in Article 5. In Article 6, we start from the argument that cohesive relations, such as identity, similarity, etc. are realised in both English and German. We therefore define the categories and subcategories reflecting the lexico-grammatical and semantic features of the devices that establish cohesive relations, and describe the strategies of their semi-automatic annotation. In Article 4, we compare German with Czech, using corpus resources annotated within different frameworks. This analysis is enabled through the formulation of more abstract linguistic categories unifying the existing ones.

In our works involving translation, we also deal with other features prone to cross-lingual variation, i.e. those reflecting context variation. These features are formulated on the basis of Systemic Functional Linguistics (Halliday, 2004, SFL) and Genre/Register theory, see Section 0.3.2 below.

0.2.2 Register/Genre Variation

In our analysis, we prefer to use the term register, and not genre, although they represent two different points of view covering the same ground, see e.g. Lee (2001). However, we refer to genre when speaking about a text as a member of a cultural category, about a register when we view a text as language. In this case, its lexico-grammatical characterisations, conventionalisation and functional configuration are determined by a context use, and its language means vary according to this situation. Different situations require different configurations of a language. According to Steiner (2012b, p. 6), register variation systematically links up with the linguistic system and its multi-functional grammar on the one hand, and with the context of culture on the other.

For the analysis of register variation, we adopt the theoretical background from SFL and Register/Genre theory. This combination of frameworks offers the abstraction level needed for our purposes. In terms of SFL, registers are referred to as linguistic variation according to use in context, i.e. contextual factors systemically influence linguistic variation, giving rise to registers. These contexts influence the distribution of particular lexico-grammatical patterns which manifest language registers. For example, language may vary according to the activity of the involved participants, production varieties (written vs. spoken) of a language, or the relationship between speaker and addressee(s). To account for a contextual configuration, SFL provides three variables characterising the level of context: field, tenor and mode of discourse. These variables are associated with the corresponding lexico-grammatical features. Field of discourse relates to processes and participants (e.g., Actor, Goal, Medium), as well as circumstantials (Time, Place, Manner etc.). Linguistically it is realised in term patterns, func-
0.2 **Main Concepts**

tional verb classes, e.g. activity (*approach, supply*, etc.), communication (*answer, inform, suggest*, etc.), argument structure and adverbial types. Tenor of discourse relates to roles and attitudes of participants, and is realised linguistically in stance expressions used by speakers to convey personal attitude to the given information, e.g. adverbs like *actually, certainly, amazingly, importantly*, or modality expressed by modal verbs such as *can, may, must*. Mode relates to the role of the language in the interaction and is realised at the grammatical level in Theme-Rheme and Given-New constellations as well as cohesive relations at the textual level. Generally speaking, in lexico-grammar, the contextual variables correspond to sets of specific lexico-grammatical features, and different registers vary in the distribution of these features.

The differences between registers can be identified through a corpus-based analysis of phonological, lexico-grammatical and textual (cohesion) features in these registers; see the studies on linguistic variation by Biber (1995) or Biber et al. (1999). Also linguistic variation among registers can be traced in the distribution of those features which are expressed in lexico-grammatical patterns.

Multilingual register studies also concern linguistic variation across languages, comparing the register settings specific for the languages under analysis, e.g. Biber (1995) on English, Nukulaelae Tuvaluan, Korean and Somali, and Hansen-Schirra et al. (2012) and Neumann (2013) on English and German. The latter two also consider this type of linguistic variation in translations. Other translation scholars e.g. Steiner (2004) and House (2014), also pay attention to register variation when analysing language in a multilingual context of translation. However, they either do not account for the distributions of register features, or analyse individual texts only. In the works by De Sutter et al. (2012) and Delaere and De Sutter (2013), register differences are also described for translated texts. Yet, these differences are identified on the level of lexical features only.

The features that are most frequently used in studies on variation (not only register variation) in corpus-based approaches are of shallow character and include lexical density (LD), type-token-ratio (TTR), and part-of-speech (POS) proportionality. Steiner (2012a) uses these features to characterise profiles of various subcorpora distinguished by language (English and German), text production type (translation and original) and eight different registers. The author defines a number of contrast types including register controlled ones which implies 1) contrasts within one register between English and German, and 2) contrasts between registers within each of the languages, see (Steiner, 2012a, p. 72). Both contrast types are relevant for our analysis, since we are interested in both the variation within and across languages. Moreover, we also consider register variation within different text production types, e.g. in translation (addressed in Section 0.2.4 below).

Applying a quantitative approach, Neumann (2013) analyses an extensive set of linguistic patterns reflecting register variation and shows the discrepancies be-
tween the two languages under analysis. The author also demonstrates to what degree translations are adapted to the requirements of different registers, showing how both register and language typology are at work.

Kunz et al. (2016) show that register variation is also relevant for a number of textual phenomena. They analyse structural and functional subtypes of coreference, substitution, discourse connectives and ellipsis on a dataset of several registers in English and German. They are able to identify contrasts and commonalities across the two languages and registers with respect to the subtypes of all textual phenomena under analysis. The authors show that these languages differ as to the degree of variation between individual registers in the realisation of the phenomena under analysis, i.e. there is more variation in German than English. They attest the main differences in terms of preferred meaning relations: a preference for explicitly realising logico-semantic relations by discourse markers and a tendency to realise relations of identity by coreference. Interestingly, similar meaning relations are realised by different subtypes of discourse phenomena in different languages and registers.

Our research strategy is also corpus-based and involves various linguistic features related to register variation. In Article 5 of this volume, we use three main types of cohesive devices in combination: co-reference, substitution and conjunction Halliday and Hasan (1976) and analyse their variation across spoken and written registers of English and German, showing that the mode of production can also play an essential role for the grouping of particular registers in the two languages. In Article 9, we analyse discourse phenomena in a dataset of various registers containing not only originally written texts in English and German, but also translations. We demonstrate that languages, even such closely related ones as English and German, have different preferences in the usage of discourse properties, which are also prone to interlingual variation in terms of registers.

0.2.3 Text Production Type

We understand variation along text production type as differences between translated texts and those that were originally authored in the source or the target language. This type of variation has been widely analysed in corpus-based translation studies for different language pairs, e.g. Laviosa (1996) for English translations from a variety of source languages, Mauranen (2000) for English-Finnish translations, Teich (2003), Steiner (2004), House (2014) for English and German translations. They consider translations to have their own specific properties which distinguish them from originals: both their source texts and comparable texts in the target language. These features establish the specific language of translations which is called translationese as defined by Gellerstam (1986). Comparing Swedish translations from English with Swedish non-translated texts, the
author stated significant differences between them, whereas not all of them were attributable to the source language. This coincides with what Frawley (1984) called third code, describing features of translational language which are supposed to be different from both source and target languages.

Later, Mona Baker emphasised general effects of the process of translation that are independent of source language, e.g. in (Baker, 1993). However, analysing characteristic patterns of translations, she excluded the role of both source and target languages. Within this context, she proposed translation universals – linguistic features which typically occur in translated rather than original texts. According to Baker (1993), they are independent of the influence of the specific language pairs involved in the process of translation. Other scholars, e.g. Toury (1995) or Chesterman (2004), operate with other terms – laws or regularities. We prefer to use the term translation features or phenomena in the present study: to claim the features ‘universal’ we would need to analyse more language pairs and translation directions, and to call them ‘laws’ and ‘regularities’, we would need to test more conditions, e.g. cognitive factors, status of translation, etc., which is not possible with the dataset at hand.

These features can be classified according to different parameters. For instance, Chesterman (2004) draws a distinction between S-universals and T-universal: the first comprise differences between translations and their source texts, and the second cover the differences between translations and comparable non-translated texts. A more fine-grained classification includes the following features: explicitation, a tendency to spell things out rather than leave them implicit; simplification, a tendency to simplify the language used in translation; normalisation, a tendency to exaggerate features of the target language and to conform to its typical patterns; levelling out, in which individual translated texts are more alike than individual original texts; and interference, where features of the source texts are observed in translations. For levelling out, we prefer the term convergence proposed by Laviosa (2002), which implies a relatively higher level of homogeneity of translated texts with regard to their own scores of universal features, e.g. lexical density, sentence length, etc., in contrast to originals. For interference, we prefer to use the term shining through defined by Teich (2003).

Some recent corpus-based studies on translation (Baroni and Bernardini, 2006, Ilisei et al., 2010, Koppel and Ordan, 2011) have shown that it is possible to automatically predict whether a text is an original or a translation. Furthermore, automatic classification of original vs. translated texts found application in machine translation, especially in studies showing the impact of the nature (original vs. translation) of the text in translation and language models used in SMT. Kurokawa et al. (2009) show that for an English-to-French MT system, a translation model trained on an English-to-French data performs better than one trained on French-to-English translations. However, the ‘better performance’ of an SMT
system is measured by BLEU scores (Papineni et al., 2002a), indicating to which extend an SMT output complies with a reference, which is a translation itself. The impact of the original source language on French-English phrase-based SMT was also investigated by Ozdowska and Way (2009). They are able to show that the data containing original French and English translated from French is optimal when building a system translating from French into English. Conversely, using data comprising exclusively French and English translated from several other languages is suboptimal regardless of the translation direction. Inspired by Kurokawa et al. (2009)’s work, Lembersky et al. (2012) show that the BLEU score can be improved if they apply language models compiled from translated texts and not non-translated ones. They also show that language models trained on translated texts fit better to reference translations in terms of perplexity. In fact, this only indicates that machine translations comply more with translated rather than with non-translated texts produced by humans. This results in the improvement of the BLEU score, and not necessary leads to a better quality of machine translation.

Translations differ from originals also in terms of registers. Avner et al. (2014) report the automatic identification of translationese is not accurate when the register of the texts was modified (popular science vs. literature). Similar trends were noted by Rabinovich and Wintner (2015) who demonstrate a drop in the accuracy of the classifiers applied on other registers.

We also analyse the differences between translated and non-translated texts in our work in Article 8 and Article 9. In the latter study, as well as in Article 13 and Article 14, we also address the discrepancies between originals and translations in terms of registers. Moreover, we also point to the fact that register settings of translated texts are different from those of originals regardless the method of translation and the experience of human translators in Article 13. The assumption that translations produced by translators with different levels of expertise are different from texts originally written in the target language was also shown in Article 12, where we used information theory as well as translation quality estimation inspired features.

0.2.4 Variation within Translation

As already stated in Section 0.1, translation is a multi-faceted, and thus multidimensional phenomenon. Its multidimensionality has been investigated from various angles. The traditional focus has been on the influence of the source or the target language, i.e. language dimension, as we call it in our work. The influence of the source language (and also source text) has been analysed within the phenomenon of interference or shining through (See the clarification of the translationese terminology in Section 0.2.3). It is often combined with the analysis of normalisation, i.e. the influence of the target language and its norms, see works
by Teich (2003), Olohan (2004), Mauranen (2004) and others. The influence of
this dimension onto translation has been investigated on different language levels.
Examples of works on lexis include (Olohan, 2004, Williams, 2005) and (Baker,
2007). The latter takes lexical patterning into consideration analysing the use of
idioms. The author states that idioms are, on the one hand, linked to a normali-
sation tendency. And on the other hand, they have some features that discourage
translators to make use of them, as they are often related to creative language use,
informal registers, etc. (Baker, 2007, p. 14–15). In her analysis of English and
German translations, Teich (2003) consider various linguistic phenomena, includ-
ing transitivity and voice, mood and modality, information structure and other.
For her analyses, the author uses a corpus that consists of both parallel and com-
parable texts that allows her to trace (over-)normalisation and shining through in
both English and German translations on the basis of quantitative patterns.

In fact, variation along the dimension of language links up with the linguist-
ic systems of the languages involved and their multi-functional grammar. At
the same time, the context of culture is also involved. Matthiessen (2001), Teich
(2001), Steiner (2002) present an attempt to model translation within this dimen-
sion of variation.

As stated by Kruger and van Rooy (2012, p. 34), in recent years the focus has
broadened as more constraints of text production and perception are taken into
account, and more levels of description (linguistic, social, cognitive) are getting
involved. Translation also varies along the dimensions register. Mauranen (2000)
states that translated academic texts tend to normalise stronger to the norms of the
target language than translated popular non-fiction texts. On the one and, she re-
lates this to the prestige of the register: a higher one (academic prose) is translated
more in agreement with the target norm. Then, further dimensions influencing
translation are involved here, i.e. translation production processes and the social
position of the translator (Mauranen, 2000, p. 13). The author also claims that
similar registers in different languages tend to have similar communicative needs.
This means that they would also have similar context of use resulting in simi-
lar register variation. In the case of such cross-linguistic similarities, translations
should also be similar to both their source texts and the comparable originals in
the target language. Neumann (2013, p. 104) claims that in cases of registerial dif-
fierences, the results for the translations by default should lie between the source
texts and the comparable originals in the target language. The author analysed
eight registers in English and German focussing on how translations (both direc-
tions) were adapted to the requirements of these different registers. The default
situation also includes tendencies towards either the source or the target origi-
nals. In cases of cross-lingual differences (between source and target language
originals), we would observe both shining through and normalisation in the trans-
lations to varying degrees, as it was stated by Teich (2003). However, translations
may also vary to some degree from non-translated texts independently of the influence from language contrasts and register differences, see (Neumann, 2013, p. 105).

Further dimensions of translation variation include the differences between human and machine translations, i.e. what we call translation method which can be either manual or automatic. There are also mixed forms (as already stated in Section 0.1 above) which are, however, not within the scope of this work.

Considering existing studies, we are aware that variation according to translation method and experience has not received much attention. There are numerous studies in the context of NLP that address both human and machine translations (Papineni et al., 2002b, Babych et al., 2004). Yet they all serve the task of automatic MT system evaluation and focus solely on translation error analysis, using human translation as a reference in the evaluation of machine translation outputs. Evaluations serve the task to prove to what extent automatically translated texts (hypothesis translations) comply with the manually translated ones (reference translations). The ranking of machine-translated texts is based on scores produced with various metrics. The metrics applied in the state-of-the-art MT evaluation are automatic and language-independent: BLEU and NIST (Doddington, 2002). However, since they do not incorporate any linguistic features, BLEU scores need to be treated carefully, which was demonstrated by Callison-Burch et al. (2006) and (Vela et al., 2014). This fact has been advancing the development of new automatic metrics, such as METEOR (Denkowski and Lavie, 2014), Asiya (González et al., 2014) and VERTa (Comelles and Atserias, 2014). They incorporate lexical, syntactic and semantic information into their scores. The accuracy of the evaluation methods is usually proven through human evaluation. More specifically, the automatically provided scores are correlated with the human judgements which are realised by ranking MT outputs (Bojar et al., 2014, Vela and van Genabith, 2015) and others. Another human evaluation method includes measuring productivity within post-editing procedures, see for instance, works by Guerberof (2009), Zampieri and Vela (2014).

As already mentioned, some of the existing metrics incorporate linguistic knowledge. There are even more works on MT evaluation that operate with linguistically-motivated categories, e.g. (Popovic and Ney, 2011) or (Fishel et al., 2012). However, none of them provides a comprehensive analysis of the differences between human and machine translation in terms of specific linguistically motivated features. In fact, the knowledge on the discriminative features of human and machine translation can be derived from the studies operating with machine learning procedures for MT evaluation, such as (Stanojević and Simaán, 2014) or (Gupta et al., 2015). Corston-Oliver et al. (2001) use classifiers that learn to distinguish human translations from machine ones. These classifiers are trained with various features including lexicalised trigram perplexity, part of
speech trigram perplexity and linguistic features such as branching properties of the parse, function word density, constituent length, and others. Their best results are achieved if perplexity calculations were combined with finer-grained linguistic features. Their most discriminatory features that differentiate between human and machine translations are all of linguistic character and include the distance between pronouns, the number of second person pronoun, the number of function words, and the distance between prepositions. Volansky et al. (2011) operate with translationese-inspired features, and are able to distinguish between manual and automatic translations in their dataset with 100% accuracy. However, the manual and automatic translations they are using have different source texts. We believe that the distinction they are able to achieve is not the distinction between translation methods, but rather between different underlying texts, since their most discriminatory features are the ones that show good performance in any text classification task (token n-grams). El-Haj et al. (2014) make us of readability as a proxy for style and analyse consistency in translation style considering how readability varies both within and between translations. They compare Arabic and English human and machine translations of the originally French novel “The Stranger” (French: L’Étranger). The results show that translations by humans (both male and female) are closer to each other than to automatic translations. The authors also measure closeness of translations to the original in terms of the selected measures, which should serve as an indicator of translation quality.

None of the existing studies, known to us, consider a potential for comparability of human and machine translation with regard to experience. By experience we mean the focus on the degree of experience of human translators and the amount of training data in a statistical machine translation system. In the area of statistical machine translation, the relation between large and small data plays an important role. Gains from using large corpora for translation model training or for language model training are described in various studies (Bertoldi and Federico, 2009, Koehn and Haddow, 2012). Large training data sets make it possible to learn longer phrases in a phrase-based machine translation, see (Koehn, 2010).

In human translations, the relation between translations produced by professionals and those produced by inexperienced translators has been analysed in a number of studies (Jääskeläinen, 1997, Englund-Dimitrova, 2005, Göpferich and Jääskeläinen, 2009, Carl and Buch-Kromann, 2010). For example, Jääskeläinen (1997) describes translational behaviour of professionals and non-professionals who perform translation from English into Finnish. Her findings show that there is a difference in the kinds of knowledge the subjects with different level of professionalism used and the purposes for which they used that knowledge. The professional translators were more inclined to apply textual and extra-textual knowledge, while the non-professionals tended to remain at the linguistic surface level. Her results also show that with the experimental task quality depended more on
the kinds of knowledge a subject was applying than on his or her experience with translating.

Carl and Buch-Kromann (2010) apply psycholinguistic methods in their analysis. They present a study of translation processes for student and professional translators, relating properties of the translators’ eye movements and keystrokes to the quality of the produced translations. They are able to show that the translation behaviour of novice and professional translators differs with respect to how they use the translation phases. Thus, their focus is on the process rather than the product, and there is no description of the linguistic features that reflect the observed variation. Englund-Dimitrova (2005) develops a combined process and product analysis and compares translators with different amounts of translation experience, but concentrates only on cohesive explicitness. Göpferich and Jääskeläinen (2009) provide an overview of the studies in process-oriented investigations of translation competence, describing their advantages and drawbacks. They also summarise methods and findings gained in this field. The main focus lies, however, on one longitudinal study that investigates the development of translation competence of translation students over a period of three years and compares them to that of professional translators with more than 10 years of experience.

We assume that the amount of experience in human translation, and the amount of data in machine translation, have similar outcomes in both translation varieties. Texts translated by novice translators may have commonalities with those produced with an SMT system trained with a small amount of data, whereas texts translated by experienced translators maybe comparable to those produced with SMT systems trained with large data; see, for example, our initial results in Article 8 of this volume. In Article 12, we also address the dimension of translation experience, attempting to distinguish two levels of translation professionalism with the help of features borrowed from MT quality estimation and the uniform information density hypothesis (Frank and Jaeger, 2008).

We are not aware of many studies regarding the interplay between the described dimensions influencing translation. (Kruger and van Rooy, 2012) try to answer the question on the relationship between register and the features of translated language. Their hypothesis was that the translation-related features would not be strongly linked to register variation suggesting that in translated text reveal less register variation, or sensitivity to register, which is a consequence of translation-specific effects. However, their findings provide limited support for this hypothesis. They state that the distribution and prevalence of linguistic realisations of the features of translated language may vary according to register. Therefore, the concept of translated language should be more carefully analysed and defined in terms of registers (Kruger and van Rooy, 2012, p. 61–62).

Jenset and McGillivray (2012) analyse the productivity and use of derivational affixes in translated English applying multivariate statistical techniques. They pay
attention to the interaction between affix use and translator, source language, author, and text type in translated English. The methods applied allow them to state that the correlation between affix and text type (approximation to what we call register) seems to be a much better explanatory structure than affix and translator or affix and source language. So, affix use seems to be constrained more by the conventions imposed by registers than by source language or translators’ background. The interaction between the dimensions of register, author and translator was also analysed by Jenset and Hareide (2013) who use patterns of sentence alignment as features.

This means that the interaction between the dimensions influencing variation in translation can and should be analysed with corpus-based techniques. The research questions to answer in this context include: 1) Which dimensions come more often into interaction?; 2) Which dimension has the greatest impact on translation? and 3) What are the main implications of this interplay? We try to find answers to these questions, analysing the interplay between register and method of translation in Article 11 and Article 9. In the latter, we also pay attention to the dimension of language (both the source and the target).
0.3 Methods and Data

As already mentioned above, we operate with corpus-based methods which include compilation and annotation of bilingual corpora, definition and extraction of features from a corpus, analyses of these features with various statistical techniques including unsupervised and supervised feature classification.

0.3.1 Building Resources

RESOURCES: NOCH STATISTIKEN ZU ANNOTIERTEN STRUKTUREN ANGEBEN!

The exploration of intra- and interlingual language variation requires resources that contain original texts in at least two languages, and translations into one of this languages. There exist several multilingual resources for English and German, e.g. CroCo (Hansen-Schirra et al., 2012) and those contained in the resource OPUS (Tiedemann, 2012).

Our contrastive work consists of analyses based on GECCo, which is an English-German corpus that contains a continuum of different registers from written to spoken discourse (understood as a sub-dimension of register variation under mode of discourse). At the moment, GECCo has several versions, which are dedicated to different contrastive analysis tasks: GECCO-UPOS, GECCOCOH and GECCOCCHAIN.

GECCO-UPOS is the biggest version of the corpus. The written part (approx. 0.57 million tokens) of GECCO-UPOS was imported from the above mentioned corpus Croco. This was redesigned and restructured, received a more structured annotation framework. This subcorpus includes eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), political essays (ESSAY), corporal communication (SHARE), instruction manuals (INSTR), texts from company websites (WEB), prepared speeches (SPEECH) and fictional texts (FICTION). The latter two registers are considered to lie at the borderline between written and spoken discourse. The spoken part (approx. 0.47 million tokens) includes further six registers: academic speeches (ACADEMIC), interviews (INTERVIEW), medical consultations (MEDCONSULT), sermons (SERMON), internet forum conversations (FORUM) and talk shows (TALKSHOW). The texts for these registers were for the most part collected from existing resources, except . The detailed description of the creation of ACADEMIC and INTERVIEW are provided in Article 1 of the present volume. The German texts in ACADEMIC are transcribed lectures from different departments of the Saarland University. The lectures were recorded by the Virtual University of Saarland (VISU). In Article 1, we addressed a number of problems related to handling spoken data. These problems are, on the one hand, of practical character, e.g. formulation of transparent transcription guidelines, comparability of the transcriptions and the included tags (repairs, repeats,
etc.) with the existing English texts that were imported from other resources. On the other hand, there were some cultural differences observed. For instance, the seminars at Saarland University turn out to be less interactive and dialogic than assumed and hence, did not correspond to their English counterparts. Similar issues were handled in the creation of the FORUM register, which contains additional tags for links, comments and other specific elements. The English parts of the registers MEDCONSULT, SERMON were excerpted from BNC (BNC, 2007). The German MEDCONCSULT texts were imported from the corpus ‘Dolmetschen im Krankenhaus’ (Bührig et al., 2012), and the German SERMON texts from the ‘Predigtauch von Oberpfarr- und Domkirche zu Berlin’ (Sermon Archive of the Supreme Parish and Collegiate Church in Berlin)\(^2\). The information on the corpus size and structure is given in Table 1.

<table>
<thead>
<tr>
<th>mode</th>
<th>register</th>
<th>EO</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoken</td>
<td>ACADEMIC</td>
<td>39.975</td>
<td>41.248</td>
</tr>
<tr>
<td></td>
<td>INTERVIEW</td>
<td>36.958</td>
<td>39.953</td>
</tr>
<tr>
<td></td>
<td>FORUM</td>
<td>41.554</td>
<td>38.852</td>
</tr>
<tr>
<td></td>
<td>MEDCONSULT</td>
<td>8.927</td>
<td>19.487</td>
</tr>
<tr>
<td></td>
<td>SERMON</td>
<td>92.914</td>
<td>21531</td>
</tr>
<tr>
<td></td>
<td>TALKSHOW</td>
<td>50.003</td>
<td>38.646</td>
</tr>
<tr>
<td></td>
<td><strong>total-spoken</strong></td>
<td><strong>270.331</strong></td>
<td><strong>199.717</strong></td>
</tr>
<tr>
<td>written</td>
<td>ESSAY</td>
<td>34.815</td>
<td>35.678</td>
</tr>
<tr>
<td></td>
<td>FICTION</td>
<td>36.450</td>
<td>36.792</td>
</tr>
<tr>
<td></td>
<td>INSTR</td>
<td>36.307</td>
<td>36.894</td>
</tr>
<tr>
<td></td>
<td>POPSCI</td>
<td>34.997</td>
<td>36.186</td>
</tr>
<tr>
<td></td>
<td>SHARE</td>
<td>35.657</td>
<td>35.250</td>
</tr>
<tr>
<td></td>
<td>SPEECH</td>
<td>34.822</td>
<td>35.403</td>
</tr>
<tr>
<td></td>
<td>TOU</td>
<td>35.585</td>
<td>36.604</td>
</tr>
<tr>
<td></td>
<td>WEB</td>
<td>35.944</td>
<td>35.793</td>
</tr>
<tr>
<td></td>
<td><strong>total-written</strong></td>
<td><strong>284.577</strong></td>
<td><strong>288.600</strong></td>
</tr>
</tbody>
</table>

**Table 1:** Information on the corpus size for GECCO-UPOS

Both the English and the German subcorpora in all versions of GECCo are annotated on various linguistic levels, which again, raises the problem of comparability, since the annotated categories should be available in both languages. In Section 0.2.1 above, we already addressed the problem of comparability of POS tags available in the state-of-the-art tools. For this reason, GECCO-UPOS was

\(^2\)http://www.berlinerdom.de/
annotated with the universal POS tagset (Petrov et al., 2012). The technical details on the mapping of Penn Treebank and STTS to Universal POS tags were documented by Martinez-Martinez (2015). The comparability of the English and German POS categories was achieved here at the cost of granularity, which is sufficient for the analysis of such shallow features, as type-token-ratio (TTR), lexical density (LD), most frequent words and part-of-speech distributions, which are among the most frequently used variables characterising linguistic variation in corpora, see Biber et al. (1999) and many others.

For the analysis of more complex linguistic structures, further information is required, i.e. annotation of chunks, clause and sentence boundaries, cohesive devices and chains. These annotations are available for some parts of GECCo only. For instance, the version with the annotated cohesive devices is called GECCO-COH and contains ten registers only: all written registers and two spoken ones – ACADEMIC and INTERVIEW, see Table 2. Note that this version contains a different number of tokens, since annotations were achieved with different procedures than those in GECCO-UPOS.

<table>
<thead>
<tr>
<th>register</th>
<th>EO</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACADEMIC</td>
<td>40.559</td>
<td>43.703</td>
</tr>
<tr>
<td>ESSAY</td>
<td>34.998</td>
<td>35.668</td>
</tr>
<tr>
<td>FICTION</td>
<td>36.996</td>
<td>36.778</td>
</tr>
<tr>
<td>INSTR</td>
<td>36.167</td>
<td>36.880</td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>37.898</td>
<td>40.198</td>
</tr>
<tr>
<td>POPSCI</td>
<td>35.148</td>
<td>36.177</td>
</tr>
<tr>
<td>SHARE</td>
<td>35.824</td>
<td>35.235</td>
</tr>
<tr>
<td>SPEECH</td>
<td>35.062</td>
<td>35.399</td>
</tr>
<tr>
<td>TOU</td>
<td>35.907</td>
<td>36.574</td>
</tr>
<tr>
<td>WEB</td>
<td>36.119</td>
<td>35.779</td>
</tr>
<tr>
<td>TOTAL</td>
<td>364.678</td>
<td>372.391</td>
</tr>
</tbody>
</table>

Table 2: Information on the corpus size for GECCOCOH

A small subset of the corpus (GECCOCHAIN) contains also annotations of lexical chains, i.e. relations of synonymy, antonymy, hyponymy, etc. The registers included, and the size of annotated subsets are listed in Table 3. ESSAY and POPSCI represent written discourse, INTERVIEW represents spoken discourse, whereas FICTION is on the borderline, as it contains both written and spoken elements in the form of dialogues.

The described versions of GECCo vary in their annotations, since they serve different tasks. As already mentioned, GECCO-UPOS allows an overall comparison of English and German in terms of shallow features. This is also the biggest...
0.3 Methods and Data

<table>
<thead>
<tr>
<th>register</th>
<th>EO</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>27.171</td>
<td>31.407</td>
</tr>
<tr>
<td>FICTION</td>
<td>36.996</td>
<td>36.778</td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>30.057</td>
<td>35.036</td>
</tr>
<tr>
<td>POPSCI</td>
<td>27.055</td>
<td>32.639</td>
</tr>
<tr>
<td>TOTAL</td>
<td>121.279</td>
<td>135.860</td>
</tr>
</tbody>
</table>

Table 3: Information on the corpus size for GECCOCHAIN

version, since the available annotations can be obtained with fully automatic procedures. The smaller versions of the subcorpora are important for a deeper linguistic analysis of language contrasts. This kind of analysis often requires manual or semi-automatic annotation procedures. Since the latter are costly and time-consuming, they are applied for smaller corpora. The information on different annotation procedures is given in Section 0.3.2 below. The described versions of the GECCo corpus that differ in the annotations contained can be queried under https://fedora.clarin-d.uni-saarland.de/cqpweb/ available under CLARIN-D license.

The analysis of language variation related to translation requires not only comparable non-translated texts in the source and the target language, but also their multiple translations. In our studies, we have restricted our analyses on translations to the language pair English-German, more precisely, translations from English into German. Although closely related to English and sharing many features with it, German is more complex and often more difficult to process for many NLP tools. However, there exist not so many resources and corpus-based study for contemporary German, if compared to English.

So, we create a resource suitable for the analysis of translation variation. Its current release (VARTRA-SMALL) is described in Article 2. This corpus contains multiple translations (of the same texts) from English into German produced by: (1) human professionals (PT1), native speakers of German under unknown settings; (2) human student translators, native speakers of German, with the help of computer-aided translation tools (PT2); (3) a rule-based MT system (RBMT) and (4) two statistical MT systems (SMT1 and SMT2). The translations by professionals, were exported from the above mentioned corpus CroCo. The PT2 variant was produced by student assistants who used a translation memory with the help of the computer-aided tool ACROSS in the translation process (www.my-across.net). The RBMT variant was translated with SYSTRAN 6. For SMT, two versions were compiled: one produced with Google Translate (SMT1, http://translate.google.com/), and the other with a Moses system (SMT2). The main difference between SMT1 and SMT2 is the amount
of the training data behind it. SMT1 was trained with a huge amount of data which is unknown to us, whereas SMT2 was trained with a smaller dataset from EUROPARL Koehn (2005).

In Article 2, professional translations were called PT, whereas student translation variety was called CAT derived from ‘computer-aided translation’. For the sake of consistency in our comparison task, we decided to rename PT as PT1 and CAT as PT2, which makes them comparable to SMT1 and SMT2. Each translation variant is saved as a subcorpus and covers seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The total number of tokens in translations of VARTRA-SMALL comprises 662,851, see details on the corpus size per register in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>PT</th>
<th>CAT</th>
<th>RBMT</th>
<th>SMT1</th>
<th>SMT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>15.574</td>
<td>15.795</td>
<td>15.032</td>
<td>15.120</td>
<td>14.746</td>
</tr>
<tr>
<td>FICTION</td>
<td>11.257</td>
<td>12.566</td>
<td>11.048</td>
<td>11.028</td>
<td>10.528</td>
</tr>
<tr>
<td>SHARE</td>
<td>24.613</td>
<td>24.764</td>
<td>22.768</td>
<td>22.792</td>
<td>22.392</td>
</tr>
<tr>
<td>SPEECH</td>
<td>23.346</td>
<td>24.321</td>
<td>23.034</td>
<td>22.877</td>
<td>22.361</td>
</tr>
<tr>
<td>TOU</td>
<td>17.638</td>
<td>19.721</td>
<td>17.761</td>
<td>17.768</td>
<td>17.671</td>
</tr>
<tr>
<td>TOTAL</td>
<td>133.236</td>
<td>139.825</td>
<td>131.330</td>
<td>130.568</td>
<td>127.892</td>
</tr>
</tbody>
</table>

Table 4: Tokens per register in VARTRA-SMALL

The source texts of the VARTRA translations (EO) are also a part of CroCo (the same texts served as basis for the written part of the above described GECCo). For the analysis of variation along the dimension of text production type, we also use comparable originals in German (GO), also contained in CroCo and the written part of GECCo.

VARTRA represents a corpus of different translation varieties, containing translations of different registers produced with different methods and tools. This corpus allows for analysis of variation in terms of the influencing language (English source or German target); register typology (e.g. fiction vs. popular-scientific articles, etc.); translation methods (human vs. machine), and experience types (experienced vs. inexperienced translators or large vs. small training data).

All subcorpora of VARTRA-SMALL are tokenised, lemmatised, tagged with part-of-speech information, and segmented into syntactic chunks and sentences. The annotations were obtained with Tree Tagger (Schmid, 1994). VARTRA is
encoded in CWB and can be queried with the help of Corpus Query Processor (Evert, 2010, CQP). Parts of the meta-data, such as information on register, as well as translation method, tools used and the source language, are also included. The first version of the corpus can be queried under https://fedora.clarin-d.uni-saarland.de/cqpweb/ available under CLARIN-D license (hdl:11858/00-246C-0000-0023-8CDB-A).

As previously mentioned, corpus compilation and annotation is a time-consuming task. We therefore propose an approach towards interoperable resources for multilingual analysis, see Article 3 and Article 4 of this volume. Our main idea here is to take advantage of existing resources, which were, however, annotated on the basis of different frameworks. In these studies, we present positive results of the experiments to use an interoperable scheme unifying discourse phenomena into more abstract categories and considering only those phenomena that have a direct match in the two frameworks under analysis. We use the one based on the definition of cohesion by Halliday and Hasan (1976) for German and English and discourse phenomena in Prague Discourse Treebank described in (Zikánová et al., 2015) for Czech. Since both existing annotation schemes only account for the systemic peculiarities and realisational options of the languages analysed, they are not general enough to permit a comparison across Germanic and Slavic languages. For this reason, we unified the categories in these schemes and created interoperable one which can be applicable to multiple languages and text registers. The successful application of the scheme in the mentioned studies indicates possible interoperability in the existing resources. In this way, our methodology saves time and effort as no compilation of additional resources is required. This is valuable for corpus-based contrastive analysis that requires multilingual data annotated according to the same scheme. We believe that this approach opens up new paths for contrastive linguistics, translation studies and multilingual NLP.

0.3.2 Feature Identification

For our analyses, we selected sets of features derived from the studies on language, register and translation variation described in Section 0.2 above. They are based on several theoretical frameworks: 1) the combination of SFL and Register/Genre Theory; 2) translation features based on Translationese; 3) Information Density. These features are formulated as abstract concepts, e.g. processes and participants, shining through, textual cohesion, etc. For their exploration in a corpus, we define operationalisations, which are lexico-grammatical patterns. What is being extracted from corpora are text instances of these lexico-grammatical patterns, particular tokens, sequences of tokens or part-of-speech tags, see Section 0.2.1 above. Particular analysis tasks impose restrictions on the definition of specific features. Thus, in a register-independent study, features should be
**0.3 Methods and Data**

content-independent, i.e. they should do not contain terminology or keywords. In the following, we will outline the features used, and described procedures for their extraction from and annotation in corpora.

In Section 0.2.2 above, we described three contextual parameters from SFL that are associated with various lexico-grammatical patterns reflecting language variation. Table 5 illustrates examples of these patterns used in our analyses. The first column contains the corresponding contextual parameter of variation, the second column shows examples of features formulated in abstract categories, and the third column provides examples of lexico-grammatical patterns serving as operationalisations for the features. These features were analysed in most of the articles collected in the present volume. For instance, we concentrate on the features of the contextual parameter of mode in Article 3 and 4, where we use various categories of textual cohesion to analyse contrasts between languages. In Article 5 and 7, we use these features for a cross-lingual register analysis of English and German. Cohesive features are also applied in Article 9, where we analyse variation in translation comparing them to non-translated texts. We use nearly the whole set of the presented registerial features in the analysis of translation variation in Article 11, 13 and 14.

<table>
<thead>
<tr>
<th>contextual parameter</th>
<th>feature</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIELD</td>
<td>participants and processes</td>
<td>nominal and verbal chunks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nominal and verbal parts-of-speech</td>
</tr>
<tr>
<td></td>
<td>vocabulary and style</td>
<td><em>ung</em>-nominalisations in German and general nouns (<em>fact, plan</em>)</td>
</tr>
<tr>
<td></td>
<td>voice</td>
<td>verbs in passive/active</td>
</tr>
<tr>
<td>TENOR</td>
<td>modality</td>
<td>obligation/necessity (<em>must, ought to, should</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>permission/possibility/ability (<em>can, may</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>volition/prediction (<em>will, would, shall</em>)</td>
</tr>
<tr>
<td></td>
<td>evaluation</td>
<td>evaluative patterns (<em>it is interesting to know</em>)</td>
</tr>
<tr>
<td>MODE</td>
<td>textual cohesion</td>
<td>coreference (including distribution of pronominal/ nominal reference, personal/demonstrative pronouns)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>conjunctive relations (additive, causal, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>substitution (<em>one, those, ein/e/r</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ellipsis (<em>Little [] is the next big thing</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lexical cohesion (synonyms, hyponyms, etc.)</td>
</tr>
</tbody>
</table>

*Table 5: Features and patterns from SFL and Register Theory*

Analysing variation within translations, as well as their differences from non-
translated texts, we also include translation features (see Section 0.2.3) into our analysis. In Table 6, we outline the patterns used for the investigation of this variation type. These features were thoroughly analysed in Article 8 and 11. In the latter, we also show how translation features are related with those borrowed from SFL and Register Theory.

<table>
<thead>
<tr>
<th>translation feature</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMPLIFICATION</td>
<td>type-token-ratios</td>
</tr>
<tr>
<td></td>
<td>lexical density</td>
</tr>
<tr>
<td>EXPLICITATION</td>
<td>conjunctions</td>
</tr>
<tr>
<td></td>
<td>nominal/ pronominal reference</td>
</tr>
<tr>
<td></td>
<td>general nouns</td>
</tr>
<tr>
<td>NORMALISATION / SHINING THROUGH</td>
<td>nominal/ verbal pos and phrases</td>
</tr>
<tr>
<td></td>
<td>personal/ demonstrative pronouns</td>
</tr>
<tr>
<td></td>
<td>modal verbs</td>
</tr>
<tr>
<td>CONVERGENCE</td>
<td>various features</td>
</tr>
</tbody>
</table>

Table 6: Translation features

In Article 12, we also include features based on the surprisal measure, a measure of how predictable an expression is. In fact, this is a quantification of the information conveyed by an expression. Surprisal features are extracted using language models trained on words, delexicalised parts-of-speech and flattened syntactic trees. These features are used for the analysis of translation variation along the dimension of experience, since surprisal values may be related to such translation features as shining through (the influence of the source language), simplification or explicitation, see Section 2.2 in Article 12.

We already stated above that the information on the abstract linguistic categories (contextual parameters and their corresponding features or translation features) can be operationalised in particular lexico-grammatical patterns, which are sequences of tokens or part-of-speech tags, often enhanced with contextual restrictions, e.g. sentence or phrase position. For a contrastive analysis, text instances of these patterns are extracted from appropriate corpora along with their distributional information.

We use CQP (see Section 0.3.1) to extract most of the features for our analyses. This query tool allows definition of language patterns in form of complex regular expressions based on string, part-of-speech and chunk and further available tags, which is beneficial for extraction of linguistically motivated features. Such regular expressions can match on various annotation strings, test for membership in user-specific word lists, include special operations on feature sets, and have constraints to specify dependencies. CQP is integrated into the IMS Cor-
pus Workbench (CWB, 2010), which provides an environment for storage and querying of text corpora.

So, we start with a set of candidate lexico-grammatical patterns derived from the theoretical background, as described above, and formulate them as CQP queries. Table 7 provides examples for two queries to extract SFL-inspired features. Query (1) is a simple lexical search with a positional constraint – here, we want to find all instances of ‘additive’ conjunctions when acting cohesively (in principle a closed class the members of which we know). Query (2) is a syntactic pattern search – here we want to find all instances of the pattern with an evaluative adjective (in principle an open class of which we know some members).

<table>
<thead>
<tr>
<th>query building blocks</th>
<th>description</th>
<th>extracted text instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) &lt;s&gt; [(][0,2]</td>
<td>sentence start followed by two optional words</td>
<td>and, moreover, in addition, etc.</td>
</tr>
<tr>
<td>[lemma=RE($additive)]</td>
<td>conjunctions from the ‘additive’ list</td>
<td></td>
</tr>
<tr>
<td>(2) [word=&quot;[I</td>
<td>i]t&quot;]</td>
<td><em>it</em></td>
</tr>
<tr>
<td>[lemma=&quot;be&quot;]</td>
<td>verb <em>be</em></td>
<td><em>is</em></td>
</tr>
<tr>
<td>[pos=&quot;JJ.*&quot;]</td>
<td>adjective</td>
<td>important</td>
</tr>
<tr>
<td>[word=&quot;that\to&quot;]</td>
<td>subjunct <em>that/ particle to</em></td>
<td><em>that</em></td>
</tr>
</tbody>
</table>

**Table 7:** CQP query for conjunctive relations and evaluative patterns

Queries deliver concordances, i.e. text instances which can be sorted according to the texts, registers and subcorpora they occur. The information on distributions can, on the one hand, be extracted for particular patterns. For instance, the frequencies of the conjunction *and* or conjunction *in addition* are sorted according to their occurrences in different registers of the English subcorpus. On the other hand, the frequencies of more abstract features can be generalised with the help of the formulated queries, since not only frequencies of *and* or *in addition*, but also all other additive conjunctions in the predefined list are extracted. In this way, we obtain information on the distributions of all additive conjunctions across different registers of the English subcorpus. These distributions can be compared to the distributions of all additive conjunctions across different registers in the German subcorpus. The extracted distributional information can then be saved in tables and matrices and used for statistical analyses.

Since CWB also provides tools for annotation, we propose a method of adding a more abstract annotation level to a corpus on the basis of the pattern-based extraction. The formalism used for interactive querying in CWB can be applied for annotation. There is no conceptual difference between extraction queries and annotation rules other than their function. Query results deliver not only con-
cordances of the searched structures but also information on their corpus positions. So, we elaborate a set of semi-automatic procedures involving an iterative extraction-annotation process that is based on the framework of YAC described by Kermes and Evert (2002) and by Kermes (2003). This permits to import the information on queried data back into the corpus. We use such procedures to semi-automatically annotate cohesive devices in English and German, as described in Article 6 of the present volume.

The annotation of such linguistic categories provides a different level of abstraction to a corpus and facilitates the access to abstract linguistic information for less experienced users. Moreover, it makes the extraction of more complex linguistic phenomena easier and more efficient. While for some phenomena, simple queries on lemma and part-of-speech information are sufficient, other phenomena require more complex queries including lexical or contextual restrictions or the combination of several simple queries. In this case, the annotation can help to simplify the queries and to make the extraction process more efficient, which is especially valuable in a contrastive corpus work.

These extraction and annotation procedures are used to work with theory-driven features. In another study, which is described in Article 10 of this volume, we use a data-driven approach to feature extraction. The features include bag-of-words, word bigrams, word trigrams and word 4-grams. In some of the experiments, all nouns are substituted with place-holders in some of the experiments. This results in a semi-delexicalised text representation, which is supposed to minimise topic variation, and thus, content-dependence of the features. Text classification methods are applied to level out discriminative features of different translation varieties that intuition alone cannot grasp; thus enabling us to investigate in more detail the properties of each of them.

**0.3.3 Feature Exploration**

In the feature exploration work, we deploy various methods and techniques using R (Venables et al., 2011) and Weka (Witten et al., 2011), which are amongst the most popular open source tools for statistical computing and visualisation graphs.

We apply various descriptive data analyses (variance, median and mean) and visualisation techniques (bar and box plots) to observe frequencies and derive general tendencies in the distribution of our data, see Articles 4, 5, 7 and 8. The results are also tested for significance (with Pearson’s chi-squared test).

Exploratory techniques (applied in Articles 2, 5 and 9) allow us to observe differences between groups of texts and subcorpora. The most popular exploratory technique is principal component analysis (PCA). However, we decide for correspondence analysis (CA) that is conceptually similar to PCA, with the difference that the data is scaled so that rows and columns are treated equivalently. In this
way, can see not only which dependent variables (in our case languages, registers, methods) have similarities, but also possible correlation of these dependent variables with the independent ones. In our case, these are features under analysis that contribute to these similarities. In CA, distances between dependent and independent variables are calculated and represented in a two-dimensional map, and the larger the differences between subcorpora or texts, the further apart they are on the map, and likewise, dissimilar categories of features are further apart. The correlations between dependent and independent variables are transformed into a set of uncorrelated variables, called principal axes or dimensions. The first two principal axes account for as much variation as possible in two dimensions. In this way, we can identify new meaningful underlying variables, which ideally correlate with such variables as language, register, text production type, i.e. the dimensions of language variation in our study.

We also exploit automatic classification and clustering techniques used in both supervised and unsupervised scenarios. Unsupervised methods allow us to discover ‘interesting structures’ in the data, for instance, groupings of subcorpora according to the different dimensions of language variation. In a supervised scenario, which is more useful for a diversification of dimensions, we would rather define the classes and the features specific for them, in the unsupervised case, we are free to choose the number of clusters as we like. Applying this technique, we can find hidden structures in the data. Supervised classification performs better with single register data, so that in a supervised scenario, we need to perform several classification tasks, if we want to analyse variation along the corresponding dimension.

Hierarchical cluster analysis (HCA) was used in Articles 9 and 11 of this volume. In Article 11, we used this technique to analyse differences and similarities between translation varieties in the above described corpus VARTRA. Moreover, analysing various number of classes, we we also able to discover prominence of a one of the dimensions in our dataset. In Article 9, we observe differences between groups of translated and non-translated texts and subcorpora. The results of automatic clustering indicate differences and similarities between English and German and their registers. In the same way, differences between non-translated and (manually or automatically) translated texts can be discovered. Unsupervised techniques should perform better with multi-register data.

Supervised classification techniques were exploited in Articles 10, 12, 13 and 14. We used different types of classifiers, e.g. Naive Bayes and Likelihood-based ones in Article 10, KNN in Articles 13 and 14 and SVM in Articles 12, 13 and 14. Supervised text classification techniques are used to observe more fine-grained differences between the subcorpora under analysis with respect to the analysed features. We can inspect in detail the whole range of features that make the pre-defined variables (classes) distinct from one another. The classes defined can
be 1) languages (English and German); 2) registers (FICTION, POPSCI, etc.); 3) discourse modes (spoken vs. written); 4) text production types (originals vs. translations); 5) translation methods (human or machine). 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. The most distinctive features can then be drawn by the observation of how well they contribute to the distinction of specific classes. For this, the SVM weights can be judged: the higher the weight of a feature, the more distinctive it is for a particular class in the respective classification task, as it was done in Article 10 to sort out features for the linguistic interpretation.

Classification techniques are suitable for the analysis of interaction between the dimensions under analysis. In Article 11, where we analyse the interaction between the dimensions of register and method in translations, we deploy automatic clustering. Article 13 (where supervised classification is applied) demonstrates the distinctiveness of registers not only in translations, but also in comparable originals. Here, the interplay between text production type, register and translation method is analysed.
0.4 Findings

In the present section, we summarise the findings that are presented in the articles of the present volume. The findings are related to the language variation along the dimensions under analysis. In most cases, these findings were obtained for the language pair English and German, and additionally for German/English and Czech. We, however, believe that some of the variation phenomena observed for these language pairs can also be related to other languages, if similar features are investigated.

Section 0.4.1 is dedicated to the variation along the language dimension. Register variation is described in Section 0.4.2, whereas the results on the differences between text production types are summarised in 0.4.3. We do not provide the results on the variation within translation in a separate section, since it is mostly covered in the sections dedicated to the first three dimensions.

0.4.1 Language Contrast

Language contrasts are analysed in this volume within the variation along the language dimension. In Articles 5, 6, 7 and 9 of this volume, we present contrastive analyses on English and German, whereas their contrasts to Czech were studied in Articles 3 and 4. In most cases, our focus was on the features related to the level of text or discourse, since the other existent works have mostly been concerned with lexico-grammar, ignoring the typological, structural and functional factors affecting the linguistic level of text across different languages. Nevertheless, the observed textual contrasts on the level of discourse go hand in hand with the typological differences existing on the lexico-grammatical level. For instance, the overall predominance of the demonstrative reference in German or verbal ellipsis in English can be justified by the greater number of demonstrative devices in German and the possibilities to omit the full verb in a verbal phrase, e.g. in a negation. However, these preferences are subject to variance in different contexts that we are able to observe analysing distributions of these linguistic categories in various registers of our corpus data.

We decide for a higher level of linguistic description in our contrastive analyses. The more abstract linguistic categories in the language comparison allow us to overcome the direct influence of typological differences in the language repositories. For this, we elaborate analysis schemes which are underlying in the multilingual annotation task. For instance, in Article 6, we describe all language means that English and German have at their disposal to express five main cohesive devices: coreference, conjunctive relations, substitution, ellipsis and lexical cohesion. For each of them, we also define functional and structural subtypes, e.g. demonstrative reference with a function of head or modifier, conjunctive re-
lations with additive meaning expressed with connects or subjuncts, nominal or verbal substitution and others, see Table 3 in Article 6 for details. The definition of the more fine-grained categories in form of subtypes is important for a detailed analysis of particular phenomena. For instance, in Article 7, we provide a thorough analysis of conjunctive relations in English and German, paying attention to their structural and positional peculiarities. These features also impose us on the elaboration of procedures to effectively extract the information on them from the corpus. What we find out is that in English, most adverbials realise temporal relations, while in German highest distributions are measured for causal meanings.

In Articles 3 and 4, we also include Czech into our analysis which imposes the formulation of even higher categories. Another problem of the study in this article is the usage of resources annotated within two different frameworks. We able to solve these problem by formulation of an interoperable analysis schemes including the categories available not only in the three languages under analysis, but also in the involved annotation schemes. This scheme includes four types of relations and their subtypes: identity, non-identity, discourse (logico-semantic) relations and ellipsis as illustrated in Table 3 of Article 4. Applying this scheme, we can show that taken all the formulated categories, German and Czech show similarities in the degree of cohesiveness, which means that they do not differ in the amount of the explicit means for building up a discourse. However, the differences appear if the distributions of the four categories are concerned. Here, Czech shows preferences for identity, whereas in German logico-semantic relations predominate (at least in the data set at hand). The preference of the German texts to explicitly realise logico-semantic relations was also observed in our English-German comparison, as stated in the conclusions of Article 5. nevertheless, the creation of the identity relations are also important for German, which is confirmed in both contrastive analyses.

The differences on the level of lexico-grammar are not ignored in our analyses. On the contrary, they play a great role in the explanation of the observed discrepancies. This explanation is achieved by zooming in into the more fine-grained definition of the phenomena. For instance, in Article 4, we find that the higher frequencies of the identity relations in Czech exclusively stem from the nominal coreference (see Table 4 in Article 4). In German, like in other languages with the definite article, most coreferring expressions contain a formal definite marker which allows to (even automatically) extract most anaphors from the corpus (except named entities, which can also be automatically identified with state-of-the-art tools). The definite article does not exist in Czech (and other Slavic languages), which means that Czech has less explicit accessibility markers serving as triggers of coreferring expressions. Accessibility of referents is indicated by information structure more often than in German. In many cases, the annotation is completed on the base of semantic criteria (often depending on the annotator choice). The
0.4 Findings

differences in the category of definiteness, as well as further differences of typological character, e.g. prodrops (existing in Czech) are also influence on the properties of coreference chains in the two languages. In the data at hand, the German texts have more two-element chains (69%) than the Czech ones (62%). The qualitative analysis of the chains show that longer chains in the Czech data include a great number of named entities. In the German data, only explicitly marked named entities are taken into consideration (due to the concept of cohesive device as a trigger of relations), and the others belong to the chains of lexical cohesion. However, the main differences in the length and the number of chains which reflect the topic structuring are not of typological or methodological character. They are rather register- or, in some cases, domain-dependent.

Further differences arise when the scope of relations and their realisation is concerned. For instance, coreference to abstract entities (verbal phrases, clauses, sentences or even text passages) can be expressed through various linguistic means in English, German and Czech. However, German shows a clear preference for demonstrative heads, including pronominal adverbs, whereas Czech also extensively utilises non-modified nominal phrases (as in example (3)), which is never the case in German.

(3) *Dnes se tento počet snížil na asi pul milionu, jenže důvodem poklesu je především skutečnost, že ten, kdo není zaměstnán déle než rok, již podporu nedostane. [Today, that number dropped to about half a million, but the reason for the decline is the fact that anyone who is not employed for more than a year, gets no support anymore.]*

In English, this kind of relation is often realised with personal heads, i.e. this/that.

As we already mentioned above, English have a clear preference for verbal ellipsis, whereas German operates rather with nominal and clausal ones. In the German-Czech comparison, the German texts lose against the Czech ones in the number of nominal ellipsis. This indicates the preference for expressing comparison by fragments in both languages. However, due to the greater syntactic flexibility of Czech relative to German, this language has a stronger tendency towards implicitness.

However, the differences attested in terms of the preferred meaning relations, as well as in the structural and functional subtypes for realising similar meaning relations are subject to register, domain or mode variation. This was especially shown in Articles 5 and 9 for German and English. Generally, we can state that the contrasts are more pronounced between the languages than between the analysed registers. The mode of production also plays an essential role for the similarities between particular registers in each language separately and also across languages (see Article 5). We address the register-dependent variation in the next section.
0.4 Findings

0.4.2 Register Variation

Registerial contrasts are analysed in this volume within both intralingual and interlingual variation. Article 5, 7 and 9 focus on register variation in English and German. Articles 10 and 11 cover the variation along this dimension in translated texts. And in Articles 13 and 14, register variation in German translation is compared to this type of variation in original German texts. Overall, some of our analyses show the predominance of this dimension of variation above other dimensions. In many cases register variation is stronger than others, especially if translated texts are concerned, see Articles 10 and 11. This also refers to original, non-translated texts, see Articles 5 and 7.

The strength of the register dimensions is present in the cases where we observe differences between registers, not regarding in which language the texts are produced. In many analyses included into this volume, this observation is related to the fictional texts of our data, see Article 5, 9 and 11. Narrative style is one of the main peculiarities of FICTION. It is reflected especially in the preference for coreference with personal pronouns, as shown in Section 4 of Article 5.

The results of the CA analysis illustrated in Figure 2 in Section 4.1 of Article 9 suggests that language contrasts are stronger if the given features are concerned. The graph in Figure 2 represents two dimensions of variation: 1) the one of language along the x-axis, and 2) the one of register along the y-axis. The first dimension covers over 60% of the data variance, whereas the second one – around 20% (represented in Table 3 in Section 4.1 of Article 9). This means the predominance of the language contrasts in the data set at hand. However, we suppose that the prevalence of the the language vs. register contrasts depend on the features analysed. Moreover, the differences between languages and registers are often tied to further varying contextual configurations, i.e. those of mode, grouping registers to spoken or written ones. At the same time, more differences emerge, if the varying number of speech participants, the formal type of conversation (private vs. public), time laps between production and reception and other contextual settings are concerned. For this reason, the differences between modes are getting more prominent (see Section 4.2.2 in Article 5), and in some cases, dominating over the other dimensions of variation, e.g. language, see Section 0.5 below.

In some of the recent results on the coreference and lexical chains for German and Czech, as well as English and German, we could state register- and mode-sensitive differences (see Section 0.4.1 for the details on German-Czech). In our analyses of the corpus GECCOCHAIN described in Section 0.3.1 in terms of chains of relations (coreference and lexical), we could also observe register-specific differences between texts not regarding the language, see the CA plot in Figure 1. The blue points indicate groupings of registers, the red triangles – the features contributing to these groupings. The position of the triangles and points
0.4 Findings

Figure 1: CA for chains of relations in a multilingual dataset

also indicates the relative importance of a feature for a register.

We observe two multilingual groupings on the plot along the x-axis (that covers ca. 63% of the data variance): fictional texts and interviews on the left side, and political essays and popular-scientific texts on the right side. On the one hand, this might suggest spoken vs. written register settings (since fiction contains spoken elements). On the other hand, this might also indicate classification to further contextual parameters, e.g. narrative style on the left side, and more informational and/or expository type of production on the right side. This assumption is confirmed in terms of the contributing features: coreference related on the left side, and those related to lexical cohesion (less reduced cohesive devices) on the right side of zero. Altogether, this shows that registerial differences are more pronounced than language differences, at least if relations expressed through coreference and lexical cohesion are concerned.

Using the same set of features, we prove this tendencies with a correlation plot based on a correlation (or distance) matrix. For this, row (multilingual register) and column (chain/relation features) profiles are calculated by taking each row/column point and dividing by the sum of all row/column points. Then, the squared distance is computed between each row/column profile and the other rows/columns in the table, resulting in a distance matrix which is visualised with a correlation plot in Figure 2. The size and the colour of the circle in the plot is proportional to the magnitude of the distance between register profiles. We see that cross-lingual differences between registers (e.g. EO-ESSAY vs. GO-ESSAY, EO-
0.4 Findings

Figure 2: Correlation between multilingual register profiles

FICTION vs. GO-FICTION) are smaller than intralingual distances between registers of one language (e.g. GO-FICTION vs. GO-ESSAY or vs. EO-POPSCI). This analysis confirms our observations on the prevalence of the variable register for the features and data at hand.

The idiosyncracy of FICTION is also confirmed with our monolingual experiments in Article 13, where we use a different set of features (Table 1 in Article 13). In this study, we attempt to model German registers on the basis of original texts by using two classifiers (KNN and SVM, see Section 3.4 in Article 13). In the classification task, registers correspond to classes. As seen from Table 2 of this Article, the classifiers perform best for fictional texts, for which we achieve 100% f-measure. This means that this register seems to be easy to model, and it differs from the other registers in the data through the distributions of the features under analysis. The performance of the classifier is dependent on the nature of the registers involved. Some of them are more difficult to model than others, see the classifier results for SHARE or SPEECH. Interestingly, these registers turn to be similar in our German data, and their features of coincide in the analysed translations. In Article 11, we also observe the phenomena of shining through and normalisation for these two registers indicating that the features of both source and target languages interact in translations of these registers. In Neumann’s description of register profiles, these two registers would coincide if we take into account both languages, see profiles of the English and German original registers in Neumann, see Tables in (Neumann, 2013, p. 309 and p. 313) for details. In
0.4 Findings

Article 10, we also apply classification techniques and show that classification works better between registers than between translation methods (see Table 1 in Section 4.1 of the article). This tendency was also observed in Article 11: comparing the distance indication on the scale in the two figures (Figure 1 and 2 in Section 4.1 of the article), we find that the difference between register clusters is greater than that between translation methods. Register-dependent clusters are also observed in Figure 7 of Article 9, which confirms the assumption that register is more prominent than translation method, i.e. there are more differences between various registers than between human and machine translations in the data under analysis, if discourse properties are concerned. In general, we believe that the influence of the register dimension is stronger in translated texts than in non-translated ones, since for the latter, we observed different tendencies in our interlingual analyses depending on the features involved.

0.4.3 Text Production Type

Most all the articles on translation included in this volume are dedicated to the variation along the text production type reflected in the differences between originals and translations (Article 2, 7, 8, 9, 12, 13 and 14).

In Articles 8 and 9, we describe features of translation variants produced by humans and machines, comparing them to their English source texts, as well as comparable German originals. With the help of selected lexico-grammatical patterns, we are able to trace differences and similarities between them, which indicate features of translationese, such as shining through and normalisation, simplification and explicitation, as well as convergence. We also showed that these features may vary if we consider translation method varieties separately.

We analyse the contrasts between translated and non-translated texts applying HCA, see Section 4.2 of Article 9. Figure 4 in this article shows that the data under analysis is clustered into originals (on the right side) and translations (on the left side). Text production type is apparently the most prominent difference in this data, and we can automatically separate originals from translations. The automatic classification of translated and non-translated texts have already been shown in a number of studies, see Section 0.2.3 above. Our results show that this discrimination is also possible with discourse features, which means that translations differ from originals also in these properties. Other empirical translation studies on the English-German language pair also address shifts in cohesion as an indicator of translationese, and especially of explicitation, e.g. (Becher, 2011). However, we believe that we need to elaborate further operationalisations for the analysis of explicitation. In many studies, this translation feature was measured by the amount of explicit cohesive devices in the source and the target texts. Since the process of explicitation is often accompanied by the phenomenon of implici-
0.4 Findings

tation (as it was shown by Becher (2011)), a quantitative analysis of parallel data is required. We need to identify explicit linguistic means in the target texts and analyse if they were implicit in the source. This type of data (containing quantitative information on the implicit-explicit transfers) would provide us with the necessary information on the feature of explicitation, which is not yet possible with the data at hand.

In our studies, we also make an attempt to identify shining through or normalisation. Within the experiments in Article 9, this is achieved by answering the question: which language influences the automatic grouping – the source or the target one? For this, two additional clustering tests are separately performed: 1) for German translations and their English source texts (Figure 5 in Section 4.2 of Article 9), and the same German translations with the German comparable originals (Figure 6 in the same section of Article 9). The results show that in both cases, the data is separated into translations and originals, so, no shining through or normalisation effect can be detected in our data in terms of the discourse-related features.

We were not able to detect shining through in another study described in Article 8. Here we used a number of shallow features with the help of which we could observe simplification, explicitation, normalisation and convergence. However, neither type–token–ratio nor lexical density, which are extensively used in other corpus-based translation studies, serve as good indicators of simplification in the analysed data. The hypotheses about normalisation and shining through can be confirmed only in part. These features turn to be sensitive to translation method, since we observe some variation across our translation varieties.

At the same time, translation varieties demonstrate a great number of similarities. In article 8 we even show that the analysed translation varieties converge, as no significant difference is observed between them in terms of the analysed phenomena. This observation is also supported by another study in Article 12, in which we make use of other frameworks as features sources. We use here information theory (surprisal) and MT quality estimation inspired features with the help of which we are able to achieve good results in the automatic classification between original and translated sentences (up to ca. 70% of accuracy). This means that these features are useful in the identification of the text type. In the same paper, we perform another experiment aiming at the automatic classification of translations into those produced by students and those produced by professionals. The performance of the classifier here is much lower than in the previous test. These results show the similarity of the two translation varieties in terms of the feature involved. Evaluation of the feature groups show that surprisal features do not belong to good indicators of convergence, i.e. we are not able to show the similarities between translation varieties in terms of these features.

The features inspired by register theories seem to be shared by different trans-
lations, even those produced with different methods. We show this in Articles 13 and 14, where we train classifiers for register-based classes using German non-translated texts. These classifiers are then tested on human and machine translations. The results show that in most cases, register settings of translations do not comply with the register settings of originally written German regardless the method involved in translation process. This indicates again that translated texts are closer between each other than to the comparable originals. In Article 14, we perform an additional test, in which we use human translations to train a classifier and test it on machine translations. We show that automatic classification works better, if translated texts are used as training data. This points to more compliance between human and machine translations than between originals and machine translations.

Although human and machine translations converge in many cases, especially if the comparison to non-translated texts are involved, we also observe a number of differences between translations produced with these two translation methods. In Article 13, we show that register classification performs better for human translations. Whereas the difference is not significant in terms of the KNN classifier, it is significant for the SVM results (see Section 4.4 of Article 13). At the same time, Figure 1 shows that the differences between human and machine translation are register-sensitive. The similarities of machine-generated translations with those produced by humans can be explained by the fact that human translations are used as training data in the development of MT.

At the same time, in some of our analyses, we also noticed that only certain varieties of human and machine translations display similarities. For instance, in Article 11, we state that student translations, as well as those produced with a system trained with little data vary strongly from other translation varieties. Our assumption is that this variation can be explained by the dimension of variation that we call experience in our study (see Section 0.2.4 above). The compliance between these two translation varieties lies in the amount of experience or data involved: degree of expertise of human translators and the amount of training data in the statistical machine translation system. The variation along this dimension needs further analyses, since the features considered so far do not seem to have good discriminatory power here. As already mentioned above, even MT quality inspired features do not work well in the differentiation of the levels of translation expertise, as it was tested for human translations in Article 12.

0.5 Final Discussion and Outlook

The analyses summarised in this volume have shown that there are several dimensions influencing linguistic variation, when considered from a multilingual perspective. In many cases, it is difficult to focus on one dimension only and ex-
0.5 Final Discussion and Outlook

Incluude the others, since all of them can come into interaction promoting emergence of specific phenomena.

At the same time, we also stated the predominance of one or the other dimension in our data. According to our observation, this predominance depends on the features involved in the analysis. For example, the predominance of the register dimension in translated data was showed with the features inspired by Register/Genre Theory, and also with some discourse-related language devices. We demonstrate the assumed feature sensitiveness in an experiment performed on the corpus data from GECCOCOH2014 (described in Section 0.3.1 above). We perform automatic clustering with HCA based on bootstrap resampling (producing p-value-based clusters), see details in this technique in Section 3.3 of Article 9.

Figure 3 illustrates the dendrogram produced on the basis of various cohesive devices and their subtypes annotated in this corpus (see Table 3 in Article 6 for details). This dendrogram contains two automatically produced clusters containing both multilingual (EO-POPSCI and GO-ESSAY) and monolingual (EO-ESSAY and EO-WEB) subcorpora on their smaller nodes. In one case they are register-specific (EO- and GO-FICTION) and in two cases they are mode-specific (IN-
TERVIEW and ACADEMIC). Yet, no general assumptions can be induced from this graph.

Cluster dendrogram with AU/BP values (%)

Figure 4: HCA tree based on the cohesive triggers of similarity relations

Excluding all cohesive features except those triggering similarity (related to comparative reference, substitution and ellipsis), we produce another dendrogram illustrated in Figure 4. In this case, we get three clusters: 1 multilingual class representing the spoken mode, and two language-specific clusters containing sub-corpora of written registers. This means that features related to similarity relations can serve as good indicators for the analysis of variation along the language dimension, as well as the subdimension of mode (analysed within register variation in our study). We suppose that different features can be grouped according to the dimensions of variation they indicate. Further analyses are needed to identify the discriminative features for every dimension comprising the model of multilingual variation.

As already mentioned, register seems to be more prominent in our translated data. Interestingly, the increased prominence of the register dimension in translated data was detected with various features: 1) features derived from Register/Genre Theory; 2) discourse/cohesion-related features and 3) data-driven fea-
tures in form of delexicalised n-grams. This dimension seems to dominate over other dimensions in translation, including language, experience and especially method. One of the explanation for this lies in the similarities between the analysed languages in terms of registers involved. Due to the fact that shining through and normalisation do not show strong presence in the data, translations are equally influenced by the source and the target language and develop their own features of rather source/target-independent character. Therefore, translations have their own registers, which is also confirmed by the discrepancies between translations and originals in terms of register features that we could show with the automatic classification experiments.

Further translation varieties, e.g. those differing in the level of expertise or translation method, show lower variation than those differing in register. Thus, convergence is not so strong between translations of different registers. The compliance of machine-translated texts to human translations can be easily explained by the technology behind, especially in case of statistical MT: SMT systems are trained on parallel data that contain texts translated by humans. This results in the overlap observed in our data for hypothesis (machine) and reference (human) translations. Both of them do not correspond to comparable originals in terms of register settings. Human translation characteristics in MT are often considered to be beneficial, since they can improve the BLEU score, if used in the corresponding direction in translation models, and used instead of non-translated texts in language models. These improvements in the BLEU score (quantitative compliance with the reference) do not necessarily lead to a better quality of machine translation.

In our approach, we use quantified register profiles (calculated on non-translated texts) as standards of the target language, which can be used to measure the quality of translated texts (both human and machine). The distinction between originals and translations performed on the basis of these profiles can be correlated with the quality of the translated texts: the closer a translation is to the target language in terms of register, the better quality it has.

Following the results from our analyses, we argue that register features should be integrated into MT evaluation process, as an additional layer to the already existing automatic metrics. In our future work, this hypothesis should be tested by combining and correlating the present results with the state-of-the-art evaluation metrics. Moreover, the application of manually translated texts as a reference should be treated with caution. We believe that a closer approximation of the automatic translations to the original texts is needed, instead of its adaptation to manual translations. However, the latter is a common practice in the up-to-date MT technology, since there exist limited possibilities of creating high-level language models capturing register profiles in a target language. However, their implementation, as well as exploitation of such profiles for MT development, need
a thorough elaboration of the features beyond the scope of the articles presented in this volume. Language models whose creation is feasible on the basis of the features analysed here represent cross-linguistic tendencies that the current MT procedures are not able to handle, since they need actual cross-linguistic correlations derived from parallel data.

However, the results of the analyses presented in this volume can find application not only in theoretical translation studies and contrastive linguistics but also in NLP, e.g. in MT. For the first two, they deliver valuable knowledge on variational phenomena providing insights on the influencing factors and emerging effects. For machine translation, they provide several possibilities of application. One of the ideas is the development of new techniques to evaluate machine translation. Here, register profiles (seen as conventions of the target language) can be used to rank translated texts.

The knowledge on similarities between certain registers can be helpful in different areas of NLP, where corpus resources are used to train systems. For instance, if we know that features of two registers are shared, then we can train a system on the data from one register and test it on the other, e.g. train an MT system on one register and automatically translate texts from another register whose features turn to be close to the first one.

In our future work, we aim at implementation of these ideas. We also intend to extend feature work and include a more fine-grained discriminatory analysis of features. We need to create dimension profiles for each dimension of language variation analysed here. For this, new methods of investigation should be explored, since those used in this analysis do not allow a comprehensive discrimination of features. The supervised techniques applied in the analyses in this volume allow feature discrimination in one particular classification task. With these techniques, we ignore possible relationships among the independent and dependent variables, i.e. among the features and dimensions. In future, we should test non-linear feature selection methods, e.g. ensemble of trees or other.
0.6 Authors’s Contribution to the Jointly-Authored Publications

In the following, we will provide brief information to the authorship of the section and subsections of jointly authored publications.

Article 1
- Lapshinova: 1, 3.1, 5
- Lapshinova and Kunz: 3.2, 4 and 6
- Kunz: 2

Article 3
- Lapshinova, Kunz and Nedoluzhko: 1, 3.3, 4 and 5
- Lapshinova and Kunz: 2.1
- Lapshinova: 3.1
- Nedoluzhko: 2.2, 3.2

Article 4
- Lapshinova, Kunz and Nedoluzhko: 1, 2, 4 and 5
- Lapshinova and Nedoluzhko: 3.1 and 3.2

Article 5
- Lapshinova: 3, 4.2, 5: Statistics
- Kunz: 1, 2, 4.1, 5: Final interpretation of findings

Article 6
- Lapshinova: 1, 3
- Lapshinova and Kunz: 2, 4, 5
0.6 Authors’s Contribution to the Jointly-Authored Publications

Article 7
- Lapshinova: 3, 4: procedures, extractions and part of the interpretations
- Lapshinova and Kunz: 2.2, 5
- Kunz: 1, 2.1, 4: interpretations

Article 10
- Lapshinova: 2, 3.1, 4.1-4.4: interpretation of the numbers, 4.5
- Lapshinova and Zampieri: 1, 3.2, 5
- Zampieri: 4.1-4.4: performance of experiments

Article 12
- Lapshinova: 2.1, 2.4, 4: interpretation of results
- Lapshinova, Rubino: hypotheses in 3, 3.2
- Lapshinova, Rubino, van Genabith: 1, 2.2, 5
- Rubino: 2.3, 3.1, 3.3, 3.4, 4: experiment conduction

Article 13
- Lapshinova: 1, 2.1, 2.2, 3.1, 3.2, 3.3
- Lapshinova, Vela: 2.3, 3.4, 4.1, 4.2, 4.3, 5

Article 14
- Lapshinova: 2.2, 2.3, 3.2
- Lapshinova, Vela: 1, 3.1, 3.3, 4.1, 5
- Vela: 3.4, 3.5, 4.2
References


De Sutter, G., Delaere, I., and Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis


48
REFERENCES


REFERENCES

Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies. Mouton de Gruyter. TILSM series.


REFERENCES


approach using register features. *Journal of the Association for Information Science and Technology*, pages n/a–n/a.


COLLECTION OF ARTICLES
Building Multilingual Resources

In this section, we...
Compiling a Multilingual Spoken Corpus

Ekaterina Lapshinova-Koltunski, Kerstin Kunz, Marilisa Amoia
Saarland University
Universität Campus 66123 Saarbrücken Germany
E-mail: e.lapshinova, k.kunz, m.amoia@mx.uni-saarland.de

Abstract

The present paper describes the compilation of the spoken part of an English-German corpus, which has been created for the investigation of cohesion. The corpus is one of the few existing resources supporting contrastive studies of cohesion and, to our knowledge, the only one permitting a contrastive analysis of spoken registers in the two languages. In addition, our corpus data offer further research potentials for contrastive linguistics and translation studies as well as for numerous NLP research areas.

Keywords: corpus compilation, spoken corpus, multilingual corpus, corpus annotation, cohesion.

1. Introduction

The present paper describes the compilation of the spoken part of an English-German corpus, which has been created for the investigation of cohesion. The corpus is one of the few existing resources supporting contrastive studies of cohesion and, to our knowledge, the only one permitting a contrastive analysis of spoken registers in the two languages. In addition, our corpus data offer further research potentials for contrastive linguistics and translation studies as well as for numerous NLP research areas.

1.1 Aims

The main objective of the present paper is to compile the spoken part of a multilingual corpus to investigate cohesion in German and English. Our long-term linguistic research interest is in the analysis of cohesive resources provided by both language systems and their instantiations in texts. More precisely, we are concerned with the exploration of contrasts in form, frequency and function of cohesive devices and meaning relations established to other textual elements. We aim to analyse these phenomena across and between languages, registers and modes.

1.2 Motivation

Comprehensive accounts of cohesion are only existent from a largely systemic and monolingual perspective, see e.g. (Halliday & Hasan, 1976; Brown & Yule, 1983; Schubert, 2008 and Esser, 2009) for English, and (De Beaugrande & Dressler, 1981; Vater, 2005; Brinker, 2005) for German. Empirical analyses (both monolingual and contrastive) in the area of cohesion mainly deal with individual cohesive devices, cf. (Bosch et al., 2007) or (Gundel et al., 2004). Empirical analyses of cohesion in spoken discourse exist for German, e.g. (Ahrenholz, 2007) and English, e.g. (Gundel et al., 2004 and 2005; Eckert & Strube 2001). These however, are limited to the investigation of individual phenomena, and mostly examine personal pronouns or demonstratives. To our knowledge, there is only one contrastive empirical analysis by (Schreiber, 1999) comparing English and German. It includes a relatively broad range of cohesive phenomena, however it uses excerpts of French and German corpora to illustrate particular phenomena rather than presenting a contrastive interpretation of findings from a statistical analysis.

These studies seem to suggest that particular cohesive devices exhibit a tendency to occur either in registers of spoken language only or with a much higher frequency than in written discourse, see e.g. (Schreiber, 1992; Ahrenholz, 2007). Our preliminary extractions from registers of written language \(^1\) underpin these observations. For instance, they show that occurrences of the German demonstrative pronouns der, die, das and particular constructions of substitution are rarely traced in typical registers of written language and appear with a much higher frequency in those written registers that approximate spoken language, such as fiction or political speeches\(^2\). In addition, dialogic sequences of our fiction subcorpus point to instantiations of cohesive ellipsis which seem to be restricted to spoken discourse. These first findings call for a corpus which allows to integrate differences between written and spoken registers so as to establish a comprehensive model of cohesion in English and German. To our knowledge, there are no corpus resources to support our research goal. The existing ones are either monolingual, e.g. ICE, cf. (Greenbaum, 1996) for English or “Deutsch heute”, cf. (Brinckmann, 2008) German, or compiled for special purposes, e.g. SCOTS corpus, cf. (Anderson, 2007) or Verbmobil, (Hirichs et al. 2000). Some of them also contain non-native data, e.g. ICLE described in (Granger, 2008) and LINDSEI, cf. (Gilquin et al., 2010).

2. Theoretical Background

There are substantial gaps in the area of text-based contrastive modeling for the two languages under analysis, especially text-based empirical accounts of mechanisms of textuality are absent. System-based text/discourse grammars commonly deal with specific questions of textuality only. While the literature in English mainly focuses on linguistic resources for establishing textuality, e.g. (Halliday & Hasan, 1976; Brown & Yule, 1983; de Beaugrande & Dressler, 1981), the German literature frequently takes as its starting point general pragmatic, cognitive and semantic principles of

---

1 cf. (Kunz et al., 2009; Klein, 2007 and Birster, 2007).
2 The extractions were done on the CroCo corpus, cf. (Neumann, 2005)
coherence, which are reflected in linguistic phenomena, cf. (Linke et al., 2001; Brinker, 2005; Vater, 2001). These methodological differences lead to noticeable differences in the range of phenomena considered. In general, monolingual text–discourse treatments inform us about the coherence-building systems of a language and are structured by type and/or function of the system (e.g. (co–) reference, conjunctive relation, lexical/semantic relations, etc.). We define cohesive resources (devices) as a set of lexico-grammatical items that function as resources allowing to transcend the boundaries of the clause. For our classification of general categories, we follow the one by (Halliday & Hasan, 1976), according to which cohesion includes five categories: reference, substitution, ellipsis, conjunctive relations, lexical cohesion.

3. Corpus Compilation

3.1 Data Collection

Our multilingual spoken corpus contains two registers: interview and academic speech. These registers are added to the eight registers of written language of the already existing corpus, cf. (Kunz & Lapshinova, 2011): popular-scientific texts, tourism leaflets, political essays, corporal communication, instruction manuals and websites, prepared speeches, fictional texts. Especially the latter two registers are considered to lie at the borderline between written and spoken discourse. In order to create the German-English spoken corpus, we extract parts of already existing speech corpora and collect our own data, cf. table 1.

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>German (GO)</th>
<th>English (EO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERVIEW</td>
<td>BACKBONE-DE</td>
<td>ELISA</td>
</tr>
<tr>
<td>ACADEMIC</td>
<td>GECO</td>
<td>MICASE</td>
</tr>
</tbody>
</table>

Table 1: Sources for the GECO spoken part.

For English, we use the data of the MICASE corpus, the English part of the BACKBONE corpus and the ELISA corpus. The Michigan Corpus of Academic Spoken English (MICASE) is a collection of nearly 1.8 million words of transcribed speech – almost 200 hours of recordings) from the University of Michigan and includes lectures, classroom discussions, lab sections, seminars, and advising sessions, cf. (Simpson et al., 2002). The BACKBONE pedagogic corpus contains corpora of video-recorded spoken interviews with native speakers of various European languages, cf. (Kohn 2011). The ELISA corpus contains interviews with native speakers of English talking about their professional career (e.g. in tourism, politics, the media or environmental education), cf. (Braun 2006). The data from the corpora were extracted according to criteria such as nationality of speaker, type of speech event, degree of speaker interaction. For German, we use the German part of the BACKBONE corpus, which contains interviews with German native speakers (including variants of German).

This subset is comparable to the interviews in ELISA and the English part of the BACKBONE corpus. In addition, we compile our own corpus of spoken academic discourse consisting of lectures from all departments of the Saarland University. The lectures were recorded by VISU (Virtual University of Saarland) and have been transcribed by our team according to the transcription guidelines described below.

3.2 Problems in Spoken Data Compilation

In the process of data collection for the German part of spoken academic discourse, we have encountered a number of practical problems. For instance, we initially planned to include recordings of seminars for the analysis of dialogues. However, the seminars in Germany turn out to be less interactive and dialectic than assumed and hence, do not correspond to their English counterparts. Moreover, the collected student presentations constitute prepared speech and thus lack the authentic character of spontaneous speech. Therefore, our German academic corpus currently consists of lecture recordings which are comparable in their speech conditions to the English lectures.

Besides that, we had to apply manual transcription methods which is very labour- and time-consuming. Yet, the recorded data was found to contain too much noise to permit an automatic transcription (speech recognition). Moreover, manual transcription requires the formulation of transparent transcription guidelines. Since the English data was transcribed according to differing guidelines we elaborate a consistent scheme for tags in both languages to annotate extra-linguistic information (example (1)), linguistic variants (example (2)) and repairs and repeats (example (3)).

(1) LAUGHTER:
<laugh>text<laugh>

CONTEXTUAL EVENTS:
<writing_on_board>text </writing_on_board>
<door>text
<break type="gasp">

(2) EO-INTERVIEW
<turn speaker="Lauren">text<alternative text="yep">Yes</alternative>
<alternative text="yeah">Yes</alternative>, absolutely. <alternative text="yeah">yes</alternative>, I <break/>
<alternative text="yeah">yes</alternative>, absolutely

GO-INTERVIEW
<turn speaker="Stefan">text<alternative text="Wem">Wenn wir </alternative>die Netz</alternative> haben
<alternative>, <alternative text="Wer">werden</alternative> die Netze <alternative text="gehoben">gehoben</alternative>, es sind Stellnetze.</turn>

(3) REPEAT:
so it's <repeat text="an awful">an awful</repeat> lot of different cultures, different religions, different countries that people are from, which is great.

REPAIR:
So <repair text="they're"> <break/> they do <repair text="it's">
In order to guarantee comparability in frequency and function of cohesive devices between the written and spoken registers we had to restrict each register to 10-14 texts with around 34 thousand tokens each. The existing registers of written language contain both comparable and parallel texts of English and German. However, for the spoken registers, only comparable texts are available, cf. table 1. One possible solution for obtaining aligned texts would be to create interpretations for the existing originals. Interpreted texts, however, are produced under very specific conditions and are affected by various constraints such as time pressure, limited short-term memory capacity, linearity and others, see e.g. (Gumul, 2010) and (Pöchhacker, 2001). They are not considered as reflecting spontaneous speech on the one hand and differ considerably from translations, on the other hand. We thus consider to integrate transcriptions of films and their synchronizations in our corpus, although these are subject to other limitations described, for example, by (Herbst, 1994) and (Döhring, 2006).

4. Corpus Annotation

The spoken registers of the multilingual corpus are annotated on the same level as its written part:

1) **token level**: words, lemmas, parts-of-speech;
2) **chunk level**: sentences, syntactic and semantic chunks and their grammatical functions;
3) **cohesion level**: cohesive devices and cohesive chains;
4) **text level**: registers;
5) **extra-linguistic level**: meta information.

The automatic annotations of parts-of-speech, chunks and their grammatical functions are obtained with the help of the Stanford Parser, cf. (Marneffe et al., 2006). Cohesive devices, such as conjunctive relations, personal and demonstrative reference, substitution, ellipsis and lexical cohesion, are semi-automatically annotated with a tool based on the YAC recursive chunker, cf. (Kermes, 2003) which utilises the CWB Perl-Modules developed within the framework of YAC, cf. (Kermes & Evert, 2001) and (Kermes & Evert, 2002). We also apply the MMAX tool, cf. (Müller & Strube, 2006) for the manual correction of these annotations. Disambiguation of cohesive devices is based on the analyses described in (Kunz & Steiner, in progress) and (Kunz, 2010).

We also aim at annotating reference and lexical chains in our corpus. For this, we apply one of the existing systems for coreference resolution, the Stanford Coreference Resolution System described by (Lee et al., 2011). Our preliminary evaluation tests, see (Amoia et al., 2012), show that the system does not perform with the desired accuracy. Therefore, we also plan to manually improve annotations for this category of cohesion.

The corpus metadata include not only the information on speaker, such as age, sex (female, male, unissex, undefined), profession (translator, teacher, professor, student, etc.) and role (interviewer, interviewee, lecturer, etc.), but also the information on register analysis: field (experiential domain and goal orientation – argumentation, exposition, instruction, narration, description and persuasion), tenor (number of speakers, agentive role – monologic or dialogic, social role – equal, up or down, social hierarchy – expert to expert, expert to layperson, layperson to expert, layperson to layperson, social distance – formal or not) and mode (language role – ancillary or constitutive, channel – graphic, phonic or electronic, and medium – written, written to be spoken, spoken).

5. Corpus Querying

The corpus can be queried with the Corpus Query Processor (CQP, (Evert, 2005)), which allows us to detect candidates for cohesive devices by means of regular expressions, offering several functionalities for extraction (e.g., context expansion) and sorting purposes (e.g., counting, grouping of results). CQP allows two types of attributes: positional (e.g. for part-of-speech and morphological features) and structural (e.g. for chunks, registers or extra-linguistic information). These attributes are employed for CQP-based queries which include string, parts-of-speech, chunk, register and further constraints, cf. table 2.

<table>
<thead>
<tr>
<th>Query elements</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>word=&quot;and&quot; &amp;</td>
<td>word <em>and</em> which is</td>
</tr>
<tr>
<td>cohesive_device=&quot;conj&quot; &amp;</td>
<td>cohesive</td>
</tr>
<tr>
<td>text_register=&quot;INTERVIEW&quot; &amp;</td>
<td>conjunction</td>
</tr>
<tr>
<td>tenor_numberOfSpeakers=&quot;2&quot; &amp;</td>
<td>in interviews only</td>
</tr>
<tr>
<td>speaker_ager=&quot;31-50&quot; &amp;</td>
<td>with 2 speakers only</td>
</tr>
<tr>
<td>tenor_socialRole=&quot;equal&quot;</td>
<td>aged between 31-50</td>
</tr>
<tr>
<td></td>
<td>in an equal social role</td>
</tr>
</tbody>
</table>

Table 2: Example of a CQP query

The present CQP query delivers a list of concordances, as shown in example for the cohesive conjunction *and* (4).

(4) 8: My name’s Norma Holt *and* I actually come from the Wirral Peninsula which is on the west coast of Liverpool, which is Lancashire...

29: which is Lancashire, *and* we have Cheshire on one side and north Wales on the other.

188: the nice seaside is, if you like, all the big houses are, and it’s more countryside, more of the farming...

296: However, over the years certainly it has changed *and* now it’s very much a Liverpool accent ...

304: ... now it’s very much a Liverpool accent *and*, you know, which I’m not saying I disapprove of it ...

325: I think it’s a lazy speech *and* you need to actually think about what you’re saying.

348: My nephew sometimes’ll speak to me in the Liverpool accent *and* I’ll say, please speak to me in English *"*.
Moreover, the sorting, counting and grouping functionality of CQP allows us to extract frequency information, as shown in table 3 (for English only as the German ACADEMIC part is still under construction). The obtained frequencies of cohesive phenomena can then be evaluated in terms of their distribution across registers, languages and modes. For instance, table 3 displays the frequencies per million words of all cohesive occurrences of the form one in its function as nominal substitute. What the table nicely illustrates is that some registers show more commonalities in their distribution of cohesive one than others, and most notably that there is a considerable difference in frequency between the spoken and the written registers of our subcorpus. In addition, the two registers FICTION and speech are closer to the spoken registers then others. This may be due to the fact that FICTION contains text passages imitating spoken dialog and that SPEECH was written to be spoken. Thus, ACADEMIC seems to be at one end of the spoken written continuum of our corpus and SHARE at the other end (at least as far as cohesive one is concerned) with FICTION and SPEECH taking a somewhat middle position.

<table>
<thead>
<tr>
<th>Register</th>
<th>Cohesive one per 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoken</td>
<td></td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>949,84</td>
</tr>
<tr>
<td>ACADEMIC</td>
<td>2769,33</td>
</tr>
<tr>
<td>FICTION</td>
<td>378,42</td>
</tr>
<tr>
<td>SPEECH</td>
<td>199,65</td>
</tr>
<tr>
<td>ESSAY</td>
<td>85,72</td>
</tr>
<tr>
<td>SHARE</td>
<td>83,74</td>
</tr>
<tr>
<td>INSTR</td>
<td>110,60</td>
</tr>
<tr>
<td>TOU</td>
<td>124,02</td>
</tr>
<tr>
<td>POPSCI</td>
<td>142,26</td>
</tr>
<tr>
<td>WEB</td>
<td>166,18</td>
</tr>
</tbody>
</table>

Table 3: Frequencies delivered by CQP

6. Conclusion and Future Work

We have compiled a spoken corpus for English and German that is enhanced with annotations on several linguistic and extra-linguistic levels. Our corpus architecture not only allows a text-based contrastive analysis of cohesion in German and English but also permits a comparison of various spoken and written registers. Therefore, the findings based on our resources will not only complement the existing research gaps in cohesion but also enrich contrastive grammars with a systematic account of discourse phenomena in written vs. spoken mode. Moreover, both the developed resources as well as our findings on cohesion will provide valuable insights for language teaching and translator training and will open up new research options for various fields. In the future, we aim at expanding corpus with further registers, e.g. internet forums, TV talk shows and reports. Besides that, we will develop further procedures to automatically annotate cohesive devices and relations. We also plan to enhance our spoken corpus with translations. The corpus will be available for querying online within the CLARIN-D initiative.

7. Acknowledgements

The project GECCo (German-English Contrasts in Cohesion) is supported by a grant from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We thank our colleagues in the GECCo team – Katrin Menzel and Erich Steiner for their assistance. Besides that, we are especially grateful to Hannah Kermes for providing the necessary perl script for adaptation.

8. References


VARTRA: A Comparable Corpus for Analysis of Translation Variation

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
A 2.2 Universität Campus
66123 Saarbrücken
Germany
e.lapshinova@mx.uni-saarland.de

Abstract
This paper presents a comparable translation corpus created to investigate translation variation phenomena in terms of contrasts between languages, text types and translation methods (machine vs. computer-aided vs. human). These phenomena are reflected in linguistic features of translated texts belonging to different registers and produced with different translation methods. For their analysis, we combine methods derived from translation studies, language variation and machine translation, concentrating especially on textual and lexico-grammatical variation. To our knowledge, none of the existing corpora can provide comparable resources for a comprehensive analysis of variation across text types and translation methods. Therefore, the corpus resources created, as well as our analysis results will find application in different research areas, such as translation studies, machine translation, and others.

1 Introduction: Aims and Motivation
Comparable corpora serve as essential resources for numerous studies and applications in both linguistics (contrastive language, text analysis), translation studies and natural language processing (machine translation, computational lexicography, information extraction). Many comparable corpora are available and have been being created for different language pairs like (a) English, German and Italian (Baroni et al., 2009); (b) English, Norwegian, German and French (Johansson, 2002); (c) written or spoken English and German (Hansen et al., 2012) or (Lapshinova et al., 2012).

However, comparable corpora may be of the same language, as the feature of ‘comparability’ may relate not only to corpora of different languages but also to those of the same language. The main feature that makes them comparable is that they cover the same text type(s) in the same proportions, cf. for instance, (Laviosa, 1997) or (McEnery, 2003), and thus, can be used for a certain comparison task.

As our research goal is the analysis of translation variation, we need a corpus which allows us to compare translations, which differ in the source/target language, the type of the text translated (genre or register) and the method of translation (human with/without CAT\(^1\) tools, machine translation). There are a number of corpus-based studies dedicated to the analysis of variation phenomena, cf. (Teich, 2003; Steiner, 2004; Neumann, 2011) among others. However, all of them concentrate on the analysis of human translations only, comparing translated texts with non-translated ones. In some works on machine translation, the focus does lie on comparing different translation variants (human vs. machine), e.g. (White, 1994; Papineni et al., 2002; Babych and Hartley, 2004; Popović, 2011). However, they all serve the task of automatic machine translation (MT) systems evaluation and use the human-produced translations as references or training material only. None of them provide analysis of specific (linguistic) features of different text types translated with different translation methods.

The same tendencies are observed in the corpus resources available, as they are mostly built for certain research goals. Although there exists a number of translation corpora, none of them fits our research task: most of them include one translation method only: EUROPARL (Koehn, 2005) and JRC-Acquis (Steinberger et al., 2006) – translations produced by humans, or DARPA-94 (White, 1994) – machine-translated texts only.

\(^{1}\)CAT = computer-aided translation
Moreover, they all contain one register only and, therefore, cannot be applied to a comprehensive analysis of variation phenomena.

Therefore, we decided to compile our own comparable corpus which contains translations from different languages, of different text types, produced with different translation methods (human vs. machine). Furthermore, both human and machine translations contain further varieties: they are produced by different translators (both professional and student), with or without CAT tools or by different MT systems.

This resource will be valuable not only for our research goals, or for research purposes of further translation researchers. It can also find further applications, e.g. in machine translation or CAT tool development, as well as translation quality assessment.

The remainder of the paper is structured as follows. Section 2 presents studies we adopt as theoretical background for the selection of features and requirements for corpus resources. In section 4, we describe the compilation and design of the comparable translation corpus at hand. In section 5, we demonstrate some examples of corpus application, and in section 6, we draw some conclusions and provide more ideas for corpus extension and its further application.

2 Theoretical Background and Resource Requirements

To design and annotate a corpus reflecting variation phenomena, we need to define (linguistic) features of translations under analysis. As sources for these features, we use studies on translation and translationese, those on language variation, as well as works on machine translation, for instance MT evaluation and MT quality assessment.

2.1 Translation analysis and translationese

As already mentioned in section 1 above, translation studies either analyse differences between original texts and translations, e.g. (House, 1997; Matthiessen, 2001; Teich, 2003; Hansen, 2003; Steiner, 2004), or concentrate on the properties of translated texts only, e.g. (Baker, 1995). However, it is important that most of them consider translations to have their own specific properties which distinguish them from the originals (both of the source and target language), and thus, establish specific language of translations – the translationese.

Baker (1995) excludes the influence of the source language on a translation altogether, analysing characteristic patterns of translations independent of the source language. Within this context, she proposed translation universals – hypotheses on the universal features of translations: *explicitation* (tendency to spell things out rather than leave them implicit), *simplification* (tendency to simplify the language used in translation), *normalisation* (a tendency to exaggerate features of the target language and to conform to its typical patterns) and *levelling out* (individual translated texts are alike), cf. (Baker, 1996). Additionally, translations can also have features of “*shining through*” defined by Teich (2003) – in this case we observe some typical features of the source language in the translation. The author analyses this phenomena comparing different linguistic features (e.g. passive and passive-like constructions) of originals and translations in English and German.

In some recent applications of translationese phenomena, e.g. those for cleaning parallel corpora obtained from the Web, or for the improvement of translation and language models in MT (Baroni and Bernardini, 2005; Kurokawa et al., 2009; Koppel and Ordan, 2011; Lembersky et al., 2012), authors succeeded to automatically identify these features with machine learning techniques.

We aim at employing the knowledge (features described) from these studies, as well as techniques applied to explore these features in the corpus.

2.2 Language variation

Features of translated texts, as well as those of their sources are influenced by the text types they belong to, see (Neumann, 2011). Therefore, we also refer to studies on language variation which focus on the analysis of variation across registers and genres, e.g. (Biber, 1995; Conrad and Biber, 2001; Halliday and Hasan, 1989; Matthiessen, 2006; Neumann, 2011) among others. Register is described as functional variation, see Quirk et al. (1985) and Biber et al. (1999). For example, language may vary according to the activity of the involved participants, production varieties (written vs. spoken) of a language or the relationship between speaker and addressee(s). These parameters correspond to the variables of...
field, tenor and mode defined in the framework of Systemic Functional Linguistics (SFL), which describes language variation according to situational contexts, cf. e.g. Halliday and Hasan (1989), and Halliday (2004).

In SFL, these variables are associated with the corresponding lexico-grammatical features, e.g. field of discourse is realised in functional verb classes (e.g., activity, communication, etc) or term patterns, tenor is realised in modality (expressed e.g. by modal verbs) or stance expressions, mode is realised in information structure and textual cohesion (e.g. personal and demonstrative reference). Thus, differences between registers or text types can be identified through the analysis of occurrence of lexico-grammatical features in these registers, see Biber’s studies on linguistic variation, e.g. (Biber, 1988; Biber, 1995) or (Biber et al., 1999).

Steiner (2001) and Teich (2003) refer to registers as one of the influencing sources of the properties of translated text. Thus, we attempt to study variation in translation variants by analysing distributions of lexico-grammatical features in our corpus.

2.3 Machine translation

We also refer to studies on machine translation in our analysis, as we believe that translation variation phenomena should not be limited to those produced by humans. Although most studies comparing human and machine translation serve the task of automatic MT evaluation only, cf. (White, 1994; Papineni et al., 2002; Babych and Hartley, 2004), some of them do use linguistic features for their analysis.

For instance, Popović and Burchardt (2011) define linguistically influenced categories (inflections, word order, lexical choices) to automatically classify errors in the output of MT systems. Specia (2011) and Specia et al. (2011) also utilise linguistic features as indicators for quality estimation in MT. The authors emphasize that most MT studies ignored the MT system-independent features, i.e. those reflecting the properties of the translation and the original. The authors classify them into source complexity features (sentence and word length, type-token-ratio, etc.), target fluency features (e.g. translation sentence length or coherence of the target sentence) and adequacy features (e.g. absolute difference between the number of different phrase types in the source and target or difference between the depth of their syntactic trees, etc.).

3 Methodology

Consideration of the features described in the above mentioned frameworks will give us new insights on variation phenomena in translation. Thus, we collect these features and extract information on their distribution across translation variants of our corpus to evaluate them later with statistical methods.

Some of the features described by different frameworks overlap, e.g. type-token-ratio (TTR) or sentence length as indicator for simplification in translation analysis and as a target fluency feature in MT quality estimation; modal meanings and theme-rheme distribution in register analysis and SFL, or alternation of passive verb constructions in register analysis and translation studies.

Investigating language variation in translation, we need to compare translations produced by different systems with those produced by humans (with/without the help of CATs). Furthermore, we need to compare translated texts either with their originals in the source or comparable originals in the target language. Moreover, as we know that text type has influence on both source and target text (Neumann, 2011), we need to compare different text registers of all translation types.

This requires a certain corpus design: we need a linguistically-annotated corpus for extraction of particular features (e.g. morpho-syntactic constructions); we need to include meta-information on (a) translation type (human vs. computer-aided vs. machine, both rule-based and statistical), (b) text production type (original vs. translation) and (c) text type (various registers and domains of discourse). This will enable the following analysis procedures: (1) automatic extraction, (2) statistical evaluation and (3) classification (clustering) of lexico-grammatical features.

4 Corpus Resources

4.1 Corpus data collection

Due to the lack of resources required for the analysis of translation variation, we have compiled our own translation corpus VARTRA (VARiation in TRAnslatation). In this paper, we present the first version of the corpus – VARTRA-SMALL, which is the small and normalised version used for our
first analyses and experiments. The compilation of the full version of VARTRA is a part of our future work, cf. section 6.

VARTRA-SMALL contains English original texts and variants of their translations (to each text) into German which were produced by: (1) human professionals (PT), (2) human student translators with the help of computer-aided translation tools (CAT), (3) rule-based MT systems (RBMT) and (4) statistical MT systems (SMT).

The English originals (EO), as well as the translations by professionals (PT) were exported from the already existing corpus CroCo mentioned in section 1 above. The CAT variant was produced by student assistants who used the CAT tool ACROSS in the translation process. The current RBMT variant was translated with SYSTRAN (RBMT1), although we plan to expand it with a LINGUATEC-generated version. For SMT, we have compiled two versions – the one produced with Google Translate (SMT1), and the other one with a Moses system (SMT2).

Each translation variant is saved as a subcorpus and covers seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters of share-holders (SHARE), prepared political speeches (SPEECH), and touristic leaflets (TOU), presented in Table 1. The total number of tokens in VARTRA-SMALL comprises 795,460 tokens (the full version of VARTRA will comprise at least ca. 1,7 Mio words).

4.2 Corpus annotation

For the extraction of certain feature types, e.g. modal verbs, passive and active verb constructions, Theme types, textual cohesion, etc. our corpus should be linguistically annotated. All subcorpora of VARTRA-SMALL are tokenised, lemmatised, tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations were obtained with Tree Tagger (Schmid, 1994).

In Table 2, we outline the absolute numbers for different annotation levels per subcorpus (translation variant) in VARTRA-SMALL.

VARTRA-SMALL is encoded in CWB and can be queried with the help of Corpus Query Processor (CQP) (Evert, 2005). We also encode a part of the meta-data, such as information on register, as well as translation method, tools used and the source language. A sample output encoded in CQP format that is subsequently used for corpus query is shown in Figure 1.

In this way, we have compiled a corpus of different translation variants, which are comparable, as they contain translations of the same texts produced with different methods and tools. Thus, this comparable corpus allows for analysis of contrasts in terms of (a) text typology (e.g. fiction vs. popular-scientific articles); (b) text production types (originals vs. translations) and (c) translation types (human vs. machine and their subtypes).

Furthermore, examination of some translation phenomena requires parallel components – alignment between originals and translations. At the moment, alignment on the sentence level (exported from CroCo) is available for the EO and PT subcorpora. We do not provide any alignment for further translation variants at the moment, although we plan to align all of them with the originals on word and sentence level.

4.3 Corpus querying

As already mentioned in 4.2, VARTRA-SMALL can be queried with CQP, which allows definition of language patterns in form of regular expressions based on string, part-of-speech and chunk tags, as well as further constraints. In Table 3, we illustrate an example of a query which is built to extract cases of processual finite passive verb constructions in German: lines 1 - 5 are used for passive from a Verbzweit sentence (construction in German where the finite verb occupies the position after the subject), and lines 6 - 10 are used for Verbletzt constructions (where the finite verb occupies the final position in the sentence). In this example, we make use of part-of-speech (lines 3a, 5, 8 and 9a), lemma (lines 3b and 9b) and

<table>
<thead>
<tr>
<th>subc</th>
<th>token</th>
<th>lemma</th>
<th>chunk</th>
<th>sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>132609</td>
<td>9137</td>
<td>55319</td>
<td>6525</td>
</tr>
<tr>
<td>CAT</td>
<td>139825</td>
<td>10448</td>
<td>58669</td>
<td>6852</td>
</tr>
<tr>
<td>RBMT</td>
<td>131330</td>
<td>8376</td>
<td>55714</td>
<td>6195</td>
</tr>
<tr>
<td>SMT1</td>
<td>130568</td>
<td>9771</td>
<td>53935</td>
<td>6198</td>
</tr>
<tr>
<td>SMT2</td>
<td>127892</td>
<td>7943</td>
<td>51599</td>
<td>6131</td>
</tr>
</tbody>
</table>

Table 2: Annotations in VARTRA-SMALL

Lapshinova (2013) Article 2
chunk type (lines 2b and 6b) information, as well as chunk (lines 2a, 2c, 6a and 6c) and sentence (lines 1 and 10) borders.

<table>
<thead>
<tr>
<th>query block</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><code>&lt;s&gt;</code></td>
</tr>
<tr>
<td>2a.</td>
<td><code>&lt;chunk&gt;</code></td>
</tr>
<tr>
<td>2b.</td>
<td><code>[chunk_type=&quot;NC&quot;]+</code></td>
</tr>
<tr>
<td>2c.</td>
<td><code>&lt;/chunk&gt;</code></td>
</tr>
<tr>
<td>3a.</td>
<td><code>[pos=&quot;VAFIN&quot;&amp; lemma=&quot;werden&quot;]</code></td>
</tr>
<tr>
<td>3b.</td>
<td><code>[word != &quot;:&quot;]#</code></td>
</tr>
<tr>
<td>4.</td>
<td><code>[pos=&quot;V.*PP&quot;]</code>;</td>
</tr>
<tr>
<td>5.</td>
<td><code>wird</code></td>
</tr>
<tr>
<td>6a.</td>
<td><code>&lt;chunk&gt;</code></td>
</tr>
<tr>
<td>6b.</td>
<td><code>[chunk_type=&quot;NC&quot;]+</code></td>
</tr>
<tr>
<td>6c.</td>
<td><code>&lt;/chunk&gt;</code></td>
</tr>
<tr>
<td>7.</td>
<td><code>[word != &quot;:&quot;]#</code></td>
</tr>
<tr>
<td>8.</td>
<td><code>[pos=&quot;V.*PP&quot;]</code></td>
</tr>
<tr>
<td>9a.</td>
<td><code>[pos=&quot;VAFIN&quot;&amp; lemma=&quot;werden&quot;]</code></td>
</tr>
<tr>
<td>9b.</td>
<td><code>wird</code></td>
</tr>
<tr>
<td>10.</td>
<td><code>/s&gt;</code></td>
</tr>
</tbody>
</table>

Table 3: Example queries to extract processual finite passive constructions

CQP also allows us to sort the extracted information according to the metadata: text registers and IDs or translation methods and tools. Table 4 shows an example of frequency distribution according to the metadata information. In this way, we can obtain data for our analyses of translation variation.

5 Preliminary Analyses

5.1 Profile of VARTRA-SMALL in terms of shallow features

We start our analyses with the comparison of translation variants only saved in our subcorpora: PT, CAT, RBMT, SMT1 and SMT2. The structure of the corpus, as well as the annotations available already allow us to compare subcorpora (translation variants) in terms of shallow features, such as type-token-ration (TTR), lexical density (LD) and part-of-speech (POS) distributions. These features are among the most frequently used variables which characterise linguistic variation in corpora, cf. (Biber et al., 1999) among others. They also deliver the best scores in the identification of translationese features. We calculate TTR as the percentage of different lexical word forms (types) per subcorpus. LD is calculated as percentage of content words and the percentages given in the POS distribution are the percentages of given word classes per subcorpus, all normalised per cent. The numerical results for TTR and LD are given in Table 5.

<table>
<thead>
<tr>
<th>method</th>
<th>tool</th>
<th>register</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>Across</td>
<td>POPSCI</td>
<td>101</td>
</tr>
<tr>
<td>CAT</td>
<td>Across</td>
<td>SHARE</td>
<td>90</td>
</tr>
<tr>
<td>CAT</td>
<td>Across</td>
<td>SPEECH</td>
<td>89</td>
</tr>
<tr>
<td>CAT</td>
<td>Across</td>
<td>INSTR</td>
<td>73</td>
</tr>
<tr>
<td>RBMT</td>
<td>SYSTRAN</td>
<td>SHARE</td>
<td>63</td>
</tr>
<tr>
<td>RBMT</td>
<td>SYSTRAN</td>
<td>POPSCI</td>
<td>62</td>
</tr>
<tr>
<td>CAT</td>
<td>Across</td>
<td>TOU</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 4: Example output of V2 processual passive across translation method, tool and text register (absolute frequencies)

<table>
<thead>
<tr>
<th>subc</th>
<th>TTR</th>
<th>LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>15.82</td>
<td>48.33</td>
</tr>
<tr>
<td>CAT</td>
<td>14.10</td>
<td>44.60</td>
</tr>
<tr>
<td>RBMT</td>
<td>15.04</td>
<td>45.08</td>
</tr>
<tr>
<td>SMT1</td>
<td>14.32</td>
<td>46.03</td>
</tr>
<tr>
<td>SMT2</td>
<td>14.68</td>
<td>47.86</td>
</tr>
</tbody>
</table>

Table 5: TTR and LD in VARTRA-SMALL
Figure 1: Example of an annotated sample from VARTRA-SMALL

For the analysis of POS distribution, we decide to restrict them to nominal and verbal word classes. Tables 6 and 7 illustrate distribution of nominal – nouns, pronouns (pron), adjectives (adj) and adpositions (adp), and verbal word classes – verbs, adverbs (adv) and conjunctions (conj) – across different translation variants.

Table 6: Nominal word classes in % in VARTRA-SMALL

<table>
<thead>
<tr>
<th>subc</th>
<th>noun</th>
<th>pron</th>
<th>adj</th>
<th>adp</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>27.18</td>
<td>8.23</td>
<td>9.38</td>
<td>8.31</td>
<td>53.10</td>
</tr>
<tr>
<td>CAT</td>
<td>24.80</td>
<td>8.53</td>
<td>8.08</td>
<td>9.52</td>
<td>50.93</td>
</tr>
<tr>
<td>RBMT</td>
<td>24.80</td>
<td>8.61</td>
<td>8.91</td>
<td>9.01</td>
<td>51.32</td>
</tr>
<tr>
<td>SMT1</td>
<td>27.18</td>
<td>8.04</td>
<td>8.67</td>
<td>9.02</td>
<td>52.89</td>
</tr>
<tr>
<td>SMT2</td>
<td>29.78</td>
<td>7.28</td>
<td>10.42</td>
<td>8.64</td>
<td>56.11</td>
</tr>
</tbody>
</table>

Table 7: Verbal word classes in % in VARTRA-SMALL

<table>
<thead>
<tr>
<th>subc</th>
<th>verb</th>
<th>adv</th>
<th>conj</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>11.80</td>
<td>3.95</td>
<td>5.32</td>
<td>21.06</td>
</tr>
<tr>
<td>CAT</td>
<td>13.58</td>
<td>3.69</td>
<td>5.83</td>
<td>23.10</td>
</tr>
<tr>
<td>RBMT</td>
<td>12.90</td>
<td>2.74</td>
<td>6.34</td>
<td>21.99</td>
</tr>
<tr>
<td>SMT1</td>
<td>11.88</td>
<td>2.81</td>
<td>6.32</td>
<td>21.02</td>
</tr>
<tr>
<td>SMT2</td>
<td>9.09</td>
<td>2.52</td>
<td>6.06</td>
<td>17.67</td>
</tr>
</tbody>
</table>

5.2 Interpretation of results

According to Biber (1999), high proportion of variable lexical words in a text is an indicator of richness and density of experiential meanings. This characterises the field of discourse (see section 2.2 above), and TTR, thus, indicates informational density. In terms of translationese (see section 2.1), TTR reveals simplification features of translations. Translations always reveal lower TTR and LD than their originals, cf. (Hansen, 2003).

The highest TTR, thus, the most lexically rich translation variant in VARTRA is the one produced by human translators: PT > RBMT > SMT2 > SMT1 > CAT. It is interesting that the other human-produced variant demonstrates the lowest lexical richness which might be explained by the level of experience of translators (student...
translators). Another reason could be the strength of pronominal cohesion and less explicit specification of domains. However, the comparison of the distribution of pronouns (devices for pronominal cohesion) does not reveal big differences between PT and CAT, cf. Table 6.

Another simplification feature is LD, which is also the lowest in CAT-subcorpus of VARTRA: PT > SMT2 > SMT1 > RBMT > CAT. Steiner (2012) claims that lower lexical density can indicate increased logical explicitness (increased use of conjunctions and adpositions) in translations. CAT does demonstrate the highest number of adpositions in the corpus, although the difference across subcorpora is not high, see Table 6.

The overall variation between the subcorpora in terms of TTR and LD is not high, which can be interpreted as indicator of levelling out (see section 2.1 above): translations are often more alike in terms of these features than the individual texts in a comparable corpus of source or target language.

In terms of nominal vs. verbal word classes, there seems to be a degree of dominance of nominal classes (56.11% vs. 17.67%) in SMT2 resulting in a ratio of 3.18 compared to other subcorpora, cf. Table 8.

<table>
<thead>
<tr>
<th>subc</th>
<th>nominal vs. verbal</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>53.10 : 21.06</td>
<td>2.52</td>
</tr>
<tr>
<td>CAT</td>
<td>50.93 : 23.10</td>
<td>2.20</td>
</tr>
<tr>
<td>RBMT</td>
<td>51.32 : 21.99</td>
<td>2.33</td>
</tr>
<tr>
<td>SMT1</td>
<td>52.89 : 21.02</td>
<td>2.52</td>
</tr>
<tr>
<td>SMT2</td>
<td>56.11 : 17.67</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Table 8: Proportionality of nominal vs. verbal opposition in VARTRA-SMALL

The greatest contributors to this dominance are nouns and adjectives (Table 6 above). For CAT, we again observe the lowest numbers (the lowest noun vs. verb ratio) which means that this translation variant seems to be the most “verbal” one. According to Steiner (2012), German translations are usually more verbal than German originals. Comparing German and English in general, the author claims that German is less “verbal” than English. Thus, a higher “verbality” serves as an indicator of “shining though” (see 2.1 above), which we observe in case of CAT. However, to find this out, we would need to compare our subcorpora with their originals, as well as the comparable German originals.

5.3 First statistical experiments

We use the extracted shallow features for the first steps in feature evaluation. As our aim is to investigate the relations between the observed feature frequencies and the respective translation variants, we decide for correspondence analysis, a multivariate technique, which works on observed frequencies and provides a map of the data usually plotted in a two-dimensional graph, cf. (Baayen, 2008).

As input we use the features described in 5.1 above: TTR, LD, nouns, adjectives (adj), adpositions (adp), verbs, adverbs (adv), conjunctions (conj). Additionally, we divide the class of pronouns into two groups: personal (pers.P) and demonstrative (dem.P) – devices to express pronominal cohesion. We also extract frequency information on modal verbs which express modality.

The output of the correspondence analysis is plotted into a two-dimensional graph with arrows representing the observed feature frequencies and points representing the translation variants. The length of the arrows indicates how pronounced a particular feature is. The position of the points in relation to the arrows indicates the relative importance of a feature for a translation variant. The arrows pointing in the direction of an axis indicate a high contribution to the respective dimension. Figure 2 shows the graph for our data.

In Table 9, we present the Eigenvalues calculated for each dimension to assess how well our data is represented in the graph. We are able to obtain a relatively high cumulative value by the first two dimensions (representing x and y-axis in Figure 2), as they are the ones used to plot the two-dimensional graph. The cumulative value for the first two dimensions is 94.3%, which indicates that our data is well represented in the graph.

If we consider the y-axis in Figure 2, we see that there is a separation between human and machine translation, although SMT2 is on the borderline. CAT is also closer to MT, as it is plotted much closer to 0 than PT. Conjunctions, personal pronouns and adverbs seem to be most prominent contributors to this separation, as their arrows are...
Verbs, adjectives and nouns seem to be most prominent contributors to the other division (considering the x-axis). Here, we can observe three groups of subcorpora: CAT and RBMT share certain properties which differ them from SMT2. PT remains on the borderline, whereas SMT1 tend slightly to SMT2.

6 Conclusion and Future Work

In this paper, we presented a comparable corpus of translations from English into German, which contains multiple variants of translation of the same texts. This corpus is an important resource for the investigation of variation phenomena reflected in linguistic features of translations. The corpus architecture allows us to extract these features automatically. Our preliminary results show that there are both similarities and differences between translation variants produced by humans and machine systems. We expect even more variation, if we compare the distribution of these features across text registers available in all subcorpora.

However, there is a need to inspect the reasons for this variation, as they can be effected by translator experience, restrictions of the CAT system applied or the training material used in MT.

We believe that our resources, as well as our research results will find application not only in contrastive linguistics or translation studies. On the one hand, our corpus provides a useful dataset to investigate translation phenomena and processes,
but on the other, it can be used for the development, optimisation and evaluation of MT systems, as well as CAT tools (e.g. translation memories).

In the future, we aim at expanding it with more data: (1) more texts for the existing registers (each register should contain around 30,000 words), (2) further text registers (e.g. academic, web and news texts). We also plan to produce further human and machine-generated translations, i.e. (3) machine translations post-edited by humans, as well as translation outputs of (4) further MT systems. Moreover, we aim at adding translations from German into English to trace variation influenced by language typology.

As the automatic tagging of part-of-speech and chunk information might be erroneous, we plan to evaluate the output of the TreeTagger and compare it with the output of further tools available, e.g. MATE dependency parser, cf. (Bohnet, 2010). Furthermore, the originals will be aligned with their translations on word and sentence level. This annotation type is particularly important for the analysis of variation in translation of certain lexico-grammatical structures.

A part of the corpus (CAT, RBMT and SMT subcorpora) will be available to a wider academic public, e.g. via the CLARIN-D repository.

Acknowledgments

The project “VARTRA: Translation Variation” was supported by a grant from Forschungsausschuß of the Saarland University. We are especially grateful to Anne-Katrin Schumann, Elke Teich and Noam Ordan for their comments. Also, we wish to thank the anonymous reviewers for their suggestions for improving our paper. All remaining errors remain ours.

References


Abstract

The present paper describes an attempt to create an interoperable scheme using existing annotations of textual phenomena across languages and genres including non-canonical ones. Such a kind of analysis requires annotated multilingual resources which are costly. Therefore, we make use of annotations already available in the resources for English, German and Czech. As the annotations in these corpora are based on different conceptual and methodological backgrounds, we need an interoperable scheme that covers existing categories and at the same time allows a comparison of the resources. In this paper, we describe how this interoperable scheme was created and which problematic cases we had to consider. The resulting scheme is supposed to be applied in the future to explore contrasts between the three languages under analysis, for which we expect the greatest differences in the degree of variation between non-canonical and canonical language.

1 Aims and Motivation

The aim of the present study is to create a scheme which will allow us to use existing annotations of textual phenomena, and which will be applicable to multiple languages and genres, including non-canonical ones. The annotations were created within two separate projects: German-English Contrasts in Cohesion (GECCo, Lapshinova and Kunz (2014)) whose focus was on English and German on the one hand, and the Prague Dependency Treebank (PDT 3.0, Bejček et al. (2013)) with the analysis of Czech, on the other hand.

The resulting scheme will serve our overarching goal to unify the two approaches in a joint analysis of contrasts between English, German and Czech on the level of discourse. We assume that the greatest differences between these languages lie in the degree of variation between non-canonical and canonical language (here we especially mean spoken language). Previous findings on lexico-grammatical and also cohesive phenomena have evidenced that there is more variation between written and spoken dimensions in German than in English, even though they are closely related, cf. Mair (2006) or Kunz et al. (forthcoming). Studies with respect to spoken and written Czech (see, e.g., Cvrček et al. (2010)) suggest that the differences between written and spoken language are even more pronounced in Czech than in German, at least with respect to lexico-grammar, hence we expect that this also holds for the level of text/discourse.

We therefore suggest that if we draw a line of differences between spoken and written English, German and Czech, we would observe a continuum in the degree of variation between these languages, as seen in Figure 1. The graph also reflects the above assumption that the differences are less pronounced between English and German than if we compare English and German with Czech. The reasons for this lie in the linguistic heritage of these languages (English and German have a common West-Germanic origin while Czech is a Slavic language) and in sociolinguistic factors that influenced their evolution (for example, Czech purism at the beginning of the 20th century, described, e.g., in Havránek and Weingart (1932)). To our knowledge, there is no
research testing these assumptions. We believe that a cross-language analysis based on the interoperable scheme proposed in this work will help to fill this gap.

However, this kind of study requires corpora that are annotated for textual phenomena. As the creation of such corpora is a time-consuming task, we decide to take advantage of existing resources, i.e. corpora, which already contain annotations of these phenomena. However, while capturing the same phenomena, the annotations in the corpora at hand were created in the frame of two different projects (GECCo and PDT, see Section 2). Moreover, both existing annotation schemes only account for the systemic peculiarities and realizational options of the languages analysed and hence are not general enough to permit a comparison across Germanic and Slavic languages. For this reason, we need to unify the categories in these schemes to create an interoperable one which can be applicable to multiple languages and text registers, including spoken ones. The scheme will allow us to profit from the existing annotated resources and at the same time will enable the contrastive analysis of the languages involved. We believe that the resulting scheme will find application not only in our research, but also in further linguistic studies and in cross-language NLP applications. It is beyond the scope of this paper to include the contrastive language analysis, which will follow from the unified scheme in our future work.

2 Theoretical Background

In this section, we describe the frameworks for the analysis of English, German and Czech. They were used in the development of the resources at hand (which are described later in Section 3) and will serve as a basis for our interoperable scheme.

2.1 Frameworks for the analysis of English and German

The analysis of textual phenomena in GECCo is based on the definition of cohesion. The concept was established by Halliday and Hasan (1976) for English, in the frame of Systemic Functional Linguistics. It concerns textual relations between linguistic expressions across grammatical domains. Additionally, the categories under analysis are based on the conceptualisations of de Beaugrande and Dressler (1981), who consider cohesion as an explicit linguistic signal on the text surface to establish coherence or textuality. Cohesion always involves a linguistic trigger (cohesive device) that links up to other linguistic expressions in the same text. The main categories used in the analysis include coreference to create relations of identity, comparative reference, substitution and ellipsis to create relations of comparison between referents belonging to the same type, conjunction for logico-semantic relations between propositions, and lexical cohesion for similarity between different types of referents. The adaptation of these categories and their subcategories to the bilingual comparison of English and German have been described in Kunz et al. (forthcoming). For coreference, ellipsis and lexical cohesion, not only cohesive devices were considered, but also the linguistic expressions they tie up with as well as the cohesive relations. The relations may contain more than just two linguistic expressions and form cohesive chains that stretch over longer passages of text.

2.2 Framework for the analysis of Czech

In the framework for the analysis of Czech, the following textual phenomena are included: ellipses, information structure, grammatical and textual coreference, bridging relations (associative anaphora) and discourse relations. Their definition is based on Functional Generative Description as described in Sgall et al. (1986). The approach uses syntactic as well as semantic criteria for text analysis and considers three layers of text representation: morphological, analytical and tectogrammatical (deep syntactic). At the tectogrammatical layer, the meaning of the sentence is represented as a dependency tree structure, in which nodes represent autosemantic words and are labelled with a large set of at-
tributes. This layer of representation is especially important for elliptical constructions, as they are captured here in reconstructions (Mikulová, 2014). Besides that, the tectogrammatical layer also covers information on structural attributes (in terms of contextually bound or contextually non-bound nodes). The approach to textual phenomena exceeding the sentence boundary is two-fold for the Czech framework. On the one hand, the conception of discourse relations is based on the Penn-style discourse lexically-grounded approach, as described in Prasad et al. (2008). According to this approach, only those relations that are signaled by explicit markers (connectives) are considered as discourse relations. However, in contrast to the Penn-style approach, the set of connectives is an open list, see Poláková et al. (2013), and the treatment of coreference and bridging relations includes both explicit and implicit categories. Language expressions that refer to the same discourse entity are considered to be coreferent. As for bridging relations, their definition has been taken from Clark (1975).

### 3 Data and Experiment

As already mentioned in Section 1, we aim to take advantage of the existing corpora annotated for textual phenomena to avoid the time-consuming creation of such resources. The existing German and English data are annotated with the GECCo framework described in 2.1, whereas the data for Czech are annotated in the PDT style described in section 2.2 above. The current section provides a brief description of these resources at hand.

#### 3.1 GECCo - German and English corpora

The GECCo corpus annotated for textual phenomena with the framework described in 2.1 represents a continuum of different text types (registers in the sense of Systemic Functional linguistics) from written to spoken discourse. More precisely, it includes English and German texts of ten registers, eight of which represent written discourse and include fictional texts, political essays, instruction manuals, popular-scientific texts, letters to shareholders, prepared political speeches, tourism leaflets and texts from corporate websites. This part contains not only original texts, but also their translations in both directions. The registers of spoken discourse include recorded and transcribed interviews and academic speeches described in Lapshinova-Koltunski et al. (2012), as well as transcriptions of television talkshows, texts from internet forums, medical consultations and sermon texts. The total number of words contained in the corpus comprises ca. 1.6 Mio (including translations). The corpus is annotated on several levels, which include morphological, syntactical, structural and textual information (i.e. information on cohesion as described above). The information on the latter was annotated with the help of semi-automatic procedures described by Lapshinova-Koltunski and Kunz (2014). These result from an integration of the systemic peculiarities of English and German and at the same time account for textual variation in terms of canonical written and non-canonical spoken language. The rich annotation allows capturing information about the structural and syntactic features of cohesive devices (and also antecedents) and about how they are mapped onto information structure. Moreover, it yields information about chain features, e.g. number of elements in chains, distance between chain elements and number of different chains.

#### 3.2 Prague Dependency Treebanks

There is a number of corpora annotated according to the Prague annotation scenario described in section 2.2 above. These include PDT 3.0 – Prague Dependency Treebank (Bejček et al., 2013), PCEDT 2.0 – Prague English Dependency Treebank (Hajič et al., 2012) and PDTSL – Prague Dependency Treebank of Spoken Language (Hajič et al., 2009). All these corpora consist of original texts (Czech and English respectively) extracted from newspaper articles (PDT), Wall Street Journal (PCEDT) and transcribed and reconstructed spontaneous dialogue speech in Czech and English. PCEDT 2.0 also contains translations from English into Czech. The total number of words in written corpora comprises ca. 3.2 Mio (including translations) and spoken corpora for English and Czech total ca. 770 thousand tokens. The written corpora are annotated with morphological, analytical and tectogrammatical information, whereas each sentence is represented as a dependency tree structure. The tectogrammatical layer of PDT 3.0 also contains annotation of...
information structure attributes and the following discourse phenomena: extended (nominal) textual coreference, bridging relations, discourse connectives and the discourse units linked by them, and semantic relations between these units, see Poláková et al. (2013) for details.

3.3 Experiment settings

The creation of an interoperable scheme requires a comparison of the underlying annotations. We therefore annotate the same data set on the basis of both conceptions, and identify those categories that cover the same phenomena. For this, we have selected texts in English (both originals) belonging to two different genres – journalism and fiction and annotated them in accordance with the guidelines of the Prague and GECCo conceptions. Journalistic texts represent written discourse, whereas the fictional texts we selected are closer to spoken language and other non-canonical genres, e.g., internet blogs or tweets. They are partially narrative and partially dialogic, and hence contain turns, but also reformulations, elaboration and other spoken language features. We believe that this data constellation ensures a good base for our future analysis (aimed at comparison of spoken vs. written dimensions). We decide for texts in English, as English data is available in both underlying resources, hence allowing us to unify the annotated categories afterwards. The journalistic sample contains texts exported from PCEDT 2.0 (see section 3.2), with a size of around 100 sentences. A sample of fictional texts of the same size was exported from the GECCo corpus described in 3.1. For the sake of convenience, we used different annotation tools for the two different frameworks – TrEd (Pajas and Štěpánek, 2008) for the framework described in 2.2, as it allows visualisation of trees, and MMAX2 (Müller and Strube, 2006) for the framework described in 2.1, as this enables visualisation of longer cohesive chains. The annotations were carried out manually by four trained annotators. Then, the parallelly created annotations were compared and analysed qualitatively and quantitatively. The results of this analysis are presented in section 4 below.

4 Analyses

4.1 Overall comparison

Both GECCo and PDT frameworks include annotations of ellipses, coreference relations and discourse connectives. The category of lexical cohesion in the German-English framework (see section 2.1) can be partially mapped to bridging relations in the Czech framework (see 2.2), although lexical cohesion is much more lexically grounded than bridging. Substitution is the only phenomenon which is asymmetric in the frameworks. It is not covered by the definition of textual relations in the framework for Czech, as this device is common for English and (less so) for German but not relevant at all for Czech. We provide a mapping of the phenomena available in both frameworks in Table 1.

<table>
<thead>
<tr>
<th>GECCo</th>
<th>PDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>coreference</td>
<td>coreference</td>
</tr>
<tr>
<td>lexical cohesion</td>
<td>bridging</td>
</tr>
<tr>
<td>ellipsis</td>
<td>ellipsis (in dependency trees)</td>
</tr>
<tr>
<td>connectives, relations</td>
<td>connectives, arguments, relations</td>
</tr>
<tr>
<td>substitution</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Mapping of the phenomena

We count the occurrences of these categories in the experimental dataset and compare absolute numbers for both frameworks, see Figure 2. The numbers in Figure 2 reveal the preferences for certain types of relations in the two approaches involved. At the same time, we are able to observe the similarities between the types.

What is most evident from the figure is that the
number of discourse relations expressed by connectives\textsuperscript{1} annotated in both approaches is very similar. This is mainly due to the fact that the typology of discourse relations of the main categories is similar in both approaches. Neglecting the terminology, there are four main relations in both approaches: temporal, causal, adversative and additive. In GECCo, also modal DSDs are distinguished (such as well, sure, of course, surely, etc.). They are especially frequent in spoken genres. However, they only provide a rather vague link to the two arguments, as they primarily carry an emotional meaning. For this particular reason, this type of textual devices is not included in the PDT framework, where a DSD always requires a clear linkage of two arguments, and in which the scope of discourse arguments is taken into account. If modal DSDs were subtracted, the number of connectives for the German-English framework would slightly change. However, it does not change the comparison considerably. The other difference observed in the approach to discourse relations is that, in the Penn-style, the four main categories are further differentiated into more detailed relations, whereas in the German-English framework, only the general categories are considered.

The numbers for the other textual phenomena reveal more differences. For example, the frequencies of ellipses and coreference relations annotated within the PDT framework prevail over those of the other types. This is justified by the representation of the phenomena according to the framework: Apart from textual ellipses (Did she open the door? No, she did not [open the door]), it also contains various grammatical types of elliptical constructions, e.g. structural ellipses (ellipses of governing verbs and nouns), different kinds of anaphoric zeros (Their reaction was 0 to do nothing and 0 ride it out), including arguments with control constructions (Peter want to [Peter] sleep), general arguments (Jane sells at Bata [what] [to whom]), etc. These are reconstructed on the deep syntactic level. The GECCo approach is based on signals to textual cohesion, and therefore, ellipses are annotated only in the case of textual relations across grammatical domains. Besides, anaphoric zeros are not reconstructed in syntactic structures.

For our contrastive analysis, we will consider cases of textual ellipsis only, which are expected to contribute especially to the differences between spoken and written language. We expect textual ellipsis to be more common in spoken genres, as our previous analyses for English and German have already evidenced, cf. Kunz et al. (forthcoming). Example (1) demonstrates a case of textual ellipsis considered in both approaches.

(1) He’d never even bothered to read it. But Truman had [].

The difference here lies in the representation of the missing element. In the GECCo approach, this case is annotated as verbal ellipsis. The missing parts of the verbal phrase could either be bothered to read it or read it. In the PDT approach, the whole verbal phrase is reconstructed in the dependency tree, see Figure 3, connected to the antecedents of verbs by the arrows of grammatical and textual coreference. Note that this type of ellipsis, where only the operator is kept (termed as lexical ellipsis by Halliday & Hasan (1976)), is available in English, but neither in German nor Czech.

Figure 3: Ellipsis in the dependency tree representation (PDT-style)

The differences in the annotations of coreference are due to the diverging definitions of coreferring expressions. In GECCo, only the mentions with an explicit marker, the cohesive device (e.g. definite articles, pronouns, demonstratives, etc.), are taken into account. This implies, for instance, that relations between named entities or between nominal construc-
tions in plural which are not introduced by a determiner are excluded from the annotation of coreference. They are, however, annotated as devices of lexical cohesion (see below). Moreover, as a cohesive relation to the antecedent is indicated by a cohesive device, only this explicit marker is annotated but not the other elements of the anaphoric nominal phrase. Hence, if an anaphoric expression consists of a definite article and a nominal head, the former is annotated as corerential device and the noun as lexical cohesion (see the and manuscript in example (2)). In the PDT approach, both implicit and explicit relations of coreference are annotated, including indefinite NPs. In addition, the whole anaphoric expression is annotated as one coreferential element, as illustrated in example (2).

(2) Twenty years I have been working on [this book],” and he leaned over to rap [[the] [manuscript]] with a thick proprietary finger,” and you can sit home in Peterskill and read it when it’s published.

Lexical repetitions (which belong to the level of lexical cohesion in GECCo) are also annotated as coreferent if they refer to the same discourse entity.

We assume that the differences in the annotation of coreference are also related to the contrasts that we observe for bridging/lexical cohesion, see Figure 2. Although there is a partial intersection of sets of the relations, the different conceptions are clearly seen in the annotations: in lexical cohesion, lexico-semantic properties of mentions in text are important. The semantic relations (e.g., meronymy, hyponymy, synonymy, etc.) assigned to the mentions are based on the context-free sense relations into which lexical words or patterns can enter, whereas their contextual meaning and referential properties are neglected. By contrast, bridging relations are based on the information instantiated in the text, which means that only those conceptual relations are considered which hold between entities mentioned in the same discourse. Nevertheless, we noticed that relations not marked as lexical cohesion are compensated by the annotation of coreference relations in GECCo, and taken together, they are comparable to the relations of bridging and coreference in the PDT framework. For example, repetitions, which are a subcategory of lexical cohesion, are marked as coreference relations in the PDT framework (see above).

Summing up, there are numerous similarities and overlaps in the categories of textual phenomena in both approaches, despite of the differences discussed earlier. This leads us to conclude that textual phenomena are reflected in both approaches in a very similar way although they are annotated with diverging terminology that stems from different theoretical backgrounds. The following section (4.2) illustrates in more detail some of the cases which are especially interesting for a cross-lingual analysis of spoken and written language.

4.2 Case studies

Coreference and bridging / lexical cohesion

The interplay between coreference and bridging or lexical cohesion is especially interesting if we compare spoken and written genres, as we expect certain preferences due to contextual settings (short-time memory, presence of all speech participants in the communication situation, etc.). In Table 2, we demonstrate the statistics (numbers are counted for one journalistic text consisting of 43 sentences) for coreference chains identified with both annotation schemes.

<table>
<thead>
<tr>
<th></th>
<th>GECCo-style</th>
<th>PDT-style</th>
</tr>
</thead>
<tbody>
<tr>
<td>coref.chains</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>aver.chain length1</td>
<td>3.48</td>
<td>4.20</td>
</tr>
<tr>
<td>aver.chain length2</td>
<td>6.25</td>
<td>7.05</td>
</tr>
</tbody>
</table>

Table 2: Annotation statistics for coreference chains

We compare the total number of chains and the average chain length\(^2\) which are higher in the PDT framework than in the GECCo approach for German and English. This coincides with the results that we observed in Section 2 above, as the total number of coreference elements is much lower in the GECCo framework.

If we go into detail and analyse the subtypes of anaphora, we find some fine-grained differences in the annotation. For example, event anaphora are annotated in both frameworks. However, the largest

\(^2\)aver.chain length1 is used for all chains, whereas aver.chain length2 indicates statistics for chains containing more than two elements.
scope of the antecedent of this anaphora type is limited to the extension of a sentence in the tree-based approach while cohesion-based annotations also include larger textual antecedents.

The above mentioned (see Section 4.1) overlap between coreference and bridging can be illustrated by the example in (3). The relation in (3-a) is covered by a combination of comparative reference and lexical cohesion in the GECCo framework, and by contrastive bridging in the PDT framework. At the same time, comparative reference also includes such cases as (3-b) and (3-c), combined with lexical cohesion in (3-b) and coreference and lexical cohesion in (3-c). Both are cases of bridging anaphora and common textual coreference in the PDT framework.

(3) a. a presentation – a better presentation, an example – other examples
   b. some case – such/similar cases.
   c. one hand – the same hand

Another illustration of this overlap can be seen in (4), where she, her children, her war-damaged husband and their are marked as a bridging relation (type subset - set) in one approach, whereas she, her, her and their are annotated as coreference in the other, their with a split antecedent.

(4) Although [she] was kind and playful to [her] children, she was dreadful to [her war-damaged husband]; she openly brought her lover into [their] home.

The relation between The World War II and that in (5) shows how coreference signaled by a demonstrative pronoun in the GECCo approach may coincide with the bridging relation in the PDT approach. In the latter, an explicit anaphor is marked as signalling a bridging and not a coreference relation since it is not entirely clear whether the event (The World War II in (5)) is identical with that time.

(5) [The World War II] remained one of the most tragic events in the history. But at [[that] time] nobody thought about it.

A minor difference between the approaches can be found within the field of event anaphora annotation. In the PDT approach, an antecedent can be explicitly annotated only when it is not longer than one sentence. In the GECCo approach, the scope of the antecedent is annotated independently of the size of the antecedent.

**Discourse relations** As already mentioned above, the greatest similarities between the two approaches were observed in terms of the total number of identified discourse relations in both schemes. The differences are discovered here on the level of types of relations involved. For example, the connective and in (6) is assigned a reason-result relation in the PDT framework, while the GECCo framework considers it as an additive conjunction.

(6) William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs, [and] Gates became an industry billionaire six years after IBM adapted one of these versions in 1981.

In Table 3, we demonstrate the number of relations identified per approach and per text genre, as we suppose that the detected differences can be genre-sensitive.

<table>
<thead>
<tr>
<th>Discourse Relation</th>
<th>GECCo-style</th>
<th>PDT-style</th>
</tr>
</thead>
<tbody>
<tr>
<td>temporal</td>
<td>6 11</td>
<td>5 5</td>
</tr>
<tr>
<td>contin./caus.</td>
<td>9 6</td>
<td>4</td>
</tr>
<tr>
<td>comp./adver.</td>
<td>16 10</td>
<td>15 17</td>
</tr>
<tr>
<td>expans./addit.</td>
<td>22 24</td>
<td>19 22</td>
</tr>
<tr>
<td>modal</td>
<td>7 4</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Annotation statistics for discourse connectives

For instance, both frameworks identify approximately the same number of temporal relations in the journalistic texts. Yet, deviating numbers for this relation are obtained for the fictional texts. The same tendency is observed for relations of contrast (adversative). In case of contingency or causal relations, the situation is different: the number of relations coincide here for fiction rather than journalism.

5 Resulting Scheme and Discussion

Summarising all the cases analysed in the data that were annotated with both frameworks, we create an intersection scheme, covering all overlapping categories. This scheme is illustrated in Table 4. The main categories here are labelled as
IDENTITY, NON-IDENTITY, ELLIPSIS and DISCOURSE RELATIONS. These general categories also include subclasses on a more fine-grained level, e.g. METONYMY or CONTRAST, which can be derived from the existing annotation. For the time being, we exclude the categories without correspondence, i.e. which exist in one approach but not in the other.

As can be seen from the table, the annotation schemes based on both frameworks can be merged even though there are differences in the terminology used for specific features, in the level of granularity and in the method of annotation.

However, without the categories we had to exclude because there was no correspondence between the two approaches, we cannot cover all the cases of textual phenomena. For instance, modal discourse markers, which are especially important for spoken genres cannot be captured by our interoperable scheme for the time being.

One of the main reasons for the incompatibility of the excluded categories lies in the nature of the phenomenon itself: the GECCo approach takes a linguistic signal into account, while the PDT framework includes a more abstract level of coherence. This is especially reflected in the relations of IDENTITY which are not marked by a referring item, e.g. definite article, pronoun, etc. In turn, the GECCo framework captures more semantic relations, e.g. hyponomy, synonymy, etc. that are purely based on sense relations and not on relations between instantiated referents, thus allowing a more fine-grained view on the thematic progression in a text, see Figure 4.

As already stated above, the conceptual dissimilarities discovered in this study seem to result, at least partially, from the systemic differences between Germanic and Slavic languages with respect to the language devices available for expressing textual phenomena. For instance, English uses a very closed class of explicit markers for establishing a relation of comparison, labeled as substitution (the shirt – the red one). German is more heterogeneous with respect to the linguistic items available, while Czech has no corresponding structures and makes use of ellipsis instead. We expect that these differences will be even more apparent when integrating the analysis of non-canonical spoken varieties into our trilingual study.

Our future work will include the application of the resulting scheme to our contrastive analysis of naturally occurring texts of English, German and Czech. We are particularly interested in comparing the textual phenomena realized in texts with plain written style with those occurring in non-canonical texts that are produced spontaneously, with a high degree of interaction between varying numbers of speech participants, such as talkshows or private conversation. Moreover, we intend to investigate language production in between spoken and written, such as forums, blogs or interviews. We expect that the most significant differences between languages and genres are tied to varying contextual configurations of mode, e.g. number of speech participants, private vs. public conversation, time laps between production and reception). They may be reflected in textual phenomena with respect to their overall number, the degree of explicitness, as well as the type of textual categories that are preferred. Moreover, we intend to examine variation in the degree of dependence of these textual phenomena on lexicogrammatical constraints or pragmatic peculiarities. The scheme developed in this paper is a first step towards unifying different frameworks that result from separate analyses of Germanic languages and a Slavic language. It therefore reflects a level of generalisation that is applicable to trilingual analysis, which will, however, be broken into more delicate subcategories to permit an identification of fine-grained contrasts.

6 Acknowledgement

This work was made possible by a grant on Short Term Scientific Missions received within the Textlink Action (ISCH COST Action IS1312). We also acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658). This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society).

3http://textlinkcost.wix.com/textlink
Table 4: Categories for the language- and genre-insensitive scheme

<table>
<thead>
<tr>
<th>IDENTITY</th>
<th>Czech framework</th>
<th>German-English framework</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coreference with pronouns</td>
<td>coreference with pers. and demo. heads</td>
</tr>
<tr>
<td></td>
<td>pronouns with arrows to segments and</td>
<td>extended reference</td>
</tr>
<tr>
<td></td>
<td>events</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NP coreference</td>
<td>coreference with pers./ dem. modifiers or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>def.art.+hyperonymy/repetition/synonymy</td>
</tr>
<tr>
<td></td>
<td>coreference of NEs</td>
<td>repetitions of named entities</td>
</tr>
<tr>
<td></td>
<td>coreference with the word same</td>
<td>comp.reference with the word same</td>
</tr>
<tr>
<td></td>
<td>coreference with demonstrative local</td>
<td>coreference with demonstrative local and</td>
</tr>
<tr>
<td></td>
<td>and temporal adverbs (tam, tehdy)</td>
<td>temporal adverbs</td>
</tr>
<tr>
<td>NON-IDENTITY</td>
<td>contextual relations of MERONYMY between</td>
<td>contextual relations of MERONYMY between lexical</td>
</tr>
<tr>
<td></td>
<td>lexical items</td>
<td>items</td>
</tr>
<tr>
<td></td>
<td>bridging CONTRAST with comparative</td>
<td>comparative reference excluding cases with</td>
</tr>
<tr>
<td></td>
<td>adjective</td>
<td>the word same</td>
</tr>
<tr>
<td></td>
<td>bridging CONTRAST without comparative</td>
<td>antonyms in lex.coh</td>
</tr>
<tr>
<td></td>
<td>adjective</td>
<td></td>
</tr>
<tr>
<td>DISCOURSE</td>
<td>temporal</td>
<td>temporal</td>
</tr>
<tr>
<td></td>
<td>contingency</td>
<td>causal</td>
</tr>
<tr>
<td>RELATIONS</td>
<td>comparison (contrast)</td>
<td>adversative</td>
</tr>
<tr>
<td></td>
<td>expansion</td>
<td>additive</td>
</tr>
<tr>
<td>ELLIPSIS</td>
<td>textual ellipsis (nominal, verbal, clausal)</td>
<td>cohesive ellipsis (nominal, verbal, clausal)</td>
</tr>
</tbody>
</table>

Figure 4: Coreferential and lexical relations in both approaches

References


Jan Hajček, Petr Pajas, David Mareček, Marie Mikulová, Zdeňka Urešová, and Petr Podveský. 2009. Prague dependency treebank of spoken language (PDTSL) 0.5.


Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel, and Erich


Interlingual Variation

In this section, we...
From Interoperable Annotations towards Interoperable Resources: A Multilingual Approach to the Analysis of Discourse

Ekaterina Lapshinova-Koltunski*, Kerstin Anna Kunz**, Anna Nedoluzhko***
Saarland University*, University of Heidelberg**, Charles University in Prague, Faculty of Mathematics and Physics***
e.lapshinova@mx.uni-saarland.de, kerstin.kunz@iued.uni-heidelberg.de, nedoluzhko@ufal.mff.cuni.cz

Abstract

In the present paper, we analyse variation of discourse phenomena in two typologically different languages, i.e. in German and Czech. The novelty of our approach lies in the nature of the resources we are using. Advantage is taken of existing resources, which are, however, annotated on the basis of two different frameworks. We use an interoperable scheme unifying discourse phenomena in both frameworks into more abstract categories and considering only those phenomena that have a direct match in German and Czech. The discourse properties we focus on are relations of identity, semantic similarity, ellipsis and discourse relations. Our study shows that the application of interoperable schemes allows an exploitation of discourse-related phenomena analysed in different projects and on the basis of different frameworks. As corpus compilation and annotation is a time-consuming task, positive results of this experiment open up new paths for contrastive linguistics, translation studies and NLP, including machine translation.

Keywords: interoperability, linguistic annotation, multilingual resources, discourse, coreference, German-Czech contrasts

1. Introduction

This paper aims at a cross-lingual analysis of discourse phenomena in the two typologically different languages – German and Czech. The discourse properties in focus are relations of identity and non-identity (semantic similarity) of discourse entities, ellipsis and discourse relations (types of conjunctions). Information on differences between the two languages in terms of discourse structuring devices is beneficial to contrastive linguistics, translation studies and multilingual natural language processing.

The novelty of our analysis lies in the nature of the resources used. Quantitative contrastive analyses on the level of discourse require annotated corpora involving time-consuming compilation and annotation, especially in a multilingual setting. Therefore, we take advantage of the existing resources reflecting systemic peculiarities and realisational options of the languages under analysis. We use Czech and German data annotated on the basis of two different frameworks: Functional Generative Description, see (Sgall et al., 1986), for Czech, and textual cohesion, see (Halliday and Hasan, 1976), for German. In our previous work, see (Lapshinova et al., 2015), we have shown that annotations of the involved resources are comparable if abstract categories are used and only the phenomena with a direct match in the frameworks for German and Czech are taken into consideration. We have also shown that although being not general enough to permit a comparison across Germanic and Slavic languages, the existing annotated resources capture the same phenomena, and the creation of an interoperable scheme is possible if more abstract categories are taken into consideration. Hence, we make use of this scheme to perform a comparison between German and Czech.

Our analysis is a first step towards unifying separate analyses of discourse relations in Germanic and Slavic languages. At the same time, it demonstrates that the application of ‘theoretically’ different resources is possible in one contrastive analysis. This is especially valuable for NLP, which uses annotated resources to train language models for various tools. Training of language models with more complex linguistic annotation often requires manually annotated corpora, which is time consuming and costly, especially if more than one language is involved. Therefore, development of interoperable schemes that enable usage of the existing annotated resources is very important.

2. Related Work

Generally speaking, Slavic languages have a richer, more fusional morphology than Germanic languages. Even though German has conserved more of the inflectional morphology of Proto-Indo-European than other Germanic languages such as English, it has a more isolating character than Czech. The morphological reduction in German partially results in a less flexible constituent word order as compared to Czech, although more positional options are possible than, e.g., in English. We expect these contrasts to have an effect on the creation of discourse properties (see interpretations below).

There is a vast number of theoretical studies comparing Germanic and Slavic languages on a rather general level, such as Štúcha (2003) and Engel (1999). Apart from these general comparative studies, a special focus on anaphoric relations between Czech and German was done by Komárek (1994). Yet, quantitative comparisons of Germanic and Slavic languages are very rare. The only works, known to us, include the comparison of English and Czech by Novák and Nedoluzhko (2015) and the comparison of English, Czech and Russian by Nedoluzhko et al. (2015). There are almost no corpus-based approaches to the comparison of the language pair German and Czech, especially if the properties of discourse are concerned. A number of corpus-based analyses exists for different Germanic languages, e.g. the one for particular cohesive conjunctions or adverbs in prepared speeches by Bühirg and House (2004) or that for abstract anaphors in parliament debates by Zinsmeister et al. (2012). Other corpus-based studies compare Romance languages, e.g. (Taboada and Gómez-González,
for particular coherence relations. In addition, there are various studies related to human and machine translation, see for instance, (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012), or (Webber et al., 2013) and and (Webber et al., 2015). Although the analysis of human and machine translation is beyond the scope of this work, we do not exclude the application of our findings for these research areas.

3. Methodology

3.1. Data

For our analysis, several texts of written discourse (essays) with comparable topics on economic, political and social issues were selected. For the German data, nine texts were excerpted from the corpus GECCo, comprising 14930 tokens and 736 sentences in total, see Table 1. The whole corpus represents a continuum of different text types including written discourse, described in (Hansen-Schirra et al., 2012) and spoken discourse described in (Lapshinova-Koltunski et al., 2012). The corpus is annotated on several levels, which include morphological, syntactical, structural and textual information. The information on the latter was annotated with the help of semi-automatic procedures described by Lapshinova-Koltunski and Kunz (2014). These result from an integration of the systemic peculiarities of English and German and at the same time account for textual variation in terms of canonical written and non-canonical spoken language. Textual information is represented in form of cohesive devices, such as coreference, conjunction, substitution, ellipsis and lexical cohesion. The annotated structures contain information about morphosyntactic features of devices (including antecedents) and allow yielding information on the chain features, i.e. number of elements in chains, distance between chain elements, etc. Annotation of textual coreference contains not only relations of identity between entities but also abstract and semantic relations (i.e. morphological, syntactical, POS, textual phenomena, etc.) along the corresponding frameworks.

The Czech texts were taken from the Prague Dependency Treebank (PDT 3.0, (Bejček et al., 2013)). They are annotated with morphological, analytical and tectogrammatical information, whereas each sentence is represented as a dependency tree structure. The tectogrammatical layer of PDT 3.0 also contains annotation of information structure attributes and the following inter-sentential relations: pronominal, zero and nominal coreference, abstract anaphora, bridging relations and discourse relations (including connectives, discourse units linked by them, and semantic relations between these units), see Zikánová et al. (2015) for details.

Since texts are shorter in PDT than in GECCo, 17 texts were excerpted to arrive at a similar number of tokens and sentences (11769 and 763 respectively), see Table 2. Both German and Czech texts under analysis include all levels of annotations (i.e. morphological, syntactical, POS, textual phenomena, etc.) along the corresponding frameworks.

<table>
<thead>
<tr>
<th>textID</th>
<th>topics</th>
<th>sent</th>
<th>tok</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO1</td>
<td>Germany and social market economy</td>
<td>121</td>
<td>2035</td>
</tr>
<tr>
<td>GO2</td>
<td>Optimistic remarks on globalisation</td>
<td>47</td>
<td>971</td>
</tr>
<tr>
<td>GO3</td>
<td>Politics and globalisation</td>
<td>103</td>
<td>1871</td>
</tr>
<tr>
<td>GO4</td>
<td>Globalisation and new challenges</td>
<td>27</td>
<td>478</td>
</tr>
<tr>
<td>GO5</td>
<td>The biggest currency changeover</td>
<td>85</td>
<td>1460</td>
</tr>
<tr>
<td>GO6</td>
<td>Globalisation and market economy</td>
<td>80</td>
<td>1782</td>
</tr>
<tr>
<td>GO7</td>
<td>Global market and technical progress</td>
<td>108</td>
<td>1851</td>
</tr>
<tr>
<td>GO8</td>
<td>Economic and technological changes</td>
<td>73</td>
<td>1795</td>
</tr>
<tr>
<td>GO9</td>
<td>Doctors and medical system</td>
<td>92</td>
<td>2687</td>
</tr>
<tr>
<td>GO</td>
<td>TOTAL: all texts</td>
<td>736</td>
<td>14930</td>
</tr>
</tbody>
</table>

Table 1: German dataset

<table>
<thead>
<tr>
<th>textID</th>
<th>topics</th>
<th>sent</th>
<th>tok</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZ1-5</td>
<td>Germany, politics and history</td>
<td>170</td>
<td>687</td>
</tr>
<tr>
<td>CZ6</td>
<td>Housing</td>
<td>83</td>
<td>1644</td>
</tr>
<tr>
<td>CZ7-8</td>
<td>Technological changes</td>
<td>73</td>
<td>1795</td>
</tr>
<tr>
<td>CZ9-12</td>
<td>Politics</td>
<td>121</td>
<td>1854</td>
</tr>
<tr>
<td>CZ13-14</td>
<td>Economics</td>
<td>149</td>
<td>2568</td>
</tr>
<tr>
<td>CZ15-16</td>
<td>Unemployment</td>
<td>112</td>
<td>2252</td>
</tr>
<tr>
<td>CZ17</td>
<td>Television</td>
<td>55</td>
<td>969</td>
</tr>
<tr>
<td>CZ</td>
<td>TOTAL: all texts</td>
<td>763</td>
<td>11769</td>
</tr>
</tbody>
</table>

Table 2: Czech dataset

Although these two data sets were annotated within two different frameworks, the data is comparable, see our discussion (Lapshinova et al., 2015).

3.2. Scheme for Analysis

In (Lapshinova et al., 2015), an attempt was made to unify the Czech and the German-English frameworks for the an-
Table 3: Categories of the interoperable scheme

<table>
<thead>
<tr>
<th>featID</th>
<th>framework for Czech</th>
<th>framework for German</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td>coreference with pronouns</td>
<td>coreference with heads (no extended reference)</td>
</tr>
<tr>
<td>id2</td>
<td>pronouns with arrows to segments and events</td>
<td>reference to verb phrases and longer segments</td>
</tr>
<tr>
<td>id3</td>
<td>NP coreference</td>
<td>coreference with modifiers or def.articles</td>
</tr>
<tr>
<td>id4</td>
<td>coreference with the word same</td>
<td>general comp.reference</td>
</tr>
<tr>
<td>id5</td>
<td>coreference with local and temporal adverbs</td>
<td>coreference with local and temporal adverbs</td>
</tr>
<tr>
<td>nonid1</td>
<td>relations of MERONYMY</td>
<td>relations of MERONYMY</td>
</tr>
<tr>
<td>nonid2</td>
<td>bridging CONTRAST</td>
<td>particular comparative reference and antonyms</td>
</tr>
<tr>
<td>temp</td>
<td>temporal</td>
<td>temporal</td>
</tr>
<tr>
<td>cont</td>
<td>contingency</td>
<td>causal</td>
</tr>
<tr>
<td>comp</td>
<td>comparison (contrast)</td>
<td>adversative</td>
</tr>
<tr>
<td>expan</td>
<td>expansion</td>
<td>additive</td>
</tr>
<tr>
<td>ellipsis</td>
<td>textual ellipsis</td>
<td>cohesive ellipsis</td>
</tr>
</tbody>
</table>

notion of discourse properties. The creation of an interoperable scheme requires a comparison of the underlying annotations. So, we annotated a small corpus of comparable English texts according to the two separate frameworks. This dataset served us as basis for identifying overlapping annotation categories and creating an interoperable scheme. In the present study, we use this scheme to test whether this can be applied for contrastive analyses of Czech and German, which can be extended to more general comparisons of Germanic and Slavic languages in the future. The whole scheme is illustrated in Table 3. The main categories are labelled as IDENTITY, NON-IDENTITY, ELLIPSIS and DISCOURSE RELATIONS. The category of IDENTITY, or coreferential relations, are further specified into five groups according to the form of anaphoric expressions:

- Pronominal coreference (id1) with pronouns referring to nominal antecedents, e.g. Ludwig Erhard – er [he] in example (2).

(2) a. GO: Als Superstar der sozialen Marktwirtschaft gilt aus gutem Grund Ludwig Erhard. Er hatte... in den 50er Jahren... die produktiven Kräfte der Unternehmen entfesselt und daraus ein Wirtschaftswunder gezaubert... Ludwig Erhard is regarded as the superstar of the social market economy, and for good reasons. ...in the nineteen-fifties,... he had unleashed the productive forces of business and in this way conjured up an economic miracle...]

b. CZ: Ta přijala strategii Bílého domu v domnění, že je to nejistější cesta k vítězství. [She endorsed the White House strategy, believing it to be the surest way to victory.

- Abstract coreference (id2) with pronominal anaphors linking up to complex antecedents such as clauses, sentences and longer stretches of text, see example (1) above.

- Nominal coreference, where anaphors are realised in text by nouns with (in German, see Gewerkschaften – den Gewerkschaften in example (3-a)) or without a modifier, as in Czech, see Prahu – Prahy in example (3-b) (id3).

(3) a. GO: Staatstragender können Gewerkschaften kaum sein. Auch wenn... Ludwig Erhard von den Gewerkschaften nicht viel hielt... [Greater loyalty to the state can hardly be expected of a trade union. Despite the fact that... Ludwig Erhard did not think much of the trade unions...]

b. CZ: Zaím se posunuje stále více za Prahu... Po dálnicí bychom se měli svět z Prahy až do Českých Budějovic... [So far, people are moving away from Prague... Highways should take us from Prague all the way to České Budějovice...]

- coreference with anaphors including the word same (id4), see (4), and

(4) And then we do this process again. It’s really exactly the same process every time.

- coreference with local and temporal adverbs as anaphors (id5), e.g. Lissabon – there.

The NON-IDENTITY category includes the relations of MERONYMY (nonid1) and CONTRAST (nonid2) as these categories correspond in both frameworks. Meronymy relations are generally taken part-whole relations between lexical items, such as Germany – the Ger-
mans in (5-a), studio apartments – kitchens in (5-b) and so on.

(5) a. GO: ...praktisch wird es dazu nicht kommen –... expenditure on pharmaceutical research, as well as its share of new active-substance discoveries, is declining.

CONTRAST covers (again, generally taken) antonymy between nominal groups (such as Halbierung – Verdoppelung [halving – doubling] in example (6-a)) and relations termed as comparative reference, e.g. cars – a smaller car.

(6) a. GO: Dazu gehören zum Beispiel die Halbierung der Energie- und Rohstoffintensität bis 2020 gegenüber 1990 (bzw. 1994) und die Verdoppelung des Anteils erneuerbarer Energien am Energieverbrauch bis 2010. [For example, halving the amount of power and raw material consumption by 2020 compared to 1990 (or 1994) levels and doubling the percentage of renewable energy used as part of total energy consumption by 2010.]

b. CZ: Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 USD... I když letos a přiští rok je nutné počítat se zpomalením růstu vývozu, prognosticujeme, že saldo přesto zůstane kladné. [The balance of the current account deficit is estimated to reach $600 last year ... Although this and the next years we expect the slowdown in export growth, we forecast that the deficit will still remain positive.]

Similarly, we include four subclasses of DISCOURSE RELATIONS, i.e. logico-semantic relations that are signalled by a discourse marker or a conjunction:

- temporal relations (temp), e.g. als [when] in (7-a) for German or potom [then] in (7-b) for Czech.

(7) a. GO: Als in Osteuropa der Kommunismus stürzte, hätten viele, die dabei mitkamen, gerne etwas von ihm gerettet. [When communism collapsed in Eastern Europe, many of the people involved would gladly have kept individual aspects of it].

b. CZ: Poslušač musí přístoupit na pozici, že vše je dovoleno. Potom se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace. [The listener has to accept the fact that everything is permitted. Then he can enjoy himself and also understand that the drama symbolizes the loss of a real-life communication.]

- relations of contingency or cause (cont), e.g. deshalb [this is why] in (8-a) for German or proto [therefore] in (8-b) for Czech.

(8) a. GO: Aber nur in den wenigsten ist diese Organisation ein dynamisches Element der Volkswirtschaft. Deshalb irritiert ausländische Beobachter auch oft... [This is why foreign observers are often confused...]

b. CZ: Zatímco většina fotbalových reprezentací vystupuje do kvalifikace pro ME 1996 nyní v zátiší, boj o účast v Anglii vypukl již dříve. (...) Před opravdovým rozjedzem kvalifikace proto přinášíme přehled, jak často spolu celký v jednotlivých skupinách už v soutěžích ME a MS v minulostí hrály. [While most national football teams entered the qualification for the 1996 European Championship now, in September, the fight for a place at the competition in England started earlier. Before the real start of the qualification, we therefore provide an overview of how often the teams in each group had played each other at European and World Championships in the past.]

- relations of contrast (comp), e.g. aber in (9-a) for German, and však [however] in (9-b) for Czech.

(9) a. GO: Arbeiten wie die Polen, aber leben wie die Japaner... [Work like the Poles, but live like the Japanese...]

b. CZ: Poslední statistické sčítání dopravy proběhlo v roce 1990. Za poslední tři roky se však na českých silnicích zvýšil provoz. [The latest statistical traffic census took place in 1990. Over the past three years, however, traffic on Czech roads has increased.]

- relations of expansion or addition (expan), such as ebenso in example (10-a) in German or a [and] in (10-b) in Czech.

(10) a. GO: Tendenziell ist der Anteil der deutschen Pharmabranche an den globalen Forschungsausgaben der Branche, ebenso wie der Anteil an der Zahl neuer Wirkstoffe, aber rückläufig. [Even so, its share of global expenditure on pharmaceutical research, as well as its share of new active-substance discoveries, is declining.]
b. CZ: Vládní plán je podle Jana Švarce ambiciózní a počítá v této oblasti s investicemi 85 miliard korun do roku 2005. [According to Jan Švarc, the government plan is ambitious and it envisages in this area the investment of 85 billion crowns in 2005.]

Note that all kinds of structural types are analysed, such as connectives of main clauses, subordinators and also adverbials. NOMINAL ELLIPSIS includes only nominal constructions as this type is available in both frameworks. We demonstrate an example of a nominal elliptical construction in example (11-a).

\[(11)\]
\[\begin{align*}
\text{a. GO: All das ist eine kleine Revolution. Die grössere & ist diese: [But there is also a bigger [revolution], and it is this:]} \\
\text{b. CZ: Klienti pojištoven, které ukončí svou činnost, se automaticky vrátí k Všeobecné &]. [Clients of insurance companies which shut down will automatically return to the General one].}
\end{align*}\]

4. Analyses and Results

We now analyse the categories in both languages with respect to their overall distribution, the degree of explicitness, as well as the type of textual categories that are preferred. Moreover, we examine variation in the degree of dependence of these textual phenomena on lexico-grammatical constraints or pragmatic peculiarities.

First, we compare the distributional characteristics of the German and Czech data. We produce box plots for analysing variance and significant differences between both data sets. Box plots are median-oriented graphics that represent a convenient way of depicting groups of numerical data through their quartiles which are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. Box plots have lines extending vertically from the boxes (whiskers), indicating variability outside the upper and lower quartiles. Notched boxplots reveal if the differences between the variables under analysis are significant: if two boxes' notches overlap, then there is no 'strong evidence' that their medians differ, see Chambers et al. (1983). The boxplots in Figure 1 demonstrate that German (GO) and Czech (CZ) texts do not differ significantly in their overall degree of cohesiveness if all four categories are taken together.

The differences get pronounced if we compare the distributions for each category, see the barplot based on the normalised (per 10000) overall frequencies per relation in Figure 2. So, we observe more variation for identity and discourse relations, while the frequency distribution for ellipsis and non-identity is similarly low.

Taking a closer look into the subcategories in Table 4, illustrating overall frequencies per category (normalised per total number of words in texts), we find that the higher frequencies of IDENTITY relations in Czech exclusively stem from id3, as numbers are higher in German in all other identity types. Qualitative analyses show that more coreference relations are underspecified in Czech than German in terms of explicit accessibility markers, since the definite article does not exist in Czech and accessibility of referents is indicated by information structure more often than in German. On the one hand, the difference in the amount of identity relations is due to the discrepancies in the annotation framework: repetitions of named entities are annotated in the German framework within lexical cohesion. The other repetitions, if coreferent, are included into the annotation of identity relations, since they will be almost always modified either with a demonstrative pronoun or a definite article. Qualitative analysis of the chains in the data shows that, for instance, in the text containing the chain consisting of Gewerkschaften – die Gewerkschaften, part of which was illustrated in example (3) above, later on, there is another mention of Gewerkschaften without an article or a demonstrative, which is used in a general meaning (Gewerkschaften gibt es in vielen Ländern)
Table 4: Frequencies of discourse categories

<table>
<thead>
<tr>
<th>featID</th>
<th>German</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td>88.41</td>
<td>97.71</td>
</tr>
<tr>
<td>id2</td>
<td>38.18</td>
<td>64.58</td>
</tr>
<tr>
<td>id3</td>
<td>144.68</td>
<td>597.33</td>
</tr>
<tr>
<td>id4</td>
<td>3.35</td>
<td>0.00</td>
</tr>
<tr>
<td>id5</td>
<td>12.06</td>
<td>10.20</td>
</tr>
<tr>
<td>nonid1</td>
<td>52.91</td>
<td>88.37</td>
</tr>
<tr>
<td>nonid2</td>
<td>28.80</td>
<td>37.39</td>
</tr>
<tr>
<td>temp</td>
<td>106.50</td>
<td>14.44</td>
</tr>
<tr>
<td>cont</td>
<td>52.24</td>
<td>66.28</td>
</tr>
<tr>
<td>comp</td>
<td>79.04</td>
<td>86.67</td>
</tr>
<tr>
<td>expan</td>
<td>181.51</td>
<td>136.80</td>
</tr>
<tr>
<td>ellipsis</td>
<td>14.07</td>
<td>50.13</td>
</tr>
</tbody>
</table>

Table 4: Frequencies of discourse categories

trade unions in many countries). This would be a part of the same lexical chain as the other mentions of Gewerkschaften, but is not coreferent and, cannot be considered as an extension for the German coreference chain here. In languages with the definite article, anaphoric expressions mostly contain a formal definite marker which allows to (even automatically) extract most anaphors from the corpus. Czech, as a Slavic language without definite article, does not dispose a formal means with the help of which anaphoric expressions can be easily found and annotated. Thus, the annotation is completed on the base of semantic and referential criteria: everything that refers to the same discourse entity, according to the annotator, is marked as coreferential.

By contrast, the frequencies for DISCOURSE RELATIONS are higher in German than in Czech. As a similar tendency was observed in comparison to English (see e.g. Kunz et al. (in press)), German seems to be exceptional in signaling logico-semantic relations by an explicit discourse marker, especially in terms of temporal relations or relations of expansion, as it is seen in Figure 3.

As for NON-IDENTITY, we foresee much higher frequencies when integrating further relations in the future. Finally, the higher number of NOMINAL ELLIPSIS in Czech than German points to a higher preference for expressing comparison by fragments. This tendency towards implicitness may, however, stem from the greater syntactic flexibility of Czech relative to German.

5. Conclusion and Discussion
We have performed a cross-lingual analysis of discourse phenomena, using resources annotated along two different frameworks. Our preliminary results show that interoperable schemes like the one used here permit a multilingual analysis of discourse-annotated corpora originating from different approaches. On the one hand, we are able to validate the interoperable scheme in an application. On the other hand, the successful application of the scheme indicates possible interoperability in existing resources. In this way, our methodology saves time and effort as no compilation of additional resources is required. This is especially valuable for multilingual NLP which usually requires multilingual data sets annotated according to the same scheme to build appropriate language models. Creation of such data sets is costly and time-consuming, and our approach can be a good solution in this case. Furthermore, the results yield first insights into differences between German and Czech in terms of the annotated phenomena. At the same time, we are aware of the limitations the dataset at hand provides: although the texts are from the same text genres and have similar topics, the variation observed may be author- or source dependent, since the size of the dataset is small. Some of the differences could also be explained by the differences in the conceptualisations in the schemes. Nevertheless, our work is an important first step towards better comparing and harmonising available resources that are already enriched with annotations. Our future plans include an expansion of the analyses in terms of corpus size, languages and factors influencing variation, e.g. authors, topics. A deeper analysis of textual examples in German and Czech will help us to improve and to refine the analysed categories. Moreover, we plan to include spoken data into our analyses, and compare the distribution of discourse relations across spoken and written dimensions in both languages. Additionally, we intend to test this scheme in further applications, e.g. for machine translation or other NLP areas.

Acknowledgements
This work was made possible within the Textlink Action (ISCH COST Action IS1312). We also acknowledge support from the Grant Agency of the Czech Republic (grant 16-05394S). This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society). The German corpus can be queried via the CQPWeb interface provided through CLARIN-D (hdl: 11858/00-246C-0000-0023-8CF7-A).

6. Bibliographical References

7. Language Resource References
Cross-linguistic analysis of discourse variation across registers

Kerstin Kunz, Heidelberg University and Ekaterina Lapshinova-Koltunski, Saarland University

Abstract
The present study deals with variation in discourse relations in different registers of English and German. Our previous analyses have been concerned with the systemic contrasts between English and German, cf. Kunz & Steiner (2013 a/b), Kunz & Lapshinova (to appear) and have addressed some cross-linguistic differences with regard to textual realizations of selected subtypes of cohesion. In our current work, our focus is on the empirical analysis of cross-linguistic variation between registers. In order to obtain a more comprehensive picture, we investigate three main types of cohesion in combination: co-reference, substitution and conjunction and their subtypes, cf. Halliday & Hasan (1976). We extract instantiations of cohesive devices from an English-German corpus of spoken and written registers. The data is analyzed with statistical procedures which show that subcorpora can be grouped along particular combinations of cohesive devices.

1. Research objective
Our aim in this study is to interpret corpus data about three types of cohesion—co-reference, substitution and conjunction—as well as their subtypes. Recent years have seen an increase in the number of works employing corpus-based methods for the study of cohesion. However, multilingual works are mostly concerned with individual cohesive devices in particular registers, e.g. Neumann et al for repetition, Zinsmeister et al. (2012) for abstract anaphors, Bühlig & House (2004) for particular cohesive conjunctions or adverbs, and Taboada & Gómez-González (2012) for particular coherence relations. Most studies that analyse particular types of cohesion work with one language only. For instance, corpus-based works concerned with conjunction are Stede (2008), Dipper & Stede (2006), Bestgen et al. (2006) and works on coreference are Eckert & Strube (2000), Gundel et al. (2004). Works on substitution in the sense of Halliday & Hasan (1976) are rare (see Kunz & Steiner 2013b), also because it seems to be a less frequent cohesive phenomenon.

The main objective of our previous studies was to identify German-English contrasts in the realization of particular types and subtypes of...
Discourse variation across registers
cohesion both from a systemic and textual perspective, c.f. Kunz and
Steiner (2013a) for co-reference, Kunz and Steiner (2013b) for
substitution and Lapshinova & Kunz (2014) and Kunz & Lapshinova (to
appear), for conjunctive relations. These types will now be examined in
combination in order to obtain a more comprehensive expression of
textual features. The focus of this study is on the analysis of variation in
the registers that were collected in the GECCo corpus - a bilingual
corpus of English and German that comprises 12 different registers of
English and German. Although our current interest lies in the cohesive
devices which serve as explicit indicators of textual relations across
grammatical domains, we also consider the ties (and chains in case of co-
reference) established by these devices. There are several corpora which
have already been annotated on the level of discourse, cf. Doddington et
al. (2004) and Pradhan et al. (2011), the OntoNotes Corpus for English,
Arabic and Chinese, (Weischedel et al. 2013) and the TüBa-D/Z corpus
(Hinrichs et al. 2005) or the Prague dependency Treebank, and the Penn
Discourse Treebank for English, cf. Prasad et al. (2008). These corpora
cannot be employed for our research on English and German as (1) they
do not contain comparable registers across languages; (2) most of them
do not contain annotations of different types of cohesion; and (3) they do
not provide enough register variation to permit analyzing register as a
variable. To our knowledge, our corpus is the only existing resource that
allows for an investigation of different cohesive phenomena cross-
linguistically and across different registers at the same time.

For this study, only those ten registers from the GECCo corpus were
selected for analysis where the annotation obtained from semi-automatic
annotation procedures has already undergone in-depth correction phases
by human annotators. The data obtained has been evaluated by a
combination of different statistical evaluation procedures. This
methodology permits comparison of distinctive main types and subtypes
of cohesive devices and clustering of registers according to particular
types. It thus facilitates a sound interpretation that goes beyond the level
of grammar towards more abstract conceptual ranks.

Most importantly, it permits us to address the research objectives
pursued in the frame of the current study. The main question we are
interested in is whether contrasts are more pronounced between different
registers independent of language or whether more differences are
identified in one and the same register between English and German.
Targeting this question, several others arise. For instance, we intend to identify which registers show most similarities/differences, within one language and across languages. Next, we want to identify those registers that are most pronounced in the realization of particular features. And finally, we are also interested in the features that contribute to the observed differences/commonalities.

Before we deal with these questions, we will provide a brief definition of cohesion which is followed by a conceptual clarification of the three cohesive types under investigation. We will describe in short our corpus resource and the procedures employed for annotating co-reference, substitution and conjunction. The centerpiece of our study will be the evaluation of the obtained data via various statistical methods such as descriptive analysis and correspondence analysis and their interpretation in terms of the questions highlighted above.

2. Cohesion and cohesive subtypes
Let us now move on to a short discussion of the concept of cohesion and the subtypes under investigation. Note that details on systemic differences between English and German can be found in Kunz & Steiner (2013a) for co-reference, Kunz & Steiner (2013b) for substitution and Kunz & Lapshinova (to appear) for conjunction.

Language producers employ particular lexicogrammatical items (cohesive devices) which indicate a linguistic relation to other textual elements across grammatical domains. The explicit linguistic ties or chains which are created by language producer on the text surface (cohesive relations) help recipients in their cognitive interpretation as to how different thematic concepts are connected. Cohesion can be signaled in texts by grammatical items such as personal and demonstrative pronouns and modifiers, substitute forms, elliptical constructions and conjunctions or by lexical devices such as verbs, nouns and adjectives. These cohesive devices trigger different semantic relationships, whose borderlines however are often blurred.

As mentioned above, the focus of this study is on three main types of cohesion (see further Halliday & Hasan 1976, Halliday & Matthiessen 2013): co-reference, substitution and conjunction and on their subtypes. Therefore, we will now examine their peculiarities in more detail.
Let us start by looking at the features shared by all three types under investigation. What co-reference, substitution and conjunction have in common is that there are explicit linguistic devices signalling particular conceptual relations to linguistic elements in other clauses, sentences or paragraphs of the same text/discourse, and that the interpretation of the cohesive devices is dependent on the elements they tie up with (see Halliday & Hasan 1976). All three types are regarded in the literature as being grammar-driven (see e.g. Louwerse and Graesser 2005, Brinker 2005, Schubert 2008) since the devices that trigger the cohesive relations belong to a closed class of functional items, in contrast to devices of lexical cohesion, which comprise open classes of nouns, verbs, adjectives and adverbs. Grammar-driven items are quite often semantically weak (see Halliday & Hasan 1976) and it is this semantic reduction which initiates a search for other linguistic elements in the text on the basis of which the intended meaning can be fully interpreted. For an illustration, consider (1) to (3) below, where cohesive devices are marked in bold and antecedents/elements connected by a cohesive device are in square brackets.

(1) Do not turn on [the computer]. Turning it on before you're finished assembling the system could …

(2) [We have to be honest about the challenges facing Europe]. And [we have to listen to what Europe’s voters are telling us].

(3) ‘I don’t have a [car], ‘he said. ‘If I borrowed one, would you …?’

In (1) the referential device employed is the neuter personal pronoun it. It refers to an entity whose semantic class (e.g. computer vs. printer) can only be identified by looking at the antecedent the computer. The cohesive item and in (2) belongs to the closed class of coordinators and only signals that there is an additive relation between two entities without but does not provide any information as to which kind of entities are involved. In (3) the substitutional form one is an indefinite pronoun that indicates similarity with or selection of a referent out of a particular class but again, does not give a clue as to the intended class or referents.

Although many devices of co-reference, substitution and conjunction conform to the above descriptions, there also items which are semantically richer and therefore less grammatical. This is particularly the case with comparative co-reference and also various conjunctive
adverbials, which are on the borderline between grammatical and lexical cohesion.

Co-reference, substitution and conjunction however differ in their lexico-grammatical realizations and, most importantly, in the type of semantic relation they preferentially express. These differences are discussed in the following.

2.1 Co-reference
The conceptual relation created by co-reference is one of identity. Co-referential devices signal that they point to a referent which has already been mentioned by a referring expression in another (mostly) preceding textual part (antecedent), hence to a referent that is textually old or given (cf. Prince 1981, Gundel et al. 1993, Ariel 1990). The referential devices can therefore be considered as search instructions to other textual elements on the basis of which the intended referent is cognitively identified (cf. Schwarz 2000:43). Following the classifications by Halliday & Hasan (1976), we distinguish three subtypes of co-reference:

- **personal**: relations triggered by personal pronouns (e.g. *it/es, they/sie*), possessive pronouns and modifiers (e.g. *his/sein(e,r,s)*)
- **demonstrative**: triggered by definite article, demonstrative pronouns (e.g. *this, that/dies, das*) and modifiers (e.g. *this/diese (r,s)*), local and temporal adverbs (e.g. *here, then/hier, da*) and pronominal adverbs (e.g. *herewith/hiermit*)
- **comparative**: triggered adjectives and adverbs of comparison (such as *similar/ähnlich* or *such/solche*¹)

It has to be noted that the category of comparative reference is semantically distinct from the other two main types as it does not create identity of reference but rather evokes a relation of similarity and comparison between referents, events or propositions of the same type (see e.g. Halliday & Matthiessen 2004:560, Schubert 2008:35). For instance in (4), *another* in combination with *explanation - device of*

¹ See Kunz & Steiner (2013) for a more detailed discussion of semantic and functional differences.
Discourse variation across registers

lexical cohesion - ties with the preceding sentence but does not create identity.

(4) It is said that the French soldiers saw the Welsh women from a distance in their tall hats, thought they were soldiers and surrendered! There may be another explanation but …

Devices of comparative reference and also of demonstrative reference are combined with devices of lexical cohesion in case the referential device is a modifier (e.g. *these explanations*) or a comparative adjective (e.g. *another explanation*).

2.2 Substitution

The main difference between co-reference and substitution concerns the semantic type of relation: in contrast to co-reference the tie between the cohesive device and its antecedent does not trigger identity between instantiated referents but similarity between referents belonging to the same class (cf. Kunz & Steiner 2013b, de Beaugrande & Dressler 1981).

It therefore exhibits some similarities with comparative reference in the semantic meaning relations established. Substitution can additionally be differentiated from co-reference because of its syntactic constraints, in which it resembles cohesive ellipsis. The formal options available in English (and also in German) for establishing substitution are very limited. We analyze three main subtypes of substitution:

- **nominal**: e.g. signaled by the same and one(s)/Das Gleiche/ dasselbe, eine(r,s) in German
- **verbal**: by *do* (*so*)/ *tun* and *machen* in German
- **clausal**: mainly with the form *so* in English, more variation in German

2.3 Conjunction

The semantic relation of conjunction differs from co-reference and substitution in that conjunctive devices do not refer themselves and therefore do not have an antecedent. They indicate relations between two other textual elements and explicitate logico-semantic relations between referents, which are semantically rather complex such as states,
processes and events (cf. Pasch et al. 2003, Blühdorn 2008). Note that the term ‘conjunction’ used in this study deviates in its conceptualization from most grammars. We depart from the meaning relation established, in the sense of Halliday and Hasan (1976) and Halliday and (Matthiessen 2013: 593ff), who include all forms that signal a cohesive relation between linguistic elements. These forms are termed “conjunctive device” in our work, while “conjunction” refers to the relation as such. The relations that are explicitated by conjunctive devices can be mainly grouped into the following categories (cf. Halliday & Hasan 1976):

- **additive**: relation of addition, for two events that are true/ not true at the same time (conjunctive devices indicating such a relation are e.g. *and, in addition, und, außerdem*)
- **adversative**: relation of contrast/ alternative, for two events which are not true at the same time (*yet, although, by contrast, doch, obwohl, im Gegensatz dazu*)
- **causal**: relation of causality/ dependence between (because, therefore, that’s why, weil, deshalb, aus diesem Grund)
- **temporal**: temporal relation between events (after, afterwards, at the same time, nachdem, danach, gleichzeitig)
- **modal**: This latter category subsumes devices that are not included in most grammars. The meaning rather is an interpersonal or pragmatic one (see e.g. Martin 1992: 178ff) in which conjunctive devices connect events by an evaluation of the speaker (*well, sure, klar, sicher*). In the literature, they often fall under the category of ‘discourse markers’ and are called ‘continuatives’ by Halliday & Hasan (1976) and Halliday & Matthiessen (2013).

Apart from these semantic peculiarities conjunctive devices exhibit distinctive lexicogrammatical features. There is more variation with respect to different forms available as well as the number of elements contained in the conjunctive device. This particularly the case with conjunctive adverbials, which may consist of one adverb only, e.g. *therefore*, or may be multiword constructions, e.g. *for this reason or that’s why*, which may contain lexical as well as referential items. For this reason, there is also more variation in terms of the semantic explicitness of the devices. However, the set of devices linking main
clauses is relatively small. In our analysis we distinguish the two main structural types of conjunctive devices mentioned above:

- **coordinators**: link textual elements in a paratactic construction (e.g. *and, but, neither … nor, und, aber, weder … noch etc*)

- **adverbials**: link clause complexes (sentences) or even elements on a higher textual level (e.g. *therefore, by contrast, deshalb, im Gegensatz dazu, etc.*)

Subordinators, which link main and subordinating clauses, are not included in the present study since they are generally not regarded in the literature as establishing relations of cohesion.

Conjunctive devices also exhibit more restrictions in their syntactic function and position: coordinators do not serve as fully-fledged syntactic constituents, and conjunctive adverbials only take on the function of a syntactic adverbial. The most common position of conjunctive devices is between the first and the second textual element they link, although there is more variation for conjunctive adverbials; see Kunz & Lapshinova (to appear).

If we look at instantiations of co-reference, substitution and conjunction, we sometimes notice that the borderlines between the three cohesive types are blurred to the extent that one and the same cohesive form may serve as referential, substitutional or conjunctive device, dependent on the context in which it is realized.

(5) And we also we have the Dee river on one side of the peninsula and the Mersey on the other. - So the peninsula is between the two rivers. - Yes, yes. Right. (causal conjunction)

(6) It's a financially driven issue that local authorities who are responsible for providing care will do so of course in the least cost way possible. (clausal substitution)

(7) So pflegen Sie Ihr Gerät … : Es genügt, wenn Sie das Gerät nur feucht abwischen. (demonstrative reference or clausal substitution)

So (‘in this way’) you take care of your instrument… It suffices when you wipe the instrument damp. (literal translation)
As examples (5) to (7) illustrate, this is especially the case with the form so in English and German, and also with pronominal adverbs in German, which may either serve as devices of co-reference or as devices of conjunction.

3. Data and methods

In the following section, we will first provide information about the GECCo corpus and about the procedures to extract the different types of cohesive devices before we will shortly describe the statistical methods applied for data evaluation.

3.1 Corpus resources

As mentioned in section 2 above, our research is based on data extracted from the GECCo corpus (cf. Kunz and Lapshinova-Koltunski, 2011 and Lapshinova et al., 2012), a German–English corpus of different written and spoken registers. The corpus contains ca. 1.3m tokens. For this particular study, we analyze four subcorpora only: German written originals (GO), English written originals (EO), English spoken originals (EO-SPOKEN) and German spoken originals (GO-SPOKEN). The two written subcorpora consist of texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and corporate websites (WEB). The two spoken subcorpora contain academic speeches (ACADEMIC) and interviews (INTERVIEW). The corpus also contains two further subcorpora: English and German translations of EO and GO. However, in this study, we analyze cohesive phenomena in subcorpora of originals only, which provide a good database for our English-German contrastive analysis. The subcorpora, from which we extract frequency information on the occurrence of cohesive phenomena, are presented in Table 1.

The corpus is annotated on several levels, and annotations include information on tokens, lemmas, morpho-syntactic features (e.g. case, number, etc.), parts-of-speech, phrase chunks and their grammatical functions, as well as and sentence boundaries. The annotation of the written part was partly imported from CroCo, whereas for the spoken
Discourse variation across registers

part, we use Stanford POS Tagger (Toutanova et al., 2003) and the Stanford Parser (Klein and Manning, 2003). The corpus is encoded in the CWB format (CWB, 2010) and can be queried with Corpus Query Processor (CQP) (Evert, 2005). These annotation levels provide us with additional information on cohesive phenomena and cohesive relations, i.e. for co-reference: morpho-syntactic preferences of referring expressions, such as their positions in a clause, etc.

Table 1. Variables for corpus-based analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO</td>
<td>GO</td>
</tr>
<tr>
<td>EO_ACADEMIC</td>
<td>GO_ACADEMIC</td>
</tr>
<tr>
<td>EO_FICTION</td>
<td>GO_FICTION</td>
</tr>
<tr>
<td>EO_ESSAY</td>
<td>GO_ESSAY</td>
</tr>
<tr>
<td>EO_INSTR</td>
<td>GO_INSTR</td>
</tr>
<tr>
<td>EO_INTERVIEW</td>
<td>GO_INTERVIEW</td>
</tr>
<tr>
<td>EO_POPSCI</td>
<td>GO_POPSCI</td>
</tr>
<tr>
<td>EO_SHARE</td>
<td>GO_SHARE</td>
</tr>
<tr>
<td>EO_SPEECH</td>
<td>GO_SPEECH</td>
</tr>
<tr>
<td>EO_TOU</td>
<td>GO_TOU</td>
</tr>
<tr>
<td>EO_WEB</td>
<td>GO_WEB</td>
</tr>
</tbody>
</table>

Moreover, the corpus is annotated with information on cohesive devices. For this, semi-automatic procedures were applied, which include a rule-based tagging of cohesion candidates and their manual post-correction by humans. A description of the procedures is given in Lapshinova and Kunz (2014). For this study, we deploy the annotation of cohesive devices establishing co-reference, substitution and conjunction, whose subtypes and functions as annotated in the corpus are given in Table 2. These subtypes serve as categories for our corpus-based analysis described in section 5 below.
Table 2. Categories of phenomena under analysis

<table>
<thead>
<tr>
<th>device</th>
<th>co-reference</th>
<th>conjunction</th>
<th>substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>personal head, personal modifier demonstrative head demonstrative modifier demonstrative local demonstrative temporal pronominal adverbs definite articles comparative general comparative particular</td>
<td>additive connectives additive adverbials adversative connectives adversative adverbials causal connectives causal adverbials causal adverbials temporal adverbials modal adverbials</td>
<td>nominal verbal clausal</td>
</tr>
</tbody>
</table>

3.2 Data extraction

The instantiations of the categories presented in Table 2, can be easily extracted from the corpus, as their information is annotated and can be queried with CQP. In Table 3, we provide examples of queries used for the data extraction.

Table 3. Query examples used to extract the categories under analysis

<table>
<thead>
<tr>
<th>Query</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[_.reference_func=&quot;poss.*&quot;]</td>
</tr>
<tr>
<td>2</td>
<td>[_.reference_func=&quot;temporal&quot;]</td>
</tr>
<tr>
<td>3</td>
<td>[<em>.conj_func=&quot;additive&quot; &amp;</em>.conj_type=&quot;connect&quot;]</td>
</tr>
<tr>
<td>4</td>
<td>[<em>.conj_func=&quot;additive&quot; &amp;</em>.conj_type=&quot;adverbial&quot;]</td>
</tr>
<tr>
<td>5</td>
<td>[_.substitution_type=&quot;verbal&quot;]</td>
</tr>
<tr>
<td>6</td>
<td>[_.substitution_type=&quot;clausal&quot;]</td>
</tr>
</tbody>
</table>
Discourse variation across registers

For instance, with the help of query 1, we can identify how many referential devices function as personal modifiers (possessive determiners), whereas query 2 is used to identify all cases of referential devices with a temporal function (temporal adverbs). Queries 3 and 4 are built to differentiate between coordinating conjunctions expressing additive relations and conjunctive adverbials expressing additive relations. We apply queries like those in 5 and 6 to extract different types of substitution and ellipsis.

3.3 Statistical methods

Descriptive data analysis is employed to compare frequency distributions of main cohesive types and of cohesive subtypes. This allows us to obtain a first insight into differences and commonalities between registers within one language and across languages with respect to preferences of particular types of cohesion.

To validate our data statistically, we use an unsupervised technique – correspondence analysis (CA), cf. Baayen (2008). This statistical procedure allows us, on the one hand, to see which registers have more commonalities and which significantly differ from each other. On the other hand, CA permits identifying the features (in our case cohesive categories) which contribute to these differences or commonalities. This also allows us to distinguish features as indicators of register and language variation. Moreover, we are able to trace the interplay of categories of the cohesive devices under analysis.

We use the CA package (cf. Nenadic and Greenacre, 2007) to perform correspondence analysis in the R environment. An input for CA is frequencies of the categories under analysis across registers. The output of the correspondence analysis is plotted into a two dimensional graph with arrows representing the observed frequencies of cohesive devices and triangles representing the subcorpora. The triangle position to the arrows and their length allow us to interpret their correlation. The length of the arrows indicates how pronounced a cohesive device is, see Jenset and McGillivray (2012) for details. The position of the triangles in relation to the arrows indicates the relative importance of a cohesive device for a subcorpus. The arrows pointing in the direction of an axis indicate a high contribution to the respective dimension.
4. Data analysis

We now present the results from the statistical tests discussed above and interpret them in the light of our research questions, which, for the sake of convenience are repeated below:

a) Main question: Are contrasts more pronounced between different registers independent of language or are more differences identified in one and the same register between English and German?
b) Which register(s) are more similar to each other and which registers are more different?
c) Which register(s) is (/are) most pronounced in the realization of particular features, across English and German?
d) Which features contribute to the observed differences/commonalities?

4.1 Frequency distribution

We begin with the results obtained from the descriptive methods. Figure 1 below shows the frequency distributions of cohesive types and subtypes extracted from the subcorpora mentioned in section 4.1. The distributions are grouped according to the registers in which they occur. Subtypes of co-reference are marked in shades of blue, substitution types are marked in green, and conjunction types are marked in red.
Figure 1 first of all reveals that substitution plays a minor role in all registers of both languages, and that high frequencies can be found for co-reference and conjunction. In addition German seems to show a preference for relations of conjunction, whereas English seems to favour relations of co-reference. Another general observation is that there are considerable differences between the two languages in the distribution of subtypes. For instance, more co-reference relations seem to be realized in English by personal pronouns (personal head) and demonstrative determiners (demonstrative modifier), more substitution relations by nominal substitutes and more conjunctive relations by additive connects. By contrast more co-reference relations are expressed in the German subcorpora via demonstrative pronouns (demonstrative heads) and pronominal adverbs and comparative particular, more relations of substitution by clausal substitutes and more conjunctive relations via modal, additive and adversative adverbials. These language peculiarities
also translate into differences within one register between languages. However, we also see that registers can be identified across languages on the basis of particular cohesive subtypes. In particular, EO_FICTION and GO_FICTION stand out in terms of a heterogeneous distribution of cohesive subtypes, both are characterized by very high frequencies of personal reference and low frequencies of subtypes of comparative reference. The markedness of the register of fictional texts in terms of cohesion in both languages seems to be in line with observations by Neumann (2013:230ff) and Biber (1995: 151ff) in terms of lexicogrammatical features. Taken together, these features seem to reflect communication between non-experts and narrative style. In contrast to FICTION, EO_POPCSI and also GO_POPSCI seem to contain more even distributions of several cohesive subtypes. This goes along with higher frequencies for less reduced cohesive devices, such as demonstrative modifiers and different logico-semantic types of conjunctive adverbials, and may point to a more informational and/ or expository type of production (see Neumann 2013:231ff and Biber 1995:141ff).

4.2 Correspondence analysis
Using the descriptive methods above we obtain information on general differences in frequency distributions between languages and registers. Yet, they do not give any information on the correlation between registers and features. Moreover, it is difficult to trace the groupings of languages and registers, as well as the discriminatory features, i.e. cohesive devices responsible for these groupings. For this, a multivariate statistic method is needed. We therefore use correspondence analysis in order to identify correlations between particular cohesive subtypes and subcorpora.

4.2.1 Analysis across registers per language
We start with the analysis of the correlation of cohesive subtypes in EO and GO separately. Figure 2 demonstrates a two-dimensional graph for English originals.

The x-axis reveals a distinction between English written and spoken subcorpora, although EO_POPSCI and EO_INSTR are also situated on
Discourse variation across registers

the left side of the borderline. However, the summary information for CA reveals their low representation in the graph (mass of 40-60), which means that they are not represented well in this subdivision. The features which contribute to this division include demonstrative reference with a local function, as well as personal reference with a modifier function as being specific for written registers of English. By contrast, demonstrative reference, with head and temporal function, nominal and verbal substitution, general comparative reference and modal conjunctions are specific for the English spoken registers EO_INTERVIEW and EO_ACADEMIC. From the grouping of these features we can conclude that spoken registers can be distinguished from written registers in English by a preference for focusing on complex referents (events). These are marked as particularly relevant or important via demonstratives heads - that and these as in (8). We also notice a tendency towards expressing relations of comparison and similarity via substitution, e.g. do in (9) and comparative reference, e.g. different in (8).

In addition cohesion is often employed as a means for marking interpersonal relations, via modal conjunctions as I mean in (9).

(8) if you’re not convinced by that let me give you a second, way of packing these, if I don’t look a - if I look at a different angle you see a hexagon. [EO_ACADEMIC]

(9) Yes, I like this little figure, yes, I definitely do. Some people describe, I mean, to me is like a little puppet version of myself. OK, OK. And you do, you get quite attached. [EO_INTERVIEW]

Hence, semantically reduced forms are combined with cohesive types that mark an ‘involved’ style (Biber 1995). The latter, however, prevails in EO_INTERVIEW, as can also be seen in figure 1.

However, we also note a separation between the combination of EO_FICTION and EO_INTERVIEW and the rest of the subcorpora with respect to the y-axis. The most prominent cohesive devices in these two registers are devices of personal reference, which serve as nominal heads while the other eight English registers are characterized by the frequent use of various conjunctive subtypes. This feature distinction could mark the separation between dialogic and non-dialogic registers contained in the GECCo corpus, as fictional texts and interviews contain dialogues. However the two registers can be distinguished by clausal substitution as a typical feature of EO_FICTION and causal conjunctions as a
distinctive feature of EO_INTERVIEW, which may potentially reflect the boundary between argumentative and non-argumentative style (see Biber 1995).

(10) *So it's basically assessors coming in and looking at the school as a whole? - Yes, coming in to observe. - Yes, yes, that's it. - Is that quite stressful? Very stressful, yes. It's we had it before Christmas actually last year - and it was the most, because it was my first Ofsted, it was the most stressful thing I think I've encountered at school, so.* [EO_INTERVIEW]

Example (10) illustrates quite nicely that causal conjunctions and personal pronouns are employed in combination in EO_INTERVIEW. Quite often, very short and semantically weak forms are used, such as the neuter personal pronoun *it*, establishing a rather vague connection to previous utterances. Furthermore, conjunctions such as *so* also carry an interpersonal meaning, in that they mark the beginning or the end of a speaker turn or the willingness of the speaker to continue with his/ her speech. Hence, the involved style seems to occur in combination with argumentation here.

From these observations we can conclude that some clusters of cohesive subtypes group registers in English with respect to different modes of production (written vs. spoken) while other features of cohesion reflect other aspects of typical contexts of situation, such as speaker interaction.
The two-dimensional graph for GO registers is shown in Figure 3. Here, we also see a clear distinction between written and spoken subcorpora along the x-axis. Yet, partially differing features seem to contribute to this distinction in German, as compared to English: German written registers are mainly characterized by co-reference relations that are established by personal pronouns and modifiers. As in English, we count more distinctive features for the German spoken registers: verbal and clausal substitution, local and temporal reference and modal and temporal adverbials. Although the distinctive features partially differ from those for English, the realizations in the corpus reveal that the cohesive devices serve to express similar meaning relations.
As in English, we can attest a tendency in the spoken registers for highlighting important referents/ events by demonstratives, for expressing relations of comparison (e.g. by besser (‘better’), as in (12)) and for marking speaker turns and speaker continuity, e.g. also (‘well’) and und (‘and’) (examples 11 and 12). Yet, these devices differ from those of English in the specific meaning relations they express. In contrast to English, German spoken registers additionally show a tendency towards expressing more cohesive relations of location and time. However, this does not necessarily point to a more content-oriented style of production (see House 1997). For an illustration see the textual passages below:

(11) Und gibt es da noch sprachliche Probleme bei denen oder sind die sprachlich völlig integriert? - Also da sind die Probleme massiv und da sieht man eigentlich auch, wo des Problem bei uns in Deutschland liegt, ja, dass die einfach mit zu geringen Sprachkenntnissen [GO_INTERVIEW] …

(12) And are there also linguistic problems amongst them or are they fully integrated linguistically? Well, there the problems are
Discourse variation across registers

huge and there you can also see, what the problem really is in Germany …

(13) Und da nehme ich an, ich könnte auch sagen eine endliche Menge. Aber irgendwie ist es besser, wenn ich sage eine unendliche Menge [GO_ACADEMIC]

(14) And there I expect, I could also say a finite number. But somehow it is better if I say an infinite number …

In examples (11) and (12) we find the locative adverb da (‘there’), which is used with a very high frequency in the spoken registers of German. If we compare instantiations of da, we observe that it is employed with a multitude of cohesive meanings, sometimes it tends to express time (in the sense of then) rather than location, and sometimes its meaning is rather metaphorical and mainly textual (as in 12). These observations suggest a combination of our empirical findings with insight from quantitative analysis in the future. Statistical evidence for differences in the spoken registers between the two languages are examined with the combined analysis below.

Concerning the y-axis, we observe a similar tendency as in English: fictional texts are opposed to the other registers. GO_POPSCI, GO_INSTR and GO_ESSAY, which are located on the same side of the separation line as GO_FICTION are poorly represented in this dimension. This means that German FICTION is a very distinctive register in terms of cohesive categories and, in contrast to English, does not cluster with EO_INTERVIEW. Most prominent features, which contribute to this distinction, are cohesive personal heads and demonstrative articles. Hence, this separation is no reflection of dialogicity but may rather have to do with the narrative passages contained in the German fictional texts. The passages contain rather long co-reference chains which denote the same protagonist again and again, via semantically reduced pronouns. The protagonist is involved in events or concerned with various objects which are realized in smaller co-

\footnote{Note that the English translations for passages from the German spoken registers are literal in order to reflect German peculiarities of cohesion. The translations for the passages for German written registers are taken from the GECCo corpus.}
reference chains that are triggered by combinations of definite articles and lexical cohesion.

(15) *Man konnte [den Schatten] vor uns zusehen, wie er sich näherte, bis er unter [der Motorhaube] verschwand, kurz darauf die Motorhaube erklomm, über die Windschutzscheibe kroch, auf unsere Gesichter, und schließlich den Wagen verschluckte, rücksichtslos, wie er alles verschluckte, was vor ihm lag. [GO_FICTION]*

(16) *The shadow in front of us could be seen approaching until it vanished below the hood, climbed the hood a moment later, crawled across the windshield onto our faces, and finally swallowed the car as ruthlessly as it swallowed all that lay before it, the shadow of that wide roof, of the building that straddled the road and blocked our view. [ETRANS_FICTION]*

In the passage presented in (13), the protagonist is a personified object – der Schatten (‘the shadow’), which is taken up by masculine personal pronouns in singular (er, ihm). It performs several actions in which objects such as die Motorhaube (‘the hood’) are involved.

### 4.2.2 Analysis across registers and languages

In the next step, we compare all subcorpora under analysis, which is represented in the two-dimensional graph in Figure 4.
A clear separation between the two languages is observed along the $x$-axis. Characteristic cohesive features for German registers include demonstrative reference with local and temporal functions and pronominal adverbs, conjunctive relations expressed with adverbials as well as clausal substitution. In English, personal heads, definite articles and general comparative reference contribute to distinction. This corroborates our earlier observations based on the descriptive analyses described in section 5.1 above.

Hence, from a semantic perspective German prefers to mark recurring referents as particularly relevant or important by focus lifters such as demonstrative adverbs, e.g. *da* in (14), demonstrative pronouns, as *der* in (14), which does not have an equivalent in English and pronominal adverbs, as in (15). The high frequency of the latter is a peculiarity of the German language system.

(17) *und plötzlich verwandelt sich einer von den Protagonisten in ein Nashorn. Und plötzlich verwandeln sich immer mehr in ein Nashorn, können danach nicht mehr sprechen. Sie haben also da auch das Problem der Sprache und es gibt einen einzigen, der*
bleibt übrig. Der verwandelt sich nicht eh und der denkt darüber nach,... [GO_ACADEMIC]

(18) And suddenly, one of the protagonists turns into a hippopotamus. And suddenly, more of them are transformed into a hippopotamus, not able to speak afterwards. So there they have the problem with language and there is only one person, he remains. He does not transform and he thinks about ...

(19) Andererseits müssen wir den Dialog gerade mit der islamischen Welt verstärken und intensivieren. Dabei geht es darum - ... - die sämtlichen Weltkulturen gemeinsamen Werte sichtbar zu machen. Dazu gehört auch das unzweideutige Eintreten für die Menschenrechte ... [GO_SPEECH]

(20) On the other hand, we have to strengthen dialogue with the Islamic world. This is a matter of visualizing the values shared by all cultures of the world. This also includes defending human rights unambiguously. [ETRANS_SPEECH]

Fewer constructions are available in English, and they are furthermore instantiated less often. Hence, English texts seem to be less marked and more neutral in terms of taking up referents in the textual world. Instead personal pronouns or modifiers seem to be employed more often to establish identity between animate referents and definite articles (in combination with lexical cohesion) seem to be used more often for inanimate referents.

(21) When we left my mother said, “Of what? What does he have to be careful about? They put the tray out so people can look at the things, don’t they? So what does he have to be careful about?”[EO_FICTION]

(22) The superb river poised in such elegance and folded to roar down its fight full throat of rapid, they will try crossing here because it is so narrow, they are certain they could throw a stone over it if they still had their usual strength. [EO_FICTION]

In (16) personal pronouns are employed where the demonstrative pronouns der and die could be used in German. In (17), the meaning of the neuter pronoun it, which refers to the superb river, could be realized with the pronominal adverb darüber in German.
Furthermore, German seems to prefer more explicit devices to mark conjunctive relations than English since adverbials are less reduced semantically than coordinating conjunctions. This is illustrated in (18) and (19). Addition is more often expressed in German by a conjunctive adverbial, while English more often employs the simple coordinator *and*:

(23) **Aufgrund der veränderten Sicherheitslage konnte die Mannschaftsstärke um 40 Prozent reduziert werden. Außerdem wurden, ...., knapp 11000 ehemalige Soldaten der Nationalen Volksarmee der DDR in die nun gesamtdeutsche Bundeswehr integriert.** [GO_ESSAY]

(24) **The new security situation made it possible to reduce personnel by 40 %. Furthermore, almost 11,000 former soldiers from the GDR's National People's Army (NVA), excluding higher ranking officers, were integrated into the new all-German Bundeswehr.** [ETRANS_ESSAY]

(25) **Mr. Bush has time and again demonstrated his commitment to open trade. And he is determined to extend the benefits of open markets to the world's poorer nations.** [EO_ESSAY]

The y-axis represents the separation between spoken and written registers in both languages: we have the constellation of EO_ACADEMIC, EO_INTERVIEW, GO_ACADEMIC and GO_INTERVIEW on the one side, which are opposite to the other registers in both languages. Further registers, e.g. EO_INSTR or EO_SPEECH can be seen on the graph but are poorly represented in this dimension. Demonstrative heads and conjunctive devices expressing additive and adversative relations contribute to the commonality between English and German spoken registers, whereas comparative particular as illustrated in (20) for German and (21) for English distinguishes the written registers from the spoken registers.

(26) **Wenn sich in der Folge trotzdem ein Sinn zeigte, dann auf eine viel kompliziertere und fragwürdigere Weise.** [GO_POPSCI]

(27) **If subsequently a meaning nevertheless appeared, then in a much more complicated and dubious way.** [ETRANS_POPSCI]

(28) **Today interest rates which were 4 per cent above those of the euro area are now 1.75 per cent higher.** [EO_ESSAY]
In addition, we note that devices of substitution play a role for the distinction of spoken and written registers both in English and German although their general distributions as observed in 5.1 above are rather low in relation to devices of co-reference and conjunction. Yet, spoken English is more characterized by the use of verbal (and to a lesser degree nominal substitution) as shown in (22), whereas spoken German clearly favors clausal substitution, as illustrated with cataphoric sowas in (23).

(29) Yes, I like this little figure, yes, I definitely do. Some people describe, I mean, to me is like a little puppet version of myself. OK, OK. And you do, you get quite attached. [EO_INTERVIEW]

(30) Ehm was machen Sie damit? - Ja zum Beispiel sowas: Wir nehmen einen ganz einfachen Prozess, das Proz - dieser Prozess ist ein Streichholz, …[GO_INTEVRIEW]

(31) What do you do with it? – Well, for instance, something like this: We take a simple process, the proc (truncated word) – the process is a quick match …

The commonalities observed across languages in the spoken registers point to preferences for realizing particular semantic meaning relations, as already discussed above. They have a tendency towards using cohesive items with an interpersonal or rhetorical function. In addition higher frequencies for cohesive devices of substitution and general comparative reference seem to point to a lower degree of explicitness of cohesive devices, which may generally create more vagueness in terms of the relations expressed by cohesion in spoken as compared to written registers. Mode of production therefore seems to be the variable along which registers can be differentiated cross-linguistically.

However, the difference between the languages is greater than that between spoken and written modes. We comprehend it from the eigenvalues of the two dimensions represented in the graph: the first dimension contributes ca. 38%, whereas the second contributes around 26%.

Figure 4 additionally shows that registers in English cluster more densely than German registers and also that the distance between spoken and written registers is more pronounced in German than in English. These observations therefore seem to support earlier assumptions about contrastive tendencies in lexicogrammar by Mair (2006) and Leech et al.
(2009), and reflect a higher degree of variation between registers in general and written and spoken registers, in German than English.

5. Summary and conclusions
The study presented above has been concerned with the corpus-based analysis of cohesion in different registers of English and German. Our aim has been to identify contrasts and commonalities in the registers under investigation with respect to types and subtypes of co-reference, substitution and conjunction.

For this purpose, we have interpreted corpus data which was evaluated by different statistical methods. Descriptive methods have been employed to observe types and subtypes of cohesive devices in registers within and across languages. The results of our correspondence analyses show that we can observe additional variables, which include combinations of subcorpora under analysis: mode (spoken vs. written), language (German vs. English), registers (FICTION vs. others); and if we observe languages separately, other dimensions come into the fore, e.g. monologic vs. dialogic texts in English and narration vs. other speaker goals in German.

Concerning our main research question formulated at the outset of this study, we can generally state that contrasts are more pronounced between the two languages English and German than between registers. The main differences are attested in terms of the preferred meaning relations: a preference for explicitly realizing logico-semantic relations by cohesive devices of conjunction and tendency towards realizing more relations of identity by cohesive devices of referents. In addition, different subtypes are preferred in the two languages for realizing similar meaning relations.

Yet – and this observation targets our second question - mode of production plays an essential role for the grouping of particular registers in the two languages separately and also across languages. The spoken registers of ACADEMIC and INTERVIEW stand out in their preference for highlighting referents and events, comparing them and evaluating them by means of cohesive relations. Their lexicogrammatical realizations via cohesive devices, however, are partially language-specific.
Furthermore, the register of FICTION seems to be marked by distinct cohesive features which reflect peculiar contextual configurations in both languages. Again there are language-specific preferences which reflect distinctions on a more semantic or pragmatic level: dialogicity in English, and narration in German. These observations suggest a need for further analysis, in which we should take into account the representation of particular registers in the dimensions. For example, we should exclude POPSCI and INSTR when analyzing registers within a language, and study them separately.

Our analyses have also shown that the two languages differ as to the degree of variation between individual registers. We can find more variation in the realization of cohesive devices in German than English. In order to obtain further empirical evidence of this tendency, we will have to combine insights about cohesive devices with studies of the elements they tie up with, and the cohesive relation as such. This will yield empirical evidence in terms of distance between elements in cohesive chains, chain size and chain length. In addition, our quantitative analyses have to be accompanied by qualitative analyses. They will allow us to investigate the structural environments of the cohesive relations as well as the specific functions and meaning relations expressed.

References
Discourse variation across registers


Discourse variation across registers


Annotating Cohesion for Multilingual Analysis

Ekaterina Lapshinova-Koltunski*, Kerstin Anna Kunz**

*Saarland University
Universität Campus A2.2,
66123 Saarbrücken
e.lapshinova@mx.uni-saarland.de

**University of Heidelberg
Grabengasse 1, 69117 Heidelberg
kerstin.kunz@ied.uni-heidelberg.de

Abstract

This paper describes a set of procedures used to semi-automatically annotate a multilingual corpus on the level of cohesion, an important linguistic component of effectively organised and meaningful discourse. The annotation categories we operate with base on different degrees of granularity and account for lexico-grammatical and semantic aspects of different types of cohesion. This annotation scheme allows us to compare and differentiate cohesive features across languages, text types and in written and spoken discourse on different levels of abstraction. Our aim is to obtain a fine-grained and highly precise annotation, at the same time avoiding purely manual annotation. Therefore, we decide for corpus-based semi-automatic procedures to identify candidates expressing cohesion in English and in German. The annotated corpus is one of the few existing resources supporting contrastive studies of cohesion.

Keywords: cohesion, discourse relations, annotation, corpora

1. Introduction

Cohesion is an important component of effectively organised and meaningful discourse, as the message being communicated in discourse is not just a set of clauses, but forms a unified, coherent whole. While coherence concerns the cognitive aspects of establishing meaning relations during text processing, cohesion involves explicit linguistic means that signal how clauses and sentences are linked together to function as a whole. Both concepts have been studied in a range of disciplines, including philology, sociology, philosophy, psychology, computer science and linguistics. The latter analyses inventories of the linguistic markers that are available in a given language, see (Louwerse and Graesser, 2005). Classifications of lexico-grammatical markers and their relational potentials are quite often language specific, cf. (Halliday and Hasan, 1976; De Beaugrande and Dressler, 1981; Brinker, 2005), etc. For multilingual analysis, e.g. contrastive linguistics or translation (both human and machine) studies, it is important to establish categories which enable the comparison of inventories across languages in order to identify commonalities and contrasts. Complex annotations on higher linguistic levels which are geared towards high precision are typically carried out manually and hence, are very time-consuming. To our knowledge, existing resources provide annotations of individual cohesive phenomena only, e.g. pronominal coreference in the BBN Pronoun Coreference and Entity Type Corpus, (Weischedel and Brunstein, 2005), verbal phrase ellipsis in (Bos and Spender, 2011) or conjunctive relations in PDTB, (Prasad et al., 2008), annotation of (abstract) anaphora in (Dipper and Zinsmeister, 2009) and (Dipper et al., 2012). Most of them are monolingual and apply manual annotation procedures only.

In the present paper, we suggest procedures to semi-automatically identify and annotate cohesive phenomena.
Lexical cohesion includes similarity between referents/entities of the same type which bases on sense relations between lexical items (e.g. hypernymy, part-whole relations), as in example (5).

(1) a. Wir arbeiten für Wohlstand und Chancen, weil das richtig ist. Wir tun damit das Richtige.
b. We work for prosperity and opportunity because they’re right. It’s the right thing to do.

(2) a. Das war ein Problem. Aber keins, mit dem ich mich auseinandersetzen wollte.
b. This was a problem. But not one I chose to deal with.

(3) a. Who says that? – My parents ☺.

(4) a. Sie wollen ein starkes Europa in der Welt. Deshalb hat Großbritannien eine europäische Sicherheitspolitik mit auf den Weg gebracht.
b. They want Europe to be strong in the world. That’s why Britain has helped launch a European security policy.

(5) a. Vor allem müssen die Entwicklungsbanken ihre Bestrebungen auf... konzentrieren. Als erster sollten die Banken mehr Ressourcen für die Entwicklung von Humankapital aufwenden.
b. First and foremost, the development banks must focus their efforts... To start, the banks should devote more resources to the development of human capital.

According to Halliday and Hasan (1976), what distinguishes cohesive relations from other semantic relations is that the lexico-grammatical resources, i.e. the cohesive devices, trigger relations that transcend the boundaries of the clause.

We argue that these semantic relations are realised in both languages under investigation and also across text types including written and spoken discourse. Systemic comparisons, e.g. (Kunz and Steiner, 2012; Kunz and Steiner, 2013; Kunz and Lapshinova-Koltunski, in press), have shown that they differ in terms of lexico-grammatical patterns of realisation, as can be see from the examples above. In addition, we suggest textual contrasts in the frequency of cohesive devices, in types of preferred cohesive relations, in the strength of the cohesive relation, as well as in the breadth of variation.

Starting from these considerations, we formulate subcategories of the five phenomena of cohesion defined above, reflecting the lexico-grammatical and semantic features of the cohesive devices that establish these types. Only those categories are defined which are applicable for both English and German, see table 1.

Our analysis also includes cohesive relations, which are often described as relations across grammatical domains, e.g. in (Halliday and Hasan, 1976; Eckert and Strube, 2000; Ariel, 2001; Kunz, 2009). Here, we define two categories: coreference and lexical chains. They both involve a textual relation that is created between linguistic expressions.

<table>
<thead>
<tr>
<th>reference</th>
<th>type</th>
<th>function</th>
</tr>
</thead>
<tbody>
<tr>
<td>personal</td>
<td>head, modifier,</td>
<td>it-endophoric</td>
</tr>
<tr>
<td>demonstrative</td>
<td>head, modifier,</td>
<td>local, temporal, pronomi-</td>
</tr>
<tr>
<td></td>
<td>local</td>
<td>nal adverb</td>
</tr>
<tr>
<td>comparative</td>
<td>particular,</td>
<td>general</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>conjunction</th>
<th>synt.type</th>
<th>sem.type</th>
</tr>
</thead>
<tbody>
<tr>
<td>connects</td>
<td>additive,</td>
<td>adverb</td>
</tr>
<tr>
<td>subjuncts</td>
<td>adversative,</td>
<td>causal, temporal,</td>
</tr>
<tr>
<td>adverbials</td>
<td>modal</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>substitution</th>
<th>nominal, verbal, clausal</th>
</tr>
</thead>
</table>

| ellipsis   | nominal, verbal, clausal |

| lexical cohesion | general nouns, collocations |

Table 1: Cohesive devices and their functions

The textual relation of coreference evokes a conceptual relation of identity between discourse referents/entities (see above). A coreference relation links at least two coreferring expressions: an antecedent, i.e. a linguistic element introducing a new discourse referent, and a cohesive device of reference which functions as an anaphora (or cataphora, in the case of forward reference) and which points to the same referent again. The cohesive device of reference serves as a linguistic marker which triggers a search instruction to its antecedent e.g. a semantically weak pro-form or a deictic element. We include all categories defined for reference (see table 1) for the analysis of anaphoraphs. As several anaphoras may point to the same antecedent, several textual relations may be created for one referent in the same discourse, hence a coreference chain is the set of all coreferring expressions which refer to the same antecedent. The same applies to lexical cohesion, although the meaning relation established is a different one (see above): a lexical chain contains at least two lexical expressions in different textual parts which are linked by a semantic relation of hypernymy (e.g. a specific noun linked to a general noun), meronymy, synonymy, etc. or by repetition of the lexical base. Again, the chain may contain more elements and hence also semantic relations, which tie textual referents that belong to the same experiential or semantic domain.

3. Corpus resources

The multilingual corpus we annotate offers a continuum of different text types (registers) from written to spoken discourse. More precisely, it includes English and German texts of ten registers, eight of which represent written discourse and include fictional texts (FICTION), political essays (ESSAY), instruction manuals (INSTR), popular-scientific texts (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), tourism leaflets (TOU) and corporate websites (WEB). This part was imported from the existing corpus CroCo described in
(Hansen-Schirra et al., 2013). The written texts are saved in two subcorpora according to the language: English written texts (EO), German written texts (GO). The other registers are of spoken discourse and include recorded and transcribed interviews, as well as academic speeches, see (Lapshinova-Koltunski et al., 2012). The spoken texts are also stored in two subcorpora classified according to the language of origin: English spoken texts (EO-SPOKEN) and German spoken texts (GO-SPOKEN). The whole number of words contained in the corpus comprise ca. 730 thousand words (see table 2, although not big, but still provides a usefull data set for annotating and analysing cohesion in both languages.

<table>
<thead>
<tr>
<th>register</th>
<th>EO</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACADEMIC</td>
<td>40443</td>
<td>40986</td>
</tr>
<tr>
<td>FICTION</td>
<td>36996</td>
<td>36778</td>
</tr>
<tr>
<td>ESSAY</td>
<td>34998</td>
<td>36668</td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>37901</td>
<td>40198</td>
</tr>
<tr>
<td>INSTR</td>
<td>36167</td>
<td>36880</td>
</tr>
<tr>
<td>POPSCI</td>
<td>35148</td>
<td>36178</td>
</tr>
<tr>
<td>SHARE</td>
<td>35824</td>
<td>35235</td>
</tr>
<tr>
<td>SPEECH</td>
<td>35062</td>
<td>35399</td>
</tr>
<tr>
<td>TOU</td>
<td>35907</td>
<td>36574</td>
</tr>
<tr>
<td>WEB</td>
<td>36119</td>
<td>35779</td>
</tr>
<tr>
<td>TOTAL</td>
<td>364565</td>
<td>369675</td>
</tr>
</tbody>
</table>

Table 2: Corpus constellation and size

The corpus is pre-annotated on several levels, which include information on tokens, lemmas, morpho-syntactic features (e.g. case, number, etc.), parts-of-speech, phrase chunks and their grammatical functions, as well as and sentence boundaries. The annotation of the written part was partly imported from CroCo, whereas for the spoken part, we use Stanford POS Tagger (Toutanova et al., 2003) and the Stanford Parser (Klein and Manning, 2003). The corpus is encoded in the CWB format (CWB, 2010) and can be queried with Corpus Query Processor (CQP) (Evert, 2005). These annotation levels allow us to obtain additional information on cohesive phenomena and cohesive relations, i.e. coreference: morpho-syntactic preferences of antecedents and anaphoras, their positions in a clause, the length of chains in terms of elements, diversity of types of antecedents, their parallelism with anaphoras, etc. Furthermore, these annotation levels provide the basis for the semi-automatic procedures described in the present paper.

4. Annotation of Cohesion

4.1. Categories to annotate

In the following, we provide a more detailed description of the annotation scheme based on the categories introduced in 2. above. Note that this mainly concerns the classifications of cohesive devices, which, however, builds the basis for the analysis of cohesive relations (see below). Main categories exist for the main cohesive types reference, conjunction, substitution, ellipsis and conjunction. We distinguish subtypes, which are annotated as ‘type’ or ‘func’ feature in the corpus. They reflect general structural groupings of cohesive devices that exist in both languages.

These categories, as well as their language realisations (operationalisations) are presented in table 3.

Each subcategory of reference (type) is further subclassified according to grammatical and semantic features of the cohesive device (func). Personal reference includes personal (head) and possessive (modifier) pronouns as well as their morphological variants. For this type, we also annotate reference by it/its separately (it-endophoric) due to the ambiguity of their usage in both languages. Demonstrative reference is expressed by means of demonstrative pronouns (head) and determiners (modifier), as well as their morphological variants (in German). Moreover, we include local and temporal relations of identity, which are expressed by adverbs (see table 3) as well as pronominal adverbs (pron-adv). These exist in English and German but are employed in German with a higher frequency. Comparative reference is expressed with comparative forms of adjectives, which either trigger a general relation of comparison or a more specific one (particular).

Conjunction is classified in terms of main syntactical types: coordinating conjunctions (connects), subordinating conjunctions (subjuncts) and discourse adverbials (adverbials). They may consist of one or multi-word constructions of conjunctions, e.g. that is why, etc. see table 3. For each syntactical subcategory we provide the same semantic subclassifications, according to the main logico-semantic relations that can be established by conjunctive devices. Both in English and in German, substitution is expressed by indefinite pronouns or other nominal substitutes (nominal), substituting verbs (verbal) and different adverbials, which substitute clausal constructions, such as so in English (clausal). Ellipsis can be triggered by different lexico-grammatical means in both languages, and therefore, automatic detection still remains problematic. Nevertheless groupings can be made in terms of which structural elements are mainly omitted in relation to the preceding full textual structure. Again main categories here are nominal, verbal and clausal. Substitution and ellipsis cannot be categorised in terms of other features since their language-specific features do not allow a common subclassification. For the time being, only two aspects of lexical cohesion are categorised: Textual relations that base on the use of general nouns, for which we use lists of nouns based on those described by (Dipper et al., 2012) and repetitions of lexical bases. We plan to integrate sense relations such as hyponymy and synonymy in the future.

4.2. Annotation of Cohesive Devices

Automatic procedures To annotate the categories presented in 4.1., we elaborate a set of semi-automatic procedures, which involve an iterative extraction-annotation process. We use a method derived from the system used for the YAC chunker, see (Kermes and Evert, 2002; Kermes, 2003). The system is based on the option of the CWB tools to incrementally enhance corpus annotations, as query results deliver not only concordances of the searched structures but also information on their corpus positions. The algorithm makes use of the CWB Perl-Modules to access CQP and the encoding functionality using Perl-scripts as wrapper. Additionally, Perl modules are derived from the
As our aim is to produce a corpus with highly precise information on cohesive devices in English we combine both annotations (syntactic and semantic) to exclude non-cohesive occurrences of conjunctions, e.g. in case they link phrases and not clauses, as in the sentence in example (6).

(6) Renewables 2004 will focus on renewables and aim at strengthening the political momentum.

Classification of semantic types proceeds directly within the query which includes a simple lexical search – here, we aim to identify all cohesive instances of ‘additive’ conjunctions (a closed class of lexical items the members of which we know). The same procedure (based on lexical list) is applied to identify and annotate general nouns.

Manual procedures As our aim is to produce a corpus with highly precise information on cohesive devices in En-

### Table 3: Annotated categories of cohesion

<table>
<thead>
<tr>
<th>device</th>
<th>type</th>
<th>func</th>
<th>realisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference</td>
<td>personal</td>
<td>head</td>
<td>he/er, she/sie, they/sie, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>modifier</td>
<td>her/ihr, his/sein, their/ih, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it-endophoric</td>
<td>it/es</td>
</tr>
<tr>
<td>demonstrative</td>
<td>head</td>
<td>this/dies/das, that/jenes, etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>modifier</td>
<td>this/diese(r/s), that/jene(r/s), etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>local</td>
<td>here/hier, there/da, etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>temporal</td>
<td>now/jetzt, then/dann, etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pronom</td>
<td>herewith/hiemit, dagegen, damit, etc.</td>
<td></td>
</tr>
<tr>
<td>comparative</td>
<td>particular</td>
<td>bigger/grosser, better/besser</td>
<td></td>
</tr>
<tr>
<td></td>
<td>general</td>
<td>other/andere, such/solche</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Examples of queries and tagged structures in XML

<table>
<thead>
<tr>
<th>QP query</th>
<th>example of XML tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &lt;chunk&gt; {[.,chunk_gft=&quot;adv_temp&quot;]</td>
<td>{[.,word=&quot;then&quot;]&amp;pos=&quot;rt&quot;]} {[.,func=&quot;temporal&quot;]&lt;reference type=&quot;dem&quot; func=&quot;temporal&quot;&gt;now&lt;/reference&gt;</td>
</tr>
<tr>
<td>2 [lemma=&quot;this/these/those/that &amp;pos=&quot;dd.&quot;</td>
<td>{[.,pos=&quot;j.&quot;</td>
</tr>
<tr>
<td>3 [lemma=&quot;RE(additive)&quot;] {[.,func=&quot;additive&quot;&gt;in addition&lt;/conj&gt;</td>
<td></td>
</tr>
<tr>
<td>4 [.chunk_gft=&quot;adv.*</td>
<td>pos=&quot;kous</td>
</tr>
</tbody>
</table>

In table 4, we demonstrate examples of CQP queries to extract and annotate reference. Query 1 is designed to extract textual instances of local demonstrative reference, whereas query 2 delivers occurrences of demonstrative reference with the grammatical function of a modifier. The results do not need further processing within the annotation process, as the categorisation is encoded in the query itself. The instances are annotated as XML structures with attributes ‘type’ and ‘func’, where ‘type’ is demonstrative and ‘func’ is either temporal or modifier respectively. The tags are then imported back into the corpus and saved as CQP structural attributes.

Further two queries are built to extract semantic (query 3) or syntactic (query 4) types of conjunctions. In the final step, the information gathered using CQP queries. This permits to import the information on queried data back into the corpus. In this way, our annotation rules are defined in form of CQP queries that allow regular expressions based on string, parts-of-speech, chunk and further constraints.

Each query is applied to the corpus separately. The result is a list of corpus positions indicating the start and the end, and possibly a target position marked within the query. These corpus positions can now be used to extract additional information already encoded in the corpus (e.g., part-of-speech tags, lemma information, sentence position, etc.). If needed, this information can be evaluated against lists in order to classify or exclude them. Finally, the results and possibly additional information can be encoded using the corpus positions as anchors.
lish and German, we integrate a step of manual correction into our procedures. To facilitate this, the annotated corpus (with the structures in XML format as shown in table 4 above) is imported into MMAX2, a tool for manual annotation (Müller and Strube, 2006). Texts are corrected by at least two human annotators with linguistic background. The MMAX2 visualisation allows annotators to decide whether the candidates tagged by the automatic procedures have a cohesive function and belong to the given category. We also add an option to mark the cases as ‘problematic’ or ‘non-problematic’ to trace and analyse the reasons for annotators’ hesitation in case of a low inter-annotator agreement. This combination of automatic pre-annotation with manual post-correction is less time-consuming for human annotators as annotating raw texts. Moreover, we achieve positive results in the inter-annotator agreement (see below in this section).

Correction by human annotators allows us, on the one hand, to improve both annotations and rules for automatic procedures (rule-based identification of items can be improved on the basis of human annotators’ observations), and on the other hand, to evaluate automatic procedures.

**Evaluation** Our preliminary results show that in the automatic identification of cohesive devices, we are able to achieve a good precision for English (between 76% and 98%) and slightly lower precision for German (between 69% and 73%), shown in table 5. The lower results for German are partially caused by the multi-functionality of the lexico-grammatical means expressing cohesion. In addition, higher flexibility of ordering clausal constituents in German complicates automatic disambiguation of cohesive and non-cohesive forms on the basis of syntactic rules. In terms of recall, we are also able to achieve satisfactory results for English, e.g. 80% for reference and 73% for conjunction, and lower results for German: 60% for reference and 71% for conjunctions.

![Table 5: Precision of automatic procedures](image)

We also calculate the inter-annotator agreement 1) between human annotators (HuHu) and 2) between human annotators and automatic procedures (HuAut), see table 6. The best scores are again observed for English in the agreement between humans and the automatic system. For German, the score is lower. However, the agreement between human annotators is slightly higher in the annotation of German. This can be explained, again, by the complexity of German lexico-grammatical means expressing cohesion.

Annotating procedures are especially problematic in spoken registers, where cohesive devices are much more frequent as in written registers, see figure 1 in section 4.3, below. Spoken discourse is characterised by numerous repairs, ellipses, unclear sentence breaks and therefore. Cohesive and non-cohesive instances cannot be easily disambiguated as sentences boundaries do not play a role in spoken discourse. This all poses a real challenge for both semi-automatic and manual annotation.

![Table 6: Inter-annotator agreement for reference](image)

We calculate precision and recall for automatic reference annotation in our spoken data. As seen from table 7, the overall results for English spoken registers are better than for German. Interestingly, the register-specific results differ in both languages. Whereas in English the system performs better on academic speeches (which are mostly monologic), interviews are annotated with less errors than academic speeches in German.

### 4.3. Annotation of Coreference

For the annotation of coreference, we use the output of the semi-automatic procedures described in 4.2. above and manual annotations produced by humans. To our knowledge, none of the existing automatic procedures can fit our tasks, as most of them operate with a limited set of categories. Moreover, previous works on coreference annotation for spoken discourse, have shown that the available systems can achieve around 60% for written and ca. 50% for spoken texts, see, for instance, (Amoia et al., 2012) for the analyses with Stanford CoreNLP (Lee et al., 2011).

Therefore, we decide for manual annotation of coreference chains, which includes manual identification of antecedents by human annotators, and their linking to the cohesive devices (anaphoras) which were automatically tagged by our system described in 4.2. above, and manually corrected by human annotators. Here again, we use MMAX2 to facilitate the annotations, as this tool allows visualisation of links between two or more elements. The annotated information is then encoded as an additional attribute of ‘mention’, which is automatically provided with an identification number (id). Every expression referring to the same antecedent is also assigned with the same id.

The information is saved for every text, and then imported into the corpus. The information on the chains can then be extracted with the help of these ids. The information on the type and function of the referring expression is also integrated into this new structure, see figure 1.

![Table 7: Evaluation of procedures in spoken registers](image)

In the example presented in figure 1, the items indexed with
Dr. Hales received his M-S and B-S degrees at Stanford in nineteen eighty-two. He then went on to the Mathematical Sciences Research Institute to do post-doctoral research, and then to Harvard, where he was an assistant professor for two years under the National Science Foundation Fellowship. He completed the post-doctoral research fellowship at the Institute for Advanced Study in the following year.

Figure 1: Annotated coreference chains in the corpus

'Set_1' belong to a longer chain. Four anaphoras refer to the same antecedent, which is 'Dr Hales'. Lexical chains have not yet been annotated in the corpus. However, we aim to use the annotation of general nouns, as well as repetitions of lexical bases, and integrate semantic relations with the help of available resources, e.g. WordNet, see Fellbaum (1998). These automatic annotations will then be corrected in terms of cohesiveness by human annotators.

4.4. Annotation Availability

The annotated corpus is available in XML format and can be queried with CQP. We also provide a CQP-WEB\(^2\) version which is available via CLAIN-D project.

5. Conclusion and future work

In the present paper, we have described semi-automatic corpus-based procedures to annotate cohesive types of (co-)reference, substitution, ellipsis, conjunction and lexical cohesion. These procedures allow both automatic identification of cohesive devices and their automatic annotation, which builds the basis for further annotation of semantic relations. Moreover, the integrated procedure of manual correction enables evaluation and improvement of the automatic procedures. Furthermore, they provide a possibility of consistent annotation on the basis of the pre-defined rules, which cannot be ensured if the entire annotation is of manual character.

Our procedures concern two Germanic languages only, which have many common or comparable categories. Therefore, it would be interesting to test the proposed approach on another language pair including languages that belong to different language families. However, this is beyond the scope of the present research project.

The enriched corpus facilitates analysis of German vs. English contrasts, providing information on cohesive phenomena in both languages. Moreover, the availability of spoken material in our corpus allows the analysis of differences which result from differing conditions of speech, such as strong relation to the communication situation, direct interaction of speech participants and constraints on cognitive processing. First findings from our analyses show that mode of production plays an essential role for the grouping of particular registers in the two languages separately, and also across languages. For instance, the spoken registers in both languages exhibit a tendency towards marking important entities, comparing and evaluating them via cohesive relations. Their lexico-grammatical realisations are partially language-specific. Furthermore, we observe greater variation between written and spoken registers than in English, which may find further support in the future, when more spoken registers containing speaker turns are integrated in our corpus.

Such a resource is valuable not only for contrastive linguistics, but also for translation study, including machine translation, as well as further areas of NLP, e.g. automatic coreference resolution. The empirical data obtained from these annotations can be interpreted in terms of various linguistic aspects on different levels of granularity. It can thus be employed for further investigation and interpretation on semantic and conceptual levels of abstraction.

6. References


S. Dipper, M. Seiss, and H. Zinsmeister. 2012. The Use of Parallel and Comparable Data for Analysis of Abstract
Figure 2: Annotated corpus on CQP Web


S. Evert, 2005. The CQP Query Language Tutorial. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.


C. Müller and M. Strube. 2006. Multi-level annotation of


R. Weischedel and A. Brunstein. 2005. BBN Pronoun Coreference and Entity Type Corpus.
Conjunctive relations across languages, registers and modes: semi-automatic extraction and annotation

Abstract

In the present paper, we focus on the corpus-based analysis of conjunctions as intra- and intersentential links in texts which play an important role in text organisation. Our research goal is to explore a broad range of cohesive conjunctive relations across languages, registers and with varying mode of discourse (spoken vs. written). For this purpose, we develop semi-automatic procedures for extraction and annotation of conjunctive relations in an English-German corpus.

1. Introduction

In the present paper, we describe our corpus-based analysis of English and German conjunctive relations. Our focus is on cohesive conjunctions (further also conjunctions) as intra- and inter-sentential links in texts. We explore a broad range of conjunctive strategies across languages and registers with varying modes of discourse (spoken vs. written).

Our definition of conjunctive relations is based on Halliday / Hasan (1976), who examine conjunctive devices as explicit triggers for logico-semantic relations which establish relations of meaning between textual elements. Cohesive conjunctions thus play an
important role for text organisation and coherence. Our concept of cohesive conjunctions is an abstract one and includes different categories of items linking text passages beyond the level of the clause: connectors linking main clauses, subordinating conjunctions and conjunctive adverbials linking complex sentences. Connectors on the phrase level are excluded from our analysis, since they do not link clauses or sentences (and are thus not cohesive).

Conjunctive relations have been discussed in numerous works, though not always with a focus on cohesive aspects. Some of these approaches integrate implicit aspects of coherence, most of them however are concerned with the systemic resources to establish logico-semantic relations, below or/and beyond the level of the clause or sentence. They either look at the phenomena from a rather theoretical perspective, integrating examples from multiple languages to construe a highly abstract and rather universal model, or focus on the resources of one individual language, e.g. Halliday / Hasan (1976), Quirk et al. (1985), Pasch et al. (2003), Halliday / Matthiessen (2004), Blühdorn (2008), Mann / Thompson (1988). In addition there exist a number of contrastive studies that examine textual instantiations of (cohesive) conjunctions, also in English and German, yet they are either limited to the investigation of individual devices in particular registers or do not provide empirical evidence, e.g. Fabricius-Hansen (1999), Doherty (2006) or Becher et al. (2011). Some studies do apply corpus-based processing mechanisms for the empirical analysis of a larger set of conjunctions. For instance, Hutchinson (2005) describes automatic acquisition of knowledge about discourse connectives based on their distributions in texts and focuses on their semantic properties, and on the relationships that hold between them. Corpora are also used in works of Stede (2002), Dipper / Stede (2006), Lüngen et al. (2006) and Stede (2008a/b) who apply lexical resources (containing lexical and morpho-syntactic information) for conjunction extraction. Bestgen et al. (2006) also employs corpora to automatically determine the semantics of connectives. Another work to consider is that of Marcu (2000), who used surface-based and statistical methods to identify elementary discourse units and hypothesise coherence relations between adjacent segments.
However, these studies mainly concentrate on the description of one language only. To our knowledge, comprehensive multilingual corpus-based approaches to the analysis of cohesive conjunctive relations are still missing. Moreover, none of the existing works combine automatic extraction and annotation techniques to study conjunctions or conjunctive relations.

We suppose that the main reason for this gap is the fact that comprehensive investigations of conjunction and cohesion in general require complex analyses on different linguistic levels. Therefore, we aim at analysing a large set of cohesive conjunctions which vary across languages (English vs. German), modes (written vs. spoken) and registers (various registers available in the corpus).

The remainder of the paper is structured as follows. Section 2 presents the conceptual clarification of conjunctive relations, their difference from other cohesive devices and the classes of relations under analysis. In section 3, we describe the corpus resources and tools we are using, as well as procedures to annotate conjunctive relations. Querying and interpretation of extraction results are presented in section 4. In section 5, we draw some conclusions and provide some ideas for future work.

2. Theoretical background

As already mentioned above, our aim is to semi-automatically analyse how the resources for establishing cohesive conjunctive relations provided by English and German language systems are instantiated in naturally occurring texts of English and German. More specifically, we intend to explore contrasts between the two languages in form, frequency, function and relation across and between registers and modes. In the following subsections, we will first provide some information about our own concept of cohesion, and the peculiarities of cohesive conjunctions if compared to other cohesive devices. Then
we provide our classification of cohesive conjunctions according to different criteria.

2.1. Cohesive conjunctions vs. other cohesive devices

The present study is part of a broader contrastive research on various cohesive phenomena such as personal and demonstrative reference, substitution, ellipsis or lexical cohesion, see Kunz / Steiner (2013a, 2013b) for details.

Our concept of cohesion combines lexico-grammatical as well as textual aspects of cohesion. On the one hand, it considers contrasts in English and German in terms of the morphological, syntactical and lexical features of the cohesive devices which indicate a relation to other textual elements. On the other hand, it accounts for differences in the conceptual and semantic relation that is established on the textual level.

Cohesive conjunctions, as well as all other types of cohesive devices, establish relations of meaning between textual elements and their interpretation is dependent on the elements they tie up with. Yet, cohesive conjunctions exhibit several distinct properties on the basis of which they can be delineated from other cohesive strategies. However, cohesive conjunctions exhibit a higher degree of variation than other cohesive devices in terms of their internal structure. They range from single items that can be assigned to different parts of speech with varying morphological components (coordinating conjunctions, subordinating conjunctions, pronominal adverbs) to expressions comprising several linguistic items (prepositional phrases, non-finite and, in some cases, even finite clauses). In contrast to cohesive conjunctions, other types of cohesive devices are either modifiers of noun phrases or prepositional phrases or serve as functional/lexical heads (e.g. reference). Therefore, they are always (part of) syntactic constituents realising all kinds of functions, e.g. subject, object, adverbial, and even parts of predicates in case of substitution, ellipsis and lexical cohesion.
In contrast to the kind of relation we find with other cohesive devices, conjunctions do not trigger a search instruction to one single other textual element. In fact, conjunctions indicate a connection between two other elements in the text. Conjunctions typically realise cohesion rather locally – they are typically not employed in texts to create cohesive chains like reference and lexical cohesion. However, Martin / Rose (2003: 111) rightfully highlight their potential of sequencing. Most obvious sequences are realised with temporal relations (first ... then ... afterwards ... finally), but other logical relations may equally be signalled by “chains” of conjunctions, such as condition (if ... as a prerequisite...) or addition (and .... or ... and).

Finally, the meaning relations established between conceptual referents are primarily relational (logical). In this they differ from other cohesive strategies, which they are mainly experiential and textual in nature indicating different degrees of similarity between referents (reference: identity between individual referents substitution and ellipsis: usually identity between types of referents, lexical cohesion: similarity between types of referents, see Steiner / Kunz 2013b). By contrast, cohesive conjunctions primarily explicate logico-semantic relations between other referring expressions.

2.2. Conceptualisation and classes

There exist numerous works that address conjunctions in various areas of linguistics. The conceptualisations established as well as the phenomena examined vary as to the peculiarities of the respective languages investigated. They also result from differing research perspectives. For instance, grammar-based orientations deal with the phenomenon from below and are primarily concerned with the syntactic constraints and properties of the conjunctive items, which explicitly indicate a particular logico-semantic relation between (mostly) two other textual elements, cf. for instance, works of Pasch et al. (2003) and Blühdorn (2008). Rhetorical orientations start from above, from the level of coherence and have their foci on the kinds of logico-semantic relations between textual elements, which may be
realised explicitly by conjunctive devices or remain implicit. For instance, works in the frame of Rhetorical Structure Theory can be grouped under this orientation, cf. Mann / Thompson (1988) or Grimes (1975).

Our contrastive approach is situated between the grammar-based and the rhetorical approaches. On the one hand, it is concerned with the cohesive resources that establish logico-semantic relations between textual elements. It, therefore, intends to identify contrasts in the lexical and syntactic realisations of conjunction between English and German. In contrast to purely grammatical orientations, our concept focuses on those devices whose connectivity is cohesive rather than grammaticalized. Thus, we exclude connectors on the phrase and integrate devices that link main and subordinate clauses, clause complexes and even larger textual parts, see Halliday / Hasan (1976). On the other hand, our approach also incorporates contrasts between English and German in terms of the nature of the logico-semantic relations established, yet considering only those relations which are explicitly triggered by conjunctive devices. In this way, our research interest lies in comparing both the lexico-grammatical nature of the conjunctions as linguistic triggers but also the logico-semantic type of the relation triggered.

A comprehensive analysis of conjunctions requires their comprehensive classification. We classify conjunctions according to both grammatical and semantic (or functional) criteria. For our classification of main types of conjunctions according to grammatical criteria we follow Blühdorn (2008), since other monolingual approaches, e.g. Pasch et al. (2003) for German and Quirk et al. (1985) for English, are either too fine-grained to permit a contrastive comparison or differ largely in their level of granularity. The latter may largely be a result of varying conditions for the syntactic organisation of information extraction (e.g. less flexible constituent structure in English if we compare it to German).

Blühdorn (2008) mainly distinguishes between (1) coordinating conjunctions, which link elements in a paratactic construction (connectors in Table 1), (2) subordinating conjunctions linking elements in a hypotaxis (subjuncts) and (3) conjunctive adverbials
Conjunctive relations across languages, registers and modes

(adverbials) that indicate relations between clause complexes (sentences) or even elements on a higher textual level (see Table 1). The first two types of conjunctions are regarded in the literature both in English and German as not having a syntactic function. Indicators of this property in both languages are that coordinating conjunction cannot be preceded by another conjunction and that they always occur between the two coordinated elements. One additional criterion identifying items as proper conjunctions for German is that these items occur in the Vorfeld, followed by the first constituent of the sentence, which precedes the finite as the Vorfeld can only be occupied by one constituent. In contrast, conjunctive adverbials are considered as fully fledged syntactical constituents whose syntactical position is more flexible both in English and German as compared to the other two types of conjunctions.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>connector</td>
<td>neither ... nor, but, and, etc.</td>
<td>weder...noch, aber, doch, etc.</td>
</tr>
<tr>
<td>subjunct</td>
<td>because, even though, providing that, etc.</td>
<td>obwohl, vorausgesetzt dass, wenn, etc.</td>
</tr>
<tr>
<td>adverbial</td>
<td>also, so, for instance, by contrast, etc.</td>
<td>auch, genauso, abgesehen davon, etc.</td>
</tr>
</tbody>
</table>

Table 1. Classification of conjunctions according to their syntactic function.

As displayed in Table 1, conjunctive devices of cohesion in all of these categories may consist of a single part of speech (a conjunction or an adverb), a combination of parts of speech (e.g. prepositional phrases) or disjoint combinations of parts of speech, usually with one conjunction introducing one clause and another conjunction introducing the following clause (e.g. neither ... nor in English).

For our contrastive study of the types of relation established, all syntactic types of conjunctions are grouped with respect to the type of lexico-grammatical relation they trigger. According to Halliday / Hasan (1976), cohesive power of conjunctions does not rest in a conjunctive expression like afterwards, but in the underlying semantic
relation. Therefore “any expression of that relation, with or without a demonstrative or other reference item, will be considered to fall within the category of conjunction” (1976: 231). The cohesive function of a semantic relation, e.g. a time sequence, is called conjunction; the device that explicitly indicates this relation, e.g. afterwards, is assigned the term conjunctive or conjunctive adjunct instead. Our main semantic categories base on the classification by Halliday / Hasan (1976), but can be found in all major works on semantics (see Table 2).

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>und, weiterhin, darüberhinaus</td>
</tr>
<tr>
<td>adversative</td>
<td>aber, allerdings, dagegen</td>
</tr>
<tr>
<td>causal</td>
<td>aufgrund dessen, aus diesem Grund</td>
</tr>
<tr>
<td>temporal</td>
<td>danach, endlich, inzwischen</td>
</tr>
<tr>
<td>modal</td>
<td>sicherlich, angeblich, anscheinend</td>
</tr>
</tbody>
</table>

Table 2. Classification of conjunctions according to their function in text.

For our investigation of cohesive conjunctions, we had to integrate disambiguation mechanisms, firstly, in order to disambiguate cohesive vs. non cohesive instances, cf. examples (1) and (4), secondly, in order to assign textual occurrences in terms of the respective conjunctive category, cf. examples (3) and (6), and thirdly, to define whether the cohesive category at stake is a conjunction or has to be assigned to another cohesive category such as cohesive reference, e.g. example (2) and substitution, e.g. example (5).

(1)  Die Union plädiert dafür, dass das Betreuungsgeld so schnell wie möglich eingeführt wird (The Christian Democratic Union advocates the earliest possible introduction of child care subsidy (pleads for that child care subsidy will be introduced as soon as possible)).

(2)  Das Betreuungsgeld soll so schnell wie möglich eingeführt werden. Dafür plädiert die Union schon seit längerer Zeit (Child care subsidy should be introduced as soon as possible. The Union has been pleading for this for a long time).
Conjunctive relations across languages, registers and modes

(3) ... Immer heftiger werden dafür die Debatten um das Betreuungsgeld (The debates on child care subsidy are getting much fiercer instead).

(4) And she made no impression at all on the Strong Man, so engrossed was he in describing ... 

(5) For most people in Europe, the nation state is the center of their political allegiance and will remain so.

(6) You will be crossing some lonely mountains, so make sure you have enough petrol.

3. Resources and tools to analyse conjunctive relations

3.1. Corpus resources

To analyse cohesive conjunctive relations semi-automatically, we use GECCo1, a multilingual corpus which offers a continuum of different registers from written to spoken discourse (ca. 1,3 Mio tokens). This constellation allows capturing differences in frequency and function of different linguistic phenomena, and cohesive devices in particular, between texts of spoken and written mode and also between individual registers within this continuum. The corpus comprises four written subcorpora: GO (German Original), EO (English Original), GTrans (German Translation), ETrans (English Translation), and two spoken subcorpora: EO-SPOKEN (English Original) and GO-SPOKEN (German Original). Each written subcorpus contains a collection of texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and websites (WEB). The spoken subcorpus includes two registers: texts from interviews on private and professional life (INTERVIEW), academic speeches (ACADEMIC) cf. Table 3.

For INTERVIEW we use parts of the already existing speech corpora ELISA, cf. Braun (2006) and BACKBONE, cf. Kohn (2011). The English part of ACADEMIC contains a subset of MICASE, cf. Simpson et al. (2002) and includes data from lectures and seminars at the University of Michigan. For the German ACADEMIC subcorpus, we transcribed recordings of lectures from different departments of Saarland University (collected by VISU, the Virtual University of Saarland2 in order to obtain comparability with the English ACADEMIC subcorpus. In order to guarantee consistency, each data sample was annotated by two student assistants. INTERVIEW and ACADEMIC reflect a medium, ACADEMIC being highly monologic, and INTERVIEW being monologic with some dialogic passages. For a diversification of registers, we currently expand the corpus to more registers of spoken language, such as transcripts from talk shows and everyday conversation as well as registers in between written and spoken language, such as internet forums and chats. We do not include translations in these subcorpora as they are not professionally produced for these registers.

<table>
<thead>
<tr>
<th>Subcorpora</th>
<th>Registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO-&gt;GTRANS, GO-&gt;ETRANS</td>
<td>FICTION, ESSAY, INSTR, POPSCI, SHARE, SPEECH, TOU, WEB</td>
</tr>
<tr>
<td>EO/GO-SPOKEN</td>
<td>INTERVIEW, ACADEMIC</td>
</tr>
</tbody>
</table>

Table 3. GECCo subcorpora and registers.

2 <http://www.uni-saarland.de/info/universitaet/zentrale-einrichtungen/virtuellesaaruniversitaet.html>
The annotations available in our corpus have a complex structure, which is required for semi-automatic processing and exploitation of cohesive phenomena. GECCo is annotated on the level of 1) tokens (with information on words, lemmas, morpho-syntactic features and parts-of-speech (POS)); 2) chunks (with information on syntactic and semantic features of phrases and on sentence boundaries); 3) texts (with information on register and text boundaries); 4) extralinguistic features (with information on register analysis, language variation, speaker information, etc.) as shown in Figure 1.

The written parallel subcorpora are aligned on the sentence level (EO with GTRANS and GO with ETRANS). All subcorpora (written and spoken) are enriched with meta-information about experiential domain, goal orientation (i.e. textual function), social hierarchy, type of interaction, etc. The layer of extralinguistic information is especially complex for the spoken subcorpora since we include such specific features as speaker turn, gender and age. This information is important for certain corpus queries as shown in section 4 below.

Figure 1. GECCo structure and annotation layers.
3.2. Annotation and extraction tools

The annotations described above allow us to establish another level of linguistic annotation on top, on the level of cohesion. For these annotations, we have formulated complex semi-automatic extraction and annotation procedures on the basis of the annotations available and with the help of the corpus query processor (CQP), cf. (Evert, 2005).

We formulate queries in form of regular expressions with CQP which allows two types of attributes: positional (e.g. for part-of-speech and morphological features) and structural (e.g. for chunks, registers or extralinguistic information). The attributes are employed for CQP-based cascaded queries which include string, parts-of-speech, chunk, register and further constraints. Furthermore, the corpus architecture also enables the definition of multilayer queries in CQP, yielding data with respect to other linguistic phenomena. The data extracted with CQP deliver the following information:

- lists of textual instances of a particular pattern (with their frequencies), which can then be used for quantitative analysis and validation, e.g. in the R environment, cf. Venables / Smith (2010);
- lists of corpus positions of the extracted instances which can be used for automatic annotation.

CQP allows us to incrementally improve corpus annotations using the information on corpus positions of the extracted items. In this way, we can import the information on extracted and validated data back into the corpus. For this, we use the annotation procedures described for the YAC recursive chunker, cf. Kermes (2003), which makes use of the CWB Perl-Modules to access CQP and the encoding functionality using Perl-scripts as wrapper. The annotation of information extracted by CQP is possible with the Perl modules developed within the framework of YAC, cf. Kermes / Evert (2001, 2002).
3.3. Annotation procedures

Using the functionality of the tools described in section 3.2 above, we develop semi-automatic annotation procedures to identify and analyse cohesive conjunctions as presented in 2.1.

We therefore, expand our corpus with tags which include information, first, as to the syntactic type (conj_type) of conjunction (connector, subjunct, adverbial) and second, as to the type of logico-semantic relation (conj_func) established (additive, adversative, temporal, causal, modal). For this, we elaborate complex CQP queries which include lexico-grammatical restrictions possible due to combinations of the relevant annotation levels. This enables us to identify linguistic items that have the potential to belong to the two types of categorisations presented in Table 1 and Table 2 above.

Each query is applied to the corpus separately. The result is a list of corpus positions indicating the start and the end, and possibly a target position marked within the query. These positions are used as anchors to encode the obtained information into the corpus.

In the first step, we design queries to extract syntactic types of conjunctions. These queries contain both lexical and grammatical restrictions, as shown in an example for subjuncts in English in Table 4. Line 1 sets the constraints to start the search at the beginning of a clause. With line 2, we exclude clauses which have the grammatical function of a direct object from our search. The next line (3) is used to restrict the search to particular POS tags (ics - preposition conjunction, cc - coordinating conjunction, cs – subordinating conjunction, bto – in order to introducing infinitive, ddq – w-words, rrq – w-adverbs and if – preposition for) which can count up to 5 words (we limit the length of the subjunct to 5 elements, as our evaluation tests have shown that this is the maximal length of conjunctive expressions occurring in our corpus). We also exclude particular tokens such as and and to, since they do not establish the kind of conjunctive relation we are looking for with the help of this query (see line 4). The searched element can be followed by any instance within the clause, and should not function as a direct object (line 5). Line 6 indicates the end of the clause.
The results extracted with such a query are encoded using the corpus positions as anchors into GECCo in form of additional CQP tags. For instance, the conjunctions detected with the query in Table 4 are tagged as demonstrated in example (7).

<table>
<thead>
<tr>
<th>Query building elements</th>
<th>Explanation</th>
<th>Extracted examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 <code>&lt;chunk_type=&quot;clause&quot;&gt;</code></td>
<td>find clause start</td>
<td></td>
</tr>
<tr>
<td>2 <code>&lt;chunk_gf!=&quot;dobj&quot;&gt;</code></td>
<td>exclude direct object clauses</td>
<td></td>
</tr>
<tr>
<td>3 `[pos=&quot;ics</td>
<td>cc</td>
<td>cs.*</td>
</tr>
<tr>
<td>4 <code>_chunk_gf!=&quot;dobj&quot;&gt;</code></td>
<td>exclude and or to</td>
<td></td>
</tr>
<tr>
<td>5 <code>&lt;chunk&gt;</code></td>
<td>find clause end</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Query to extract subjuncts in English.

(7) `<conj_type="subjunct">as</conj> it moves on to this next stage of its development

The newly encoded information on syntactic types can be analysed per se, and can also be additionally applied in the second step of our procedures, cf. Table 5.

<table>
<thead>
<tr>
<th>Query building elements</th>
<th>Explanation</th>
<th>Extracted examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 <code>&lt;conj_type=&quot;subjunct&quot;&gt;</code></td>
<td>start of a conjunction restricted to a subjunct</td>
<td></td>
</tr>
<tr>
<td>2 <code>[lemma=&quot;though\although\ despite\etc...]</code></td>
<td>a list of adversative conjunctions known from grammar</td>
<td>Although</td>
</tr>
<tr>
<td>3 <code>&lt;/conj_type&gt;</code></td>
<td>end of a conjunction restricted to a subjunct</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Query to extract adversative conjunctions in English.
We include the annotated structures (line 1 and line 3) together with lexical lists (line 2) into the queries to extract additive, adversative, causal, temporal and modal conjunctive relations from the corpus.

The extracted results are also encoded into the corpus, in the same way as in the first case, see examples (8) - (11).

(8) `<conj_func additive>And</conj_func> with refinery capacity growth constrained by regulations and economics, refined products are projected to represent a growing share of these imports, reaching an estimated 20 percent of total net oil imports by 2025.

(9) conj_func="adversative">Although</conj_func> the technical hurdles of fusion energy are immense, the promise of this technology is simply too great to ignore.

(10) `<conj_func causal>As a result</conj_func>, U.S. energy policy plays an influential role in maintaining global energy security.

(11) `<conj_func temporal>Finally</conj_func>, the declaration reaffirms the importance of intellectual property protection and the important role it plays in the development of new medicines.

As our aim is to provide highly precise information on cohesive conjunctions in English and German, we integrate a step of manual correction into our procedures. To facilitate the manual correction, the annotated corpus is imported into MMAX2, a tool for manual annotation, cf. Müller / Strube (2006). Corrections by human annotators are done on the basis of elaborate annotations guidelines. They allow us, on the one hand, to check the output of automatic procedures and to correct them, whenever necessary, and to evaluate them in terms of precision and recall as well as ambiguity of the cohesive relation, on the other hand. Moreover, we are also able to improve the procedures themselves: The rules for extraction are enhanced on the basis of human annotators' observations.

Our preliminary results show that in the automatic identification of conjunctive relations, we are able to achieve the average precision of ca. 76% for English originals and of ca. 69% for the German ones. We were also able to calculate recall, that respectively estimates ca.
73% and ca. 71% for English and German texts. Interestingly, the results vary across registers. We achieve the highest precision and recall in English political essays (ca. 87% and ca. 86% respectively), whereas for German this register turns out to be more problematic in both precision and recall (56% and 62%). In terms of identified syntactic and semantic types, we can also observe differences between languages. In English, most 'noise' cases (contributing to lower precision) occur in the identification of additive connects, whereas most 'silent cases' are caused by causal subjunctions. For instance, and in (12) was automatically tagged as an additive connector, although it links phrases and not clauses. This error occurs due to the erroneous annotations on the chunk level, as we employ this information in our macros.

(12) IPHE will address the technological, financial, and institutional barriers to hydrogen <conj_type connect><conj_func develop internationally recognized technology standards to speed market penetration of new technologies.

The evaluation results for German show that all semantic classes pose a similar challenge for the detection procedure – they all equally contribute to the precision and the recall. However, the situation is different in the case of syntactic classes – here, in most cases, it is difficult to detect and correctly classify adverbials, which could be caused by ambiguity of German adverbials. For instance, pronominal adverbs dafür, dadurch, etc. can also express demonstrative reference (in this case they refer to an antecedent, another element in a text) as in (13). They can also operate as non-cohesive elements, e.g. in (14), in which darauf operates as a Korrelat (a 'place holder' for a grammatical element in German).

(13) Die EU hat entschieden, ihre Märkte für alle Produkte der am wenigsten entwickelten Länder zu öffnen: Was erwartet Europa dafür im Gegenzug? (The European Union has decided to open its markets for all products from the least developed countries. What does Europe expect in return (for this)?)

(14) Achten Sie darauf, dass die Lüftungsschlitze nicht verdeckt werden (Avoid (be mindful of) covering the ventilation slots).
The word *so* in both German and English can function not only as a cohesive conjunction as in (15) and (16), but also as clausal substitution, like in (17) and (18). However, in German, we observe more ambiguous cases, e.g. (19), resolution of which requires more context information, and thus is not possible in fully automatic annotation.

(15) So the capacity of DNA to make accurate copies of itself and to produce proteins via mRNA is very much dependent upon a highly organized context, the living cell.

(16) Breiten Raum auf der Agenda von Johannesburg nimmt eine Bestandsaufnahme ein: *So* soll bilanziert werden, was in der vergangenen Dekade hinsichtlich der Rio-Ziele erreicht (oder versäumt) wurde (The Johannesburg agenda devotes plenty of time to reviewing progress of the Rio aims to date; *(for instance)* assessing what has or has not been achieved over the past decade).

(17) Die meisten der neuen Arbeitsplätze entstehen in kleineren Firmen mit weniger als zehn Beschäftigten, vor allem in den Dienstleistungsbranchen. *So* lief es in den Zeiten, als es der Konjunktur in Deutschland gut ging (Most of the new jobs are created in smaller firms with less than 10 employees and primarily in service industries. That was *how things worked* *(so)* when Germany’s economy was still going strong).

(18) Whether on tax harmonisation, choice of Commission President or majority voting on Foreign Policy, we have the support of other Member States. For most people in Europe, the nation state is the centre of their political allegiance and will remain *so*.

(19) *So* erwarb zuletzt der US-Konzern Abbott Laboratories die Pharma-Aktivitäten der BASF *(The US group Abbott Laboratories *(for instance)* recently acquired BASF’s pharmaceutical operations vs. *In this way* the group Abbott Laboratories recently acquired BASF’s pharmaceutical operations)*.

4. Querying and analysing conjunctive relations

Having this information annotated in the corpus, we can now easily extract conjunctive relations formulated as abstracted classes with a
CQP query. We have several options for querying different aspects of cohesive conjunctions – on a general level but also on the more fine-grained levels described above.

For instance, concordances and general frequencies of all items expressing conjunctive relations can be queried by means of the following query: 

```
/region[conj].
```

Figure 2 illustrates the overall distribution of conjunction in the GECCo subcorpora. Our findings from this query already reveal for the written registers of the GECCo corpus that more cohesive conjunctions are traced in the German subcorpora than in the English subcorpora. This is illustrated in Figure 2, which displays the distribution in percentage for conjunctions in relation to all tokens in the respective subcorpora. These findings may result from a tendency to explicate logico-semantic relations by cohesive items in German which may rather stay implicit in English.

![Figure 2. Distribution of conjunctions in the written GECCo subcorpora.](image)

To gain information on the distribution of conjunctive relations in the whole corpus, we can now apply sorting and grouping options of CQP. For example, with the commands `group Last match conj_type` or `group Last match conj_func` we can obtain information on the distribution of the syntactic and semantic types.

We are also able to restrict our extractions to a certain type of relations only. For instance, in order to extract all occurrences of adversative relations we run the following query:

```
<conj_func="adversative"> [ ]* </conj>.
```
Conjunctive relations across languages, registers and modes

The output delivered by CQP is a list of concordances containing adversative conjunctive relations, cf. Figure 3:

| 395: <Although> most of the United States’ natural gas can be su |
| 444: <still> be highly dependent on energy imports to meet fut |
| 735: <But> the United States also recognizes that it must ta |
| 3991: <Nevertheless> it does mean that we have to explain even more cl |
| 10903: <in contrast>, is the fruit of human freedom. Free trade can |

Figure 3. Output of the CQP query for adversative conjunctions.

Using further options available for CQP, we can now sort results according to their occurrences across registers, as shown in Table 6.

One can observe certain tendencies in the distribution of additive vs. adversative relations having this dataset: all registers utilise more additive relations than adversative, and additionally, POPSCI and FICTION seem to use more conjunctive relations in general. This may be due to the conceptual strategy of construing logico-semantic relations in English, or point to a greater variation in meaning and functionality of additive than adversative forms.

<table>
<thead>
<tr>
<th>Register</th>
<th>Relation type</th>
<th>Freq. abs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTR</td>
<td>additive</td>
<td>745</td>
</tr>
<tr>
<td>FICTION</td>
<td>additive</td>
<td>556</td>
</tr>
<tr>
<td>WEB</td>
<td>additive</td>
<td>547</td>
</tr>
<tr>
<td>POPSCI</td>
<td>additive</td>
<td>453</td>
</tr>
<tr>
<td>ESSAY</td>
<td>additive</td>
<td>407</td>
</tr>
<tr>
<td>SPEECH</td>
<td>additive</td>
<td>386</td>
</tr>
<tr>
<td>POPSCI</td>
<td>adversative</td>
<td>373</td>
</tr>
<tr>
<td>FICTION</td>
<td>adversative</td>
<td>350</td>
</tr>
</tbody>
</table>

Table 6. Results for additive and adversative conjunctions and counted by CQP according to written registers in English.

We can further validate these data statistically, for instance, with the help of the R environment mentioned in section 3.2 above. We
perform the Pearson's chi-square test to compare English and German originals in terms of types of conjunctive relations. If the estimated p-value is < 0.05, the difference in the distribution of conjunctive relations is significant. In Table 7, we exemplify our analysis for different combinations of conjunctive relations.

<table>
<thead>
<tr>
<th>relations</th>
<th>p-value</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive vs. adversative</td>
<td>&lt; 2.2e-16</td>
<td>yes</td>
</tr>
<tr>
<td>additive vs. causal</td>
<td>0.09405</td>
<td>no</td>
</tr>
<tr>
<td>causal vs. temporal</td>
<td>0.006506</td>
<td>yes</td>
</tr>
<tr>
<td>all relations</td>
<td>&lt; 2.2e-16</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 7. P-values for EO vs. GO in terms of conjunctive relations.

In terms of all types of conjunctive relations, English and German originals show significant contrasts. However, the situation is different if we take into account the distribution of certain types only. Thus, EO and GO do not differ if we compare the distribution of additive and causal relations.

The presence of further annotation levels in our corpus allows us to construct queries for extraction of relations for more specific contexts. For example, if we want to compare the usage of connectors with the additive function in certain registers, e.g. in FICTION vs. POPSCI, we can apply query 1 demonstrated in Table 8.

<table>
<thead>
<tr>
<th>CQP queries</th>
<th>Explanation for restrictions and commands</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  [._conj_type=&quot;connect&quot;&amp;_conj_func=&quot;additive&quot;]</td>
<td>connector</td>
</tr>
<tr>
<td>2  group Last match text_register</td>
<td>group the frequencies according to registers</td>
</tr>
<tr>
<td>3  group Last match lemma by match text_register</td>
<td>group the frequencies of lemmas according to registers</td>
</tr>
</tbody>
</table>

Table 8. Query to extract additive connectors from FICTION and POPSCI.
This query is designed to extract the additive connectors that occur either in fictional texts or in popular-scientific articles. This proves to be extremely useful not only for the contrastive analysis of textual phenomena but also for language learning and teaching purposes, translator training or social linguistics, as the conjunctive relations are easily accessed by categories and patterns one is familiar with. The information on the distribution of this category and certain lexical items within this category can be obtained by queries 2 and 3 in Table 8. The output of queries 2 and 3 is presented in Table 9:

<table>
<thead>
<tr>
<th>Lemma</th>
<th>FICTION</th>
<th>POPSCI</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>and</td>
<td>462</td>
<td>201</td>
<td>663</td>
</tr>
<tr>
<td>or</td>
<td>27</td>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td>nor</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>either</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>neither</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>TOTAL</td>
<td>503</td>
<td>219</td>
<td>722</td>
</tr>
</tbody>
</table>

Table 9. Query results for additive connectors from FICTION and POPSCI in English.

Table 9 displays the distributions of various forms of connectors in the two registers FICTION and POPSCI in English. Although the frequencies still have to undergo statistical evaluation, we already can state that FICTION exhibits a clear preference for connectors if compared to POPSCI. Closer investigation reveals a linguistic manifestation of several influencing factors: A combination of narration and dialogic passages in FICTION causes less variation and more oral style, whereas a tendency for exposition and argumentation in POPSCI engenders a more phrasal organization of information together with a usage of more specific/explicit conjunctions (e.g. in form of conjunctive adverbials).
Thus, the extraction and annotation procedures described above facilitate our analysis of contrasts for cohesive relations in different languages, registers and modes. The findings gained with our corpus-based analyses will complement our observations about contrasts in the systemic resources in English and German and will enable the verification of hypotheses as to their instantiations in texts of the two languages. These hypotheses about contrasts in cohesive conjunctions between English and German cover a broad range of aspects on different linguistic levels such as:

- syntactic properties of conjunctive devices: position, parts of speech, syntactic types,
- types of logico-semantic relations established (see above),
- scope of conjunctive devices, and
- textual and lexico-grammatical properties of connected elements.

Annotations of both syntactic and semantic types of conjunctions allow us to contrastively analyse preferences of specific syntactic types in terms of the logico-semantic relations established. This information can be obtained by the queries illustrated in Table 10:

<table>
<thead>
<tr>
<th>CQP queries</th>
<th>Explanation for restrictions and commands</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 /region[conj]</td>
<td>all conjunctions</td>
</tr>
<tr>
<td>2 group Last match conj_func by match conj_type</td>
<td>syntactic and semantic types</td>
</tr>
</tbody>
</table>

Table 10. Query to extract both semantic and syntactic types.
Table 11. Preferences of syntactic types for logico-semantic relations in English and German originals.

Table 11 presents an example of the extracted results for connectors in English and German. It illustrates that both languages show a clear preference for expressing additive relations via connectors. Yet, we observe considerable contrasts between the two languages for the other two syntactic types: In English, most adverbials realise temporal relations, while in German highest distributions are measured for

<table>
<thead>
<tr>
<th>Syntactic</th>
<th>Semantic</th>
<th>EO in %</th>
<th>GO in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>connect</td>
<td>additive</td>
<td>79.57</td>
<td>74.01</td>
</tr>
<tr>
<td>connect</td>
<td>adversative</td>
<td>20.43</td>
<td>20.73</td>
</tr>
<tr>
<td>connect</td>
<td>causal</td>
<td>0.00</td>
<td>5.13</td>
</tr>
<tr>
<td>connect</td>
<td>temporal</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td>connect</td>
<td>modal</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TOTAL connect</td>
<td></td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>subjunct</td>
<td>additive</td>
<td>12.78</td>
<td>7.51</td>
</tr>
<tr>
<td>subjunct</td>
<td>adversative</td>
<td>12.04</td>
<td>33.79</td>
</tr>
<tr>
<td>subjunct</td>
<td>causal</td>
<td>41.53</td>
<td>18.34</td>
</tr>
<tr>
<td>subjunct</td>
<td>temporal</td>
<td>33.17</td>
<td>14.42</td>
</tr>
<tr>
<td>subjunct</td>
<td>modal</td>
<td>0.48</td>
<td>25.94</td>
</tr>
<tr>
<td>TOTAL subjunct</td>
<td></td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>adverbial</td>
<td>additive</td>
<td>18.44</td>
<td>19.26</td>
</tr>
<tr>
<td>adverbial</td>
<td>adversative</td>
<td>14.36</td>
<td>16.65</td>
</tr>
<tr>
<td>adverbial</td>
<td>causal</td>
<td>8.21</td>
<td>0.00</td>
</tr>
<tr>
<td>adverbial</td>
<td>temporal</td>
<td>36.12</td>
<td>28.74</td>
</tr>
<tr>
<td>adverbial</td>
<td>modal</td>
<td>22.87</td>
<td>35.36</td>
</tr>
<tr>
<td>TOTAL adverbial</td>
<td></td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
modal adverbials (*deshalb*, *deswegen*). Furthermore, in terms of subjuncts, it is English, that exhibits the highest values for causal relations (*since, as, because*), while German shows a preference for adversative subjuncts. Further qualitative analyses are required here for an interpretation of explanatory backgrounds.

In addition, we are able to identify co-occurrences, frequencies and functions of cohesive conjunctions as indicators of register or features of translations. The annotation and extractions procedures described above are applied in a similar way for corpus-based analyses of other cohesive phenomena such as reference and substitution, cf. Lapshinova / Kunz (2013). Furthermore, they provide information about co-occurrences of the different types and subtypes of cohesion and therefore prove to be a very valuable methodology for gaining an overall insight into textual realisations of a broad range of cohesive phenomena.

5. Conclusion

In the present paper, we have described automatic procedures to extract and annotate cohesive conjunctive relations which can be classified according to different criteria. These procedures allow both automatic extraction of cohesive conjunctions from corpora, as well as their automatic annotation in corpora. This methodology facilitates our corpus-based contrastive analysis of cohesive conjunctions, as we can easily access information on different linguistic levels.

Moreover, we manually improve the annotations of ambiguous cases that cannot be filtered by the automatic procedures highlighted above. The first evaluations of our procedures have shown that we are able to achieve over 70% of precision and recall. As shown in our examples, most problematic cases are caused by conjunctive adverbials, and their automatic identification is especially challenging as some forms can serve different non-cohesive and cohesive
functions. The final results will then undergo statistical validation and evaluation procedures in order to check their significance.

Future work will also include more fine-grained annotations of cohesive conjunctions, for the spoken registers in particular. First observations lead us to suggest considerable differences in frequency and function between our spoken and written registers, which result from differing conditions of speech such as strong relation to the communication situation, direct interaction of speech participants and constraints on cognitive processing.

Acknowledgements

The project GECCo: German-English Contrasts in Cohesion is supported by a grant from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). We thank our colleagues in the GECCo team – Marilisa Amoia, Katrin Menzel and Erich Steiner for their assistance. Furthermore, we are especially grateful to Hannah Kermes for providing the necessary perl script for adaptation.

References


Conjunctive relations across languages, registers and modes


Lüngen, Harald / Puskas, Csilla / Bärenfänger, Maja / Hilbert, Mirco / Lobin, Henning 2006. Discourse Segmentation of German Written Text. In Salakoski, Tapio et al. (eds) *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*. Berlin: Springer.


Intralingual Variation

In this section, we...
Chapter 5

Variation in translation: Evidence from corpora

Ekaterina Lapshinova-Koltunski

The present paper describes a corpus-based approach to study variation in translation in terms of translation features. We compare texts, which differ in the source/target texts (English vs. German), production types (original vs. translation) and method of translation (human, computer-aided = CAT, machine) in terms of a theoretically-motivated set of features. In this study, we decide for the features which can be easily obtained on the basis of automatic corpus annotations, i.e. tokens, lemmas and part-of-speech tags. Our results show that there is variation in the mentioned translations in terms of the features under analysis.

1 Introduction: Aims and Motivation

In this paper, we apply corpus-based methods to analyse translation variants – translations from English into German produced with different translation methods.

Although numerous studies on translation operate with corpus-based methods, most of them concentrate on the questions concerning the nature of translations and their specific features, (e.g. Baker 1993; 1995; Laviosa 2002; Chesterman 2004) and others. The majority of them tried to generalise translation by defining certain rules or regularities of translated texts. Moreover, they mostly compare translations with originals, i.e. differences or similarities between translations and their source texts or comparable non-translated texts, ignoring variation which can be observed in different translation variants. Corpus-based studies dedicated to the analysis of variation phenomena involving translations, (e.g. Teich 2003; Steiner 2004; Neumann 2013), etc. concentrate on the analysis of human translations only. However, nowadays, translations are produced not only
Ekaterina Lapshinova-Koltunski

by humans but also with machine translation (mt) systems. Furthermore, new
variants of translation appear due to the interaction of both, e.g. in computer-
aided translation or post-editing.

In some works on machine translation the focus lies on comparing different
translation variants, such as human vs. machine, as in (White 1994; Papineni et
al. 2002; Babych & Hartley 2004; Popovic 2011). However, they all serve the task
of automatic mt system evaluation and use the human-produced translations as
references or training material only. None of them provide an analysis of specific
linguistically motivated features of different text types translated with different
translation methods, which is the aim of the present analysis.

In this study, we aim to apply corpus-based methods to prove the knowledge of
translation features on a new dataset which contains different variants of trans-
lations, including human and machine translation.

The remainder of the paper is structured as follows. §2 presents studies we
adopt as theoretical background for the selection of features under analysis. In
§3.1, we describe the resources and methods used. In §4, we present the results
of our analyses and their discussion, and in §5, we draw some conclusions and
provide more ideas for future work.

2 Theoretical Background

Since the present study concentrates on the analysis of linguistic features of dif-
ferent translation variants, we address the existing studies on translation for their
definition.

2.1 Related Feature Work

As already mentioned in §1 above, in most cases, these studies either analyse dif-
ferences between original texts and translations (House 1997; Matthiessen 2001;
Teich 2003; Hansen 2003; Steiner 2004), or concentrate on the properties of trans-
lated texts only (Baker 1995). Nevertheless, an important point is that most of
them consider translations to have their own specific properties which distin-
guish them from the originals: both their source texts and comparable texts in
the target language. These features establish the specific language of translations
which is called translationese (Gellerstam 1986). Comparing Swedish translations
from English with Swedish original texts, the author stated significant differ-
ences between them, whereas not all of them were attributable to the source
language. This coincides with what Frawley (1984) called "third code", describ-
5 Variation in translation

ing features of translational language which are supposed to be different from both source and target languages.

Later, Mona Baker emphasised general effects of the process of translation that are independent of source language, e.g. in Baker (1993; 1995). Analysing characteristic patterns of translations, she excluded the influence of the source language on a translation altogether. Within this context, she proposed translation universals – linguistic features which typically occur in translated rather than original texts. According to Baker (1993), they are independent of the influence of the specific language pairs involved in the process of translation. Other scholars (e.g. Toury 1995 or Chesterman 2004) operate with other terms – “laws” or “regularities”. We prefer to use the term “translation features” or “phenomena” in the present study: to claim the features “universal” we would need to analyse more language pairs and translation directions, and to call them “laws” and “regularities”, we would need to test more conditions, e.g. cognitive factors, status of translation, etc., which is not possible with the bilingual dataset at hand.

Translation features can be classified according to different parameters. For instance, Chesterman (2004) makes a distinction between S-universals and T-universals: the first comprises differences between translations and their source texts, and the second covers the differences between translations and comparable non-translated texts. A more fine-grained classification includes the following features: explicitation – tendency to spell things out rather than leave them implicit, simplification – tendency to simplify the language used in translation, normalisation – a tendency to exaggerate features of the target language and to conform to its typical patterns, levelling out – individual translated texts are more alike than individual original texts, in both source and target languages, and interference – features of the source texts are observed in translations. For the second last, we prefer the term convergence proposed by Laviosa (2002), which implies a relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features, e.g. lexical density, sentence length, etc. in contrast to originals. For the last feature, we also prefer to use the term shining through defined by Teich (2003).

All these features have been widely analysed in corpus-based translation studies for different language pairs, e.g. in Laviosa (1996) for English translations from a variety of source languages, in Mauranen (2000) for English–Finnish translations, in Teich (2003) for English and German translations, and others. Yet, all of them concentrate on human translations only.

Moreover, some recent corpus-based studies applied machine learning supervised methods to automatically differentiate between translations and originals.
Ekaterina Lapshinova-Koltunski

(e.g. Baroni & Bernardini 2006). These approaches found application in some recent works on natural language processing, e.g. those on cleaning parallel corpora obtained from the Web, or improvement of translation and language models in MT (e.g. Kurokawa, Goutte & Isabelle 2009; Koppel & Ordan 2011; Lembersky, Ordan & Wintner 2012).

We employ the knowledge from these studies, as well as techniques applied to explore the differences between translation variants under analysis, including the features related to their source texts as well as those of comparable target texts.

2.2 Translation Features and their Operationalisation

We group the features described above into three classes according to their correlations, especially in their operationalisation: 1) simplification, 2) explicitation, 3) normalisation vs. shining through and 4) convergence. Simplification can be analysed on different levels, e.g. lexical, syntactic or semantic. If core patterns of lexical use are observed (see Laviosa 1998), we can identify simplification comparing the proportion of content vs. grammatical words. Translated texts have a relatively low percentage of content words, and the most frequent words are repeated more often. This means, that both lexical density and type–token–ratio of translations are lower than those of their source texts and the comparable texts in the target language. Besides, more general terms are expected to be used in translations. On the level of syntax, one can observe short sentences which replace long ones and a lower average sentence length in general.

Explicitation involves the addition and specification of lexical and grammatical items, with the help of which implicit information in the source text is “spelled out” in its translation. The indicators of this feature include a higher ratio of function words which make grammatical relations explicit, specific terms replacing more general terms (the opposite of simplification), disambiguation of pronouns, increased use of cohesive devices, e.g. conjunctions, and others. In terms of cohesion, one would also expect more nominal (expressed with nominal phrases) than pronominal reference (expressed with personal pronouns) in translations.

Simplification and explicitation features correlate and may be just the opposite of each other. For example, if we observe more specific terms replacing general terms in translation, we face the feature of explicitation, and not simplification. Normalisation and “shining through” can also be measured on different levels, depending on the languages involved. Both features depend on the contrasts between these languages: normalisation implies the exaggerated use of the patterns typical for the target languages, whereas “shining through” involves the
patterns typical for the source language (but not specific for the target language) that can be observed in translations. For instance, normalisation can be verified by a great number of typical collocations and neutralised metaphoric expressions. Baker (1996) claims that influence of normalisation depends on the status of the source language: “the higher the status of the source text and language, the less the tendency to normalise”. We assume that the languages with a higher status also tend to “shine through” more often. For example, if we analyse translations from English, we would probably observe more “shining through” than normalisation, as English has the highest world language status.

And finally, convergence is a homogeneity feature of translations: they reveal less variation if we compare them to original texts. Convergence can also be observed on all levels of a language system. In accordance with the convergence phenomenon, one would expect that the lexical, grammatical and syntactic features under analysis will reveal smaller differences in translations than in originals.

2.3 Hypotheses

For our analysis of translation variants, we select a set of operationalisation of the features described in §2.2 above.

1. **Simplification** - We expect that our translated texts have a lower percentage of content words vs. grammatical words than their English source texts and the comparable German texts. Also, words are repeated more often in translations. Thus, we observe lower lexical density and type–token–ratio in our translations. In the analysis of English to German translations, we exclude sentence length as operationalisation for simplification. Due to the systemic differences in the morphology, German sentences are generally shorter than those in English, as they contain one-word compounds. To measure this uniformly, we need to split compounds and measure their parts as tokens, which is not feasible within this study.

2. **Explicitation** - Our translated texts reveal more cohesive explicitness than English and German originals: we can observe more conjunctions, less pronominal reference and less general nouns in translations than in English and German originals.

3. **Normalisation/ shining through** - If the translations under analysis demonstrate features more typical for English than for German, we observe “shining through”. If there are more features typical for German originals,
then our translations demonstrate normalisation. Here, we use the knowledge from contrastive analysis, e.g. German–English contrasts described in Hawkins (1986), König & Gast (2007), Steiner (2012). For example, we know that English is more “verbal” than German. This can be proved by comparing the distribution of nominal and verbal phrases in both translations and originals. English originals are expected to contain more verbal than nominal phrases. The phenomenon of “shining through” will be confirmed in our data if translations contain more verbal phrases than German originals. On the contrary, if they contain less verbal phrases than German originals, the normalisation hypothesis will be confirmed.

4. Convergence - The variation of the features in 1 to 3 is not great if we compare translation variants: they are similar to each other, i.e. the features are distributed homogeneously.

3 Resources, Methods and Tools

To prove the hypotheses formulated in §2.3, we need to compare the distribution of the features under analysis across translation variants, their English sources as well as comparable German originals. For this, we analyse frequency distribution information of lexico-grammatical patterns which serve as operationalisation for these features. The patterns are extracted from a corpus at hand, and evaluated with univariate statistical methods (e.g. significance analysis).

3.1 Corpus Resources

For our investigations, we use VARTRA-SMALL, (see Lapshinova-Koltunski 2013), a translation corpus which contains German translation variants from English produced with different translation methods: by (1) human professionals (pt), (2) human inexperienced translators (cat), with (3) rule-based MT systems (rbmt) and (4) two statistical MT systems (smt1 and smt2). Translations by professionals (pt) were exported from the already existing corpus CroCo (Hansen-Schirra, Neumann & Steiner 2013). The same corpus provides source English texts (eo) and comparable German originals (go). Thus, we can compare source English texts with their multiple translations into German, as well as to comparable German originals.

The cat variant was produced by trained translators with at least BA degree, who have no/little experience in translation. All of them applied computer-aided
tools while translating the given texts.\footnote{We used the open source tool \textsc{across}, see www.my-across.net.} The rule-based machine translation variant was translated with \textsc{systran} (\textsc{rbmt}),\footnote{\textsc{systran} 6.} whereas for statistical machine translation we have two further versions – the one produced with Google Translate\footnote{http://translate.google.com/.} (\textsc{smt1}), and the other – with a self-trained Moses system (\textsc{smt2}) (see Koehn et al. 2007).

The analysed dataset covers seven registers of written language: political essays (essay), fictional texts (fiction), manuals (instr), popular-scientific articles (popsci), “letters to share-holders” (share), prepared political speeches (speech), and tourism leaflets (tou). The size of all translation variants in \textsc{vartra-small} comprises approx. 600 thousand tokens. The subcorpora of originals from \textsc{CroCo} comprise around 250 thousand words each.

All subcorpora under analysis are tokenised, lemmatised and tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations of the \textsc{vartra-small} subcorpora were obtained with Tree Tagger (see Schmid 1994). The availability of these annotation levels in both corpora allows us to analyse certain lexico-grammatical patterns – operationalisation of the translation features under analysis, defined in §2.3.

The subcorpora are encoded in \textsc{cwb} format (\textsc{cwb}, 2010) and can be queried with the help of the \textsc{cqp} regular expressions described in Evert (2005).

Alignment on sentence level is available for professional translations only: each translation is aligned with its English source on sentence level. No alignment is provided for further translation variants at the moment. However, this annotation level is not necessary for the extraction of the operationalization used in the present paper.

3.2 Feature Extraction

As already mentioned in §3.1 above, the corpus at hand can be queried with \textsc{cqp}, which allows the definition of language patterns in form of regular expressions based on string, part-of-speech and chunk tags as well as further constraints.

To prove the hypothesis for simplification indicated by lexical density (proportion of content words), we extract information on the distribution of content words in our corpus, for which the query 1 in Table 1 is used.

To extract the corpus evidence of explicitation, we apply queries 2 to 5. Query 2 is used to extract all occurrences of coordinating and subordinating conjunctions,
Table 1: Queries for feature extraction

<table>
<thead>
<tr>
<th>query element</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 [pos=&quot;vv.*</td>
<td>n.*</td>
</tr>
<tr>
<td>2 [pos=&quot;kon</td>
<td>kous&quot;]</td>
</tr>
<tr>
<td>3 (&lt;np&gt;[pos=&quot;ppe.*&quot;]&lt;/np&gt;)</td>
<td>nominal phrase filled with a pronoun</td>
</tr>
<tr>
<td>4 (&lt;np&gt;[ ]+&lt;/np&gt;)</td>
<td>any nominal phrase</td>
</tr>
<tr>
<td>5 [lemma=re($general)] [pos=&quot;n.*&quot;]</td>
<td>nouns from a list</td>
</tr>
<tr>
<td>6 (&lt;np&gt;[ ]+&lt;/np&gt;(&lt;pp&gt;[ ]+&lt;/pp&gt;))</td>
<td>nominal phrase or prepositional phrase</td>
</tr>
<tr>
<td>7 (&lt;vp&gt;[ ]+&lt;/vp&gt;)</td>
<td>verbal phrase</td>
</tr>
</tbody>
</table>

whereas queries 3 and 4 are used for extraction of information on pronominal vs. nominal reference in the corpus.

We calculate this as proportion of nominal phrases filled with personal pronouns (query 3) to all nominal phrases in the corpus (query 4). Query 5 is used to extract occurrences of general terms in order to compare their proportion to all nouns in the dataset. For this, we use a simple lexical search – we extract a closed class of lexical items of which we know the members. Here, we use lists of general nouns as defined in (Dipper, Seiss & Zinsmeister 2012). For normalisation/shining through, we extract all occurrences of nominal and prepositional phrases (query 6) vs. verbal phrases (query 7). Convergence is proved with the help of all patterns described above.

As we operate with low-level features which do not require formulation of complex lexico-grammatical patterns, we believe that our feature extraction procedures are adequate for the present task. Its only shortcoming is the potential noise caused by tagging errors, especially in case of machine translation. In the latter, we observe a number of untranslated words which are tagged as named entities by automatic part-of-speech taggers. In the longer run, we aim to include deeper structures into our analysis which would require parsed data.

4 Results and their Interpretation

4.1 Simplification

In the first step, we want to test if lexical density and type–token–ratio are lower in translation variant than in EO and GO.
5 Variation in translation

Table 2: sttr and ld in vartra-small

<table>
<thead>
<tr>
<th></th>
<th>EO</th>
<th>GO</th>
<th>PT</th>
<th>CAT</th>
<th>HU-x</th>
<th>RBMT</th>
<th>SMT1</th>
<th>SMT2</th>
<th>MT-x</th>
<th>Trans-x</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD</td>
<td>45.72</td>
<td>45.49</td>
<td>46.23</td>
<td>44.60</td>
<td>45.64</td>
<td>45.08</td>
<td>46.02</td>
<td>47.86</td>
<td>46.30</td>
<td>45.97</td>
</tr>
<tr>
<td>STTR</td>
<td>367.5</td>
<td>369.9</td>
<td>360.8</td>
<td>336.4</td>
<td>348.6</td>
<td>335.2</td>
<td>350.4</td>
<td>309.0</td>
<td>331.5</td>
<td>338.4</td>
</tr>
</tbody>
</table>

As already mentioned above, lexical density (ld in Table 2) is measured as a proportion of all content words in our corpus. Unexpectedly, average lexical density in translations (Trans-x in Table 2) does not differ from that of both source and comparable originals. Moreover, if we consider the mean values for human and machine translations separately (HU-x and MT-x respectively); the latter demonstrates even higher LD than human translations and English and German originals. The lowest figure is obtained for CAT, which demonstrates a value below the average. The highest value is observed for SMT2 (47.86).

We explain this by the lexical constraints of the Moses-based system: this system depends on the parallel data used for its training. If the parallel data does not contain translations for some words in a text to be translated, the system keeps them untranslated. In the automatic part-of-speech tagging, these words are then tagged as proper nouns (ne) which leads to their high amount in texts, as seen in example (1).

However, the overall difference between originals and translations is not great, which means that lexical density is not an indicator of simplification in our dataset, as the translated texts show an amount of content words similar to that of the source and comparable originals.

(1) Wenn Sie strongly, believe, wie ich, dass Großbritannien einen zentralen Platz einnehmen müssen in Europe’s decision-making...

Another indicator of simplification is type–token–ratio which we measure as standardised type–token–ratio (STTR) – a percentage of different lexical word forms (types) per text. As expected, on average, translations show lower STTR than their source texts and comparable originals, see Table 2. Mean value of human translations is also higher than that of machine (348.6 vs. 338.6 respectively). Within translations, the highest STTR, thus, the most lexically rich translation...
variant in our corpus, is the one produced by professional human translators (360.8), followed by smt1 (350.4), and cat (336.4). The level of the latter is close to the average of all translations but lower than that of human translations. The lowest figure is obtained for smt2 (309.0). This can once again be explained by the fact that this translation variant contains a great deal of untranslated English words, the lemmas of which cannot be identified by the lemmatiser and thus is replaced with “<unknown>”, see example (2).

(2) Closing die Gap Zwischen Supply und Die Nachfrage
<unknown> d <unknown> zwischen <unknown> und d Nachfrage nach A balanced , umfassende Energiepolitik ist dringend nach A <unknown> , umfassend Energiepolitik sein dringend erforderlich , die langfristige Stärke der amerikanische wirtschaftlichen erforderlich d langfristig Stärke d amerikanisch wirtschaftlich und nationalen.
und national <unknown>

Interestingly, student translations are closer to the rbmt translation variant in terms of both sttr (336.4 vs. 335.2) and ld (44.60 vs. 45.08). Analysing human and machine translation separately, we observe the same ranking in terms of both indicators: pt > cat, whereas it is not stable in machine translation: while smt2 ranks first in ld, it occupies the last position with its sttr value.

4.2 Explicitation

To analyse this feature in our corpus, we measure cohesive explicitness in all subcorpora. Here, we calculate the relative frequencies for conjunctions (conj in Table 3, normalised to the total number of words per thousand), proportion of nominal phrases filled with pro-forms vs. full nominal phrases (pronnp in Table 3, normalised per thousand), as well as proportion of general nouns vs. all noun occurrences (gennoun in Table 3 normalised per thousand) in translations and English and German originals.

According to our hypothesis in §2.3, we expect more conjunctions, less pronominal reference and less general nouns in translations than in originals. If we compare the values of all translations (Trans-\(\bar{x}\)) with those of their originals, our hypothesis can be confirmed for pronominal reference and general nouns only: Trans-\(\bar{x}\) (137.76) < eo (204.67) and Trans-\(\bar{x}\) (20.51) < eo (48.71). Translations demonstrate a lower and not higher distribution of conjunctions, Trans-\(\bar{x}\) (50.67) < eo (53.80), contrary to what was expected. If we consider human and machine
5 Variation in translation

Table 3: Explicitation indicators

<table>
<thead>
<tr>
<th></th>
<th>conj</th>
<th>PRONNP</th>
<th>gennoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO</td>
<td>53.80</td>
<td>204.67</td>
<td>48.71</td>
</tr>
<tr>
<td>GO</td>
<td>43.58</td>
<td>127.14</td>
<td>23.85</td>
</tr>
<tr>
<td>PT</td>
<td>47.58</td>
<td>232.76</td>
<td>19.64</td>
</tr>
<tr>
<td>CAT</td>
<td>49.67</td>
<td>139.12</td>
<td>19.93</td>
</tr>
<tr>
<td>HU-*</td>
<td>48.33</td>
<td>184.67</td>
<td>19.74</td>
</tr>
<tr>
<td>RBMT</td>
<td>53.32</td>
<td>144.46</td>
<td>23.18</td>
</tr>
<tr>
<td>SMT1</td>
<td>52.54</td>
<td>143.15</td>
<td>21.22</td>
</tr>
<tr>
<td>SMT2</td>
<td>53.69</td>
<td>39.85</td>
<td>19.46</td>
</tr>
<tr>
<td>MT-*</td>
<td>53.18</td>
<td>107.65</td>
<td>21.19</td>
</tr>
<tr>
<td>Trans-*</td>
<td>50.76</td>
<td>137.76</td>
<td>20.51</td>
</tr>
</tbody>
</table>

Translation separately, we see that values for machine translation are much closer to EO (53.18 vs. 53.80), which means that in these translation variants, cohesive relation expressed via conjunctions, were preserved similarly to their English originals. Conversely, fewer conjunctions were used in human translation. The number is still higher than observed in German originals (43.58); therefore, we cannot assume the phenomenon of normalization here. This means that, in our dataset, human translators tend to keep that relation implicit, as seen in example (3).

(3) a. Negative molecules moved into the nurse cells if the egg was made negative, while positive molecules stayed put (EO-PopsCI).

b. Wenn das Ei auch negativ war, bewegten sich negativ geladene Moleküle in die Nährzellen, positiv geladene Moleküle blieben an Ort und Stelle (PT-PopsCI).

Admittedly, our extractions exclude occurrences of adverbial conjunctions (as we extract coordinating and subordinating conjunctions only). Previous analyses (e.g. Kunz & Lapshinova-Koltunski 2014) show that this syntactic type of conjunction is highly frequent in German. We suppose that English coordinating and subordinating conjunctions are in some cases translated with adverbials in German.

Example (4) extracted from our corpus demonstrates variants of translation of the English subordinating conjunction “while”. In both human translations (b.
and c.), conjunctive relation is transferred with an adverbial phrase. In machine translated variants (d. to f.), “while” is translated directly with **während**, so the type of cohesive conjunction is preserved as it was in the original.

(4)  

a. And *while* this will vary from quarter to quarter based on large cash outlays such as tax payments and end-of-year compensation payments, we were pleased with our average positive cash flow for the year from operations of $1.5 billion per quarter. (po-share).


d. Und basiert *während* dieses von Viertel zu das Viertel schwankt, das auf grossen Barauslagen wie Steuerzahlungen und Jahresendeausgleichszahlungen, wurden wir mit unserem durchschnittlichen positiven Cashflow für das Jahr von den Operationen von $1,5 Milliarden pro Viertel gefallen (rbmt-share).

e. Und *während* dies von Quartal zu Quartal basierend auf grosse Barauslagen wie Steuer-Zahlungen und End-of-Jahres-Ausgleichszahlungen variieren, wurden wir mit unserer durchschnittlichen positiven Cashflow für das Jahr aus dem operativen Geschäft von 1,5 Milliarden Dollar pro Quartal (smt1-share).

f. Und *während* diese je nach Viertel bis Viertel auf der Grundlage grosse Geld ausgegeben wie Steuerzahlungen und abschliessende Entschädigung payments, freuen wir uns mit unseren durchschnittliche positive Cashflow für das Jahr von Maßnahmen der $1.5 Milliarden pro quarter (smt2-share).

In some cases, cohesion might be expressed with different cohesive devices in the two languages under analysis. For instance, the conjunction “while” in the
5 Variation in translation

source sentence in example (5) is substituted with a reference expressed with the pronominal adverb *dabei* in pt, see (5-b). Pronominal adverbs expressing a reference are typical for German and are rare in English. At the same time, we observe the adoption of the cohesive device used in the source sentence also in other translation variants (c. to f.).

(5) a. My father preferred to stay in a bathrobe and be waited on for a change *while* he lead the stacks of newspapers *me* and *my* grandmother saved for him (eo-fiction).

b. Mein Vater ist lieber im Bademantel geblieben und hat sich zur Abwechslung mal bedienen lassen und *dabei* die Zeitungsstapel durchgelesen, die ich und meine Großmutter für ihn aufgehoben haben (pt-fiction).


e. Mein Vater lieber im Bademantel bleiben und werden wartete auf eine Veränderung, *während* er die Stapel von Zeitungen mich und meine Großmutter für ihn gerettet führen (smt1-fiction).


In terms of reference, translations demonstrate less noun phrases filled with pronouns than their source texts in English: 137.76 (Trans-x) vs. 204.14 (eo), whereas the opposite phenomenon is observed, if we compare them to the original texts in German. In this case, we observe more pronominal reference in translations than in comparable originals (137.76 vs. 127.14). However, variation is observed across translation variants: while in human translations pronominal reference is much higher and tends to the values of eo, machine translation shows values which are lower when compared to both eo and go. This low value is obviously caused by the small amount of pronominal phrases in smt2. Here, we suppose that many pronouns remained untranslated in certain registers, as seen in example (5-f) above, and were wrongly tagged in the part-of-speech an-
notation. Moreover, we observe a high number of pronominal references in Pt (232.76), which contradicts the hypothesis in §2.3.

The figures obtained for general nouns confirm our hypothesis about their low frequency in translations. On average, translations demonstrate a lower amount of general nouns than eo and go (20.51 vs. 48.71 and 23.85 respectively). RBMT is the only translation variant whose distribution of general nouns is similar to that of go. As seen from the values for the originals, there are more general nouns in eo than in go. This means that this type of nouns is more typical for English than for German. Hence, we observe normalisation in terms of general nouns in all translation variants of our corpus.

In the analysis of explicitation in translations from English into German, one should also take into account the fact that German is more explicit than English, which could also have influenced on the results obtained.

4.2.1 Normalisation and “shining through”

To analyse normalisation and “shining through”, we extracted all occurrences of nominal and prepositional phrases and compared them with the occurrences of verbal phrases. Table 4 demonstrates the proportions of nominal (nominal and prepositional phrases) and verbal (verbal phrases) classes across all subcorpora under analysis. As already mentioned in §2.3 above, German is less “verbal” than English, which is confirmed in our data: go contains less verbal phrases than eo.

The mean value of verbal phrases for all translations comprises 28.63, which is much lower than that of go. This indicates the phenomenon of normalisation in this case. Comparing the values across translation variants, we observe variation in the degree of normalisation – it is less pronounced in human than in machine translation (33.59 vs. 24.64 respectively). Moreover, human translations produced by professionals are very close to German originals in terms of the distribution of nominal vs. verbal phrases, which means that they demonstrate neither normalisation nor “shining through” if we consider the indicators under analysis.

The higher noun-verb-ratio (nvratio in Table 4) is observed for smt2. The reason for it could once again be the erroneous part-of-speech tagging which results from the gaps in training data used for smt2. Most untranslated verbs (e.g. promote, report, import) or verbal forms (recognising, closing, helping, etc.) were tagged as nouns or adjectives.

Overall, the results are rather surprising. Analysing examples in our corpus, we notice that source verbal phrases in human translations from English into German are often translated as nominal phrases, see examples (6-a), (6-b) and
5 Variation in translation

Table 4: Proportionality of nominal vs. verbal opposition in vartra-small

<table>
<thead>
<tr>
<th>subc</th>
<th>nominal</th>
<th>verbal</th>
<th>NVRATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO</td>
<td>59.45</td>
<td>40.55</td>
<td>1.47</td>
</tr>
<tr>
<td>GO</td>
<td>61.95</td>
<td>38.05</td>
<td>1.63</td>
</tr>
<tr>
<td>PT</td>
<td>61.92</td>
<td>38.08</td>
<td>1.63</td>
</tr>
<tr>
<td>CAT</td>
<td>71.87</td>
<td>28.13</td>
<td>2.56</td>
</tr>
<tr>
<td>HU-\text{\textbar}</td>
<td>66.41</td>
<td>33.59</td>
<td>1.98</td>
</tr>
<tr>
<td>RBMT</td>
<td>72.42</td>
<td>27.58</td>
<td>2.63</td>
</tr>
<tr>
<td>SMT1</td>
<td>74.38</td>
<td>25.62</td>
<td>2.90</td>
</tr>
<tr>
<td>SMT2</td>
<td>79.54</td>
<td>20.46</td>
<td>3.89</td>
</tr>
<tr>
<td>MT-\text{\textbar}</td>
<td>75.35</td>
<td>24.64</td>
<td>3.06</td>
</tr>
<tr>
<td>Trans-\text{\textbar}</td>
<td>71.36</td>
<td>28.63</td>
<td>2.49</td>
</tr>
</tbody>
</table>

(6-c). However, they are often left as verbal phrases in machine translation, as in examples (6-d), (6-e) and (6-f). Therefore, we would expect machine-produced translations to have a lower noun-verb-ratio, which is not the case in the quantitative data. To analyse the correspondences between source and target phrases we need to align our subcorpora, which is not available at the moment.

(6) a. Settings changed here override settings changed anywhere else (EO-instr).
   b. Die hier vorgenommenen Änderungen setzen alle anderen Änderungen außer Kraft (PT-instr).
   c. Hier vorgenommene Einstellungsänderungen sind allen anderen Einstellungsänderungen übergeordnet (CAT-instr).
   d. Die Einstellungen, die hier geändert werden, heben die Einstellungen auf, die irgendwoanders geändert werden (RBMT-instr).
   e. Hier geänderten Einstellungen überschreiben Einstellungen, die anderswo geändert (SMT1-instr).
   f. ... bei dem Sie überhaupt hier über Rahmenbedingungen geändert Settings überall else (SMT2-instr).
4.3 Convergence

In our last hypothesis, we test if the analysed translations exhibit convergence – the variation of the features across translation variants in our corpus is not high. For this purpose, we consider the indicators analysed in §1, §2 and §3 above: sttr, ld, conj, pronnp, gennoun and nvratio. The overall variation between the subcorpora is relatively low for all features, except for pronominal reference and noun-verb-ratio (see Figure 1), which means that translation variants in our corpus are alike in terms of the features considered. Most prominent indicators for convergence are that of simplification. We remove pronominal reference and noun-verb-ratio from the data matrix and calculate p–values using the Pearson’s chi–square test, which is a univariate statistical method to reveal significant differences between variables. If p–value is < 0.05, then the difference between the compared subcorpora (translation variants) is not significant.

![Figure 1: Levelling out in vartra-small](image)

We calculate p–value for all pairs of subcorpora in vartra. The results confirm our assumptions, as p–value is above 0.05 in almost all cases (see Table 5). An exception is pair pt-smt2, where we observe a p–value of approx. 0.01. Our translation variants therefore converge, as expected, as there is no significant difference between almost all subcorpora; they are alike in terms of the analysed features.
5 Variation in translation

Table 5: p–values for comparison of translation variants

<table>
<thead>
<tr>
<th>subc</th>
<th>p–value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT vs. CAT</td>
<td>0.8863</td>
</tr>
<tr>
<td>PT vs. RBMT</td>
<td>0.5663</td>
</tr>
<tr>
<td>PT vs. SMT1</td>
<td>0.8806</td>
</tr>
<tr>
<td>PT vs. SMT2</td>
<td>0.0142</td>
</tr>
<tr>
<td>CAT vs. RBMT</td>
<td>0.9307</td>
</tr>
<tr>
<td>CAT vs. SMT1</td>
<td>0.9986</td>
</tr>
<tr>
<td>CAT vs. SMT2</td>
<td>0.0980</td>
</tr>
<tr>
<td>RBMT vs. SMT1</td>
<td>0.9373</td>
</tr>
<tr>
<td>RBMT vs. SMT2</td>
<td>0.1771</td>
</tr>
<tr>
<td>SMT1 vs. SMT2</td>
<td>0.0731</td>
</tr>
</tbody>
</table>

phenomena, which are indicators of simplification, explicitation and normalisation.

4.4 Summary

Summarising the obtained results, we found that not all hypotheses formulated in §2.3 above can be applied to our dataset. Both type–token–ratio as well as lexical density do not serve as good indicators of simplification in this case. In terms of explicitation, we should also think of further operationalisation, as those chosen reveal rather other phenomena (e.g. normalisation). The hypotheses about normalisation and “shining through” can be confirmed only in part and reflect high variations across translation varieties. The only assumption confirmed by our data is that of convergence. The analysed translation variants converge, as there is no significant difference between them in terms of the analysed phenomena.

5 Conclusion and future work

In this paper, we analysed translation variants produced by humans and machine systems and compared them to their English source texts, as well as comparable German originals. With the help of lexicogrammatical patterns, we were able
Ekaterina Lapshinova-Koltunski

to trace differences and similarities between them, which indicate the following translation features: simplification, explicitation, normalisation and convergence. Although our analysis includes translations from English into German, we could not detect “shining through” – at least with the indicators at hand. The analysed features vary if we consider translation variants or their groups separately, e.g. in terms of explicitation or normalisation. At the same time, we observe convergence in translation, especially if we take simplification into account.

We believe that we should include more factors into the analysis to explain the variation observed. For example, in some cases, we should revise our hypotheses and their operationalisation, as contrasts between languages should be taken into account. We also need to look at the “experience” factor – this could verify the differences between two human translations observed for some features. Furthermore, restrictions of the translation memory applied in cat or the training material used in smt can also have an influence on the distribution of lexico-grammatical patterns. For this, a closer inspection of correlations between translation memory as well as applied smt training material (parallel corpora) is required, which is planned for our future work.

We also plan to align originals with their translations on word and sentence level to allow analysis of certain phenomena involved, e.g. translation of ambiguous cases, direct translation solutions, see 4.3 and their multiple variants.

Acknowledgments

The project vartra: Translation Variation was supported by a grant from Forschungsausschuss of the Saarland University.

References


Ekaterina Lapshinova-Koltunski


5 Variation in translation


Exploration of Inter- and Intralingual Variation of Discourse Phenomena

Ekaterina Lapshinova-Koltunski
Saarland University
FR4.6, University Campus, D-66123 Saarbrücken
e.lapshinova@mx.uni-saarland.de

Abstract

In this paper, we analyse cross-linguistic variation of discourse phenomena, i.e. coreference, discourse relations and modality. We will show that contrasts in the distribution of these phenomena can be observed across languages, genres, and text production types, i.e. translated and non-translated ones. Translations, regardless of the method they were produced with, are different from their source texts and from the comparable originals in the target language, as it was stated in studies on translationese. These differences can be automatically detected and analysed with exploratory and automatic clustering techniques. The extracted frequency-based profiles of variables under analysis (languages, genres, text production types) can be used in further studies, e.g. in the development and enhancement of MT systems, or in further NLP applications.

1 Introduction

Although considerable research aiming at enhancing machine-translated texts with discourse properties achieved positive results in recent years, see e.g. (Webber et al., 2013; Hardmeier, 2014) or (Meyer et al., 2015), some document-wide properties of automatically translated texts still require improvement, as translation models are induced from stand-alone pairs of sentences. Moreover, target language models approximate the target language on the string level only, whereas target texts have properties that go beyond those of their individual sentences and that reveal themselves in the frequency and distribution of certain structures. These frequency- and distribution-based properties of translated and non-translated texts are in focus of corpus-based translation studies. However, these properties (in form of higher-level language models) may also be useful for natural language processing (NLP), including machine translation (MT).

In this paper, we show an example of a corpus-based analysis of interlingual (between English and German) and intralingual (across different genres) variation of discourse properties in translated and non-translated texts. In particular, this paper will focus on various types of discourse relational devices, pronominal referring expressions, as well as modal meanings expressed with particular modal verbs. The frequencies of these discourse features will be automatically extracted from English-German comparable corpora which also contain multiple translations produced with several methods, including manual and automatic ones. We will compare the distributions of these features in both languages, as well as in translations from English to German, paying attention to their variation across genres available in the dataset. We will also consider differences in their distributions in human and machine translation. For our analysis, we apply exploratory and unsupervised classification techniques. The obtained information on the frequency-based interlingual and intralingual differences may be valuable for linguistic studies on language contrasts, human translation, and may find application in NLP and especially MT.

2 Related Work

2.1 Discourse properties in English and German

Various discourse phenomena have been in focus of several translation studies and those on language contrasts dealing with English and German. Recent years have seen an increase in the number of works employing corpus-based methods for their analysis. However, multilingual stud-
ies are mostly concerned with individual phenomena in particular genres, see e.g. (Bührig and House, 2004) for particular cohesive conjunctions or adverbs in prepared speeches, (Zinsmeister et al., 2012) for abstract anaphora in parliament debates, and (Taboada and Gómez-González, 2012) for particular coherence relations. The latter, however, considers two modes: spoken and written, and states that the differences between modes are more prominent than between languages. Kunz and Lapshinova-Koltunski (2015) and Kunz et al. (2015) show that distributions of different discourse phenomena are not only mode- but also genre-dependent. The authors show this for a number of textual phenomena, analysing structural and functional subtypes of coreference, substitution, discourse connectives and ellipsis. Their dataset includes several genres, and they are able to identify contrasts and commonalities across languages (English and German) and genres with respect to the subtypes of all textual phenomena under analysis, showing that these languages differ as to the degree of variation between individual genres. Moreover, there is more variation in the realisation of discourse devices in German than English. The authors attested the main differences in terms of preferred meaning relations: a preference for explicitly realising logico-semantic relations by discourse markers and a tendency to realise relations of identity by coreference. Interestingly, similar meaning relations are realised by different subtypes of discourse phenomena in different languages and genres.

2.2 Discourse properties in human and machine translation

Cross-lingual contrasts stated on the basis of non-translated data are also of great importance for translation. Kunz et al. (2015) suggest preferred translation strategies on the basis of contrastive interpretations for the results of their quantitative analysis, which show that language contrasts are even more pronounced if we compare languages per genre. These contrasts exist in the features used for creating textual relations. Therefore, they suggest that, for instance, when translating popular science texts from English into German translators should more extensively use linguistic means expressing textual relations. Overall, they claim that translators should use more explicit devices translating from English into German, e.g. demonstrative pronouns should be used more often instead of personal pronouns (e.g. dies/das instead of es/it). The opposite translation strategies should be used when translating from German to English.

However, studies of translated language show that translators do not necessarily apply such strategies. For instance, Zinsmeister et al. (2012) demonstrate that translations in general tend to preserve the source language anaphor’s categories, functions and positions, which results in the shining through effect (shining through of the source language preferences, see (Teich, 2003)) in both translation directions. Additionally, due to the tendency to explicate textual relations, translators tend to use more nominal coreference instead of pronominal one. Explicitation (tendency of translations to be more explicit than their sources, see (Vinay and Darbelnet, 1958) and (Blum-Kulka, 1986)) along with shining through belong to the characteristics of translated texts caused by peculiarities of translation process. A number of works on discourse connectives, e.g. (Becher, 2011; Bisiada, 2014; Meyer and Webber, 2013) and (Li et al., 2014), show implicit/explicit discourse expression divergence in both human and machine translation. There are several studies that attempt to incorporate information on discourse relations or other discourse properties into MT, see for instance, those by Le Nagard and Koehn (2010), Hardmeier and Federico (2010) and Guillou (2012), or those presented within the first DiscoMT workshop, see (Webber et al., 2013). Most of them employ parallel corpora, thus, the approximation of the target language is based on translations, which, however, possess characteristics that differ them from non-translated texts originally written in a target language, also in terms of discourse properties. This paper will consider discourse-related characteristics that differ translation from non-translated texts, and also differentiate human from machine translations.

3 Methodology

3.1 Data

As we focus on variation of discourse phenomena in English and German, as well as English-German translations, our data should contain both English-German parallel texts and non-translated comparable texts in German. Furthermore, as we are also interested in linguistic variation in terms of genre, the texts should be from different gen-
res. For this reason, we had to dismiss the typical corpora used in MT, e.g. Europarl (Koehn, 2005) or TED talks, as translated texts in these resources are not comparable. The latter contains multilingual subtitles which are produced under different restrictions than those of translations. We also expect that some of the phenomena under analysis might be omitted in the subtitles, as this is recommended in the guidelines\(^1\). So, we select two corpora which contain English-German parallel and comparable texts from different genres. English and German originals (EO and GO) were extracted from CroCo (Hansen-Schirra et al., 2012), whereas German translations originate from the VARTRA corpus (Lapshinova-Koltunski, 2013), as it contains multiple translations of the CroCo English originals produced both manually and automatically (HU and MT).

The whole dataset totals 406 texts which cover seven genres: political essays (ESS), fictional texts (FIC), instruction manuals (INS), popular-scientific articles (POP), letters to shareholders (SH), prepared political speeches (SP), and tourism leaflets (TOU). The decision to include this wide range of genres is justified by the need for heterogeneous data for our experiment. The number of words per genre in comprises ca. 36 thousand tokens. We tag both English and German data with the TreeTagger tools (Schmid, 1994).

### 3.2 Feature selection

Linguistic relations between textual elements help recipients in their cognitive interpretation as to how different thematic concepts are connected. These relations are indicated by particular structures that language producers employ, e.g. grammatical items such as connectives, personal and demonstrative pronouns, substitute forms, elliptical constructions and lexical items, such as nouns, verbs and adjectives. As already mentioned in Section 1 above, we will analyse discourse relations, coreference and modality.

For discourse relations, we will analyse connectives classified according to the semantic relations they convey. Our classification is based on semantic relations defined by Halliday and Hasan (1976) and includes additive (relation of addition, e.g. and, in addition, moreover), adversative (relation of contrast/alternative, e.g. yet, although, by contrast), causal (relation of causality/dependence, e.g. because, therefore, that’s why), temporal (temporal relation between events such as after, afterwards, at the same time) and modal relations (expressing rather a pragmatic meaning, in which evaluation of the speaker is involved, e.g. unfortunately, surely).

Demonstrative and personal pronouns (such as this, that, she, his, theirs, it, etc.) will serve as triggers of coreference. We also consider distributions of general nouns, e.g. plan, case, fact, which commonly function as abstract anaphora (Zinsmeister et al., 2012). For the analysis of modality, we consider frequencies of modal verbs grouped according to the modal meanings defined by Biber et al. (1999): permission (can/could, may/might), volition (will, would, shall) and obligation (must, ought to, should, need to, have got to, suppose to).

<table>
<thead>
<tr>
<th>Feature pattern</th>
<th>Discourse property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permission</td>
<td>Modality</td>
</tr>
<tr>
<td>Obligation</td>
<td></td>
</tr>
<tr>
<td>Volition</td>
<td></td>
</tr>
<tr>
<td>Additive</td>
<td>Discourse relations</td>
</tr>
<tr>
<td>Adversative</td>
<td></td>
</tr>
<tr>
<td>Causal</td>
<td></td>
</tr>
<tr>
<td>Temporal</td>
<td></td>
</tr>
<tr>
<td>Modal</td>
<td></td>
</tr>
<tr>
<td>General nouns</td>
<td>Coreference</td>
</tr>
<tr>
<td>Perspron</td>
<td></td>
</tr>
<tr>
<td>Dempron</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1: Features under analysis**

The set of 11 selected features is outlined in Table 1. The first column denotes the extracted and analysed feature patterns, the second represents the corresponding discourse property. For the extraction of the frequencies of these feature patterns, we use a number of regular expressions based on string, part-of-speech and chunk tags, as well as further constraints, e.g. position in a sentence or in a text. Frequency information is collected both per text, and per subcorpus (e.g. per genre in a certain language).

### 3.3 Methods

For our analysis, we use exploratory and also unsupervised classification (automatic clustering) techniques which will allow us to observe differences between groups of texts and subcorpora, and also to discriminate between them on the basis of discourse features described in Section 3.2.

We apply correspondence analysis (CA) (Venables and Smith, 2010; Baayen, 2008; Greenacre,
that is conceptually similar to principal component analysis (PCA), with the difference that the data is scaled so that rows and columns are treated equivalently. Thus, this technique will help us to see not only which variables (e.g. languages or genres) have similarities, but also possible correlation of these variables with discourse features contributing to these similarities, as distances between between dependent and independent variables are calculated. These distances are then represented in a two-dimensional map, and the larger the differences between subcorpora or texts, the further apart they are on the map. Likewise, dissimilar categories of discourse phenomena are further apart. Proximity between subcorpora and discourse features in the merged map is as good an approximation as possible of the correlation between them. In computing this low-dimensional approximation, CA transforms the correlations between rows and columns of our table into a set of uncorrelated variables, called 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. In this way, we can identify new meaningful underlying variables, which ideally correlate with such variables as language or genre, indicating the reasons for the similarities or differences between these subcorpora. The length of the arrows in the graph indicates how pronounced a discourse feature is, see (Jenset and McGillivray, 2012) for details. The position of the points in relation to the arrows indicates the relative importance of a feature for a subcorpus. The arrows pointing in the direction of an axis indicate a high correlation with the respective dimension, and thus, a high contribution of the feature to this dimension.

The results of automatic clustering will indicate differences and similarities between the languages (English and German) and their varieties (genres). Moreover, we can also discover differences between non-translated and (manually or automatically) translated texts. We decide for unsupervised techniques, in favour of different genres contained in our data, and supervised classification performs better with single genre data, so that in a supervised scenario, we would need to perform several classification tasks. We apply hierarchical cluster analysis (HCA), see (Hothorn and Everitt, 2014) and (Everitt et al., 2011). This clustering technique is connectivity-based as its core idea is that objects are more related to nearby objects than to objects farther away. Objects, in our case texts and subcorpora, are connected to form clusters based on their distance measured here on the basis of the feature distributions. We calculate the distance by the Euclidean distance which is one of the most straightforward and generally accepted ways of computing distances between objects in a multi-dimensional space. The results of hierarchical clusters are represented graphically in a dendrogram, which is a branching diagram that represents the relationships of similarity among a group of entities. The arrangement of the branches tells us which texts/subcorpora (on leaves) are most similar to each other. The height of the branch points indicates how similar or different they are from each other. Ward’s method (also called Ward’s minimum variance method) is employed to perform clustering. This method minimises the total within-cluster variance after merging.

The main drawback of this technique is that the number of clusters needs to be specified in advance. Therefore, we apply a technique based on bootstrap resampling, with the help of which we are able to produce $p$-value-based clusters, i.e. that are highly supported by the data will have large $p$-values. The output dendrogram demonstrates two types of $p$-values: AU (Approximately Unbiased) $p$-value and BP (Bootstrap Probability) value. AU $p$-value, which is computed by multi-scale bootstrap resampling, is a better approximation to unbiased $p$-value than BP value computed by normal bootstrap resampling.

4 Analyses

4.1 Discourse properties in English and German

First, we analyse English and German non-translated texts, to define the differences between these languages in terms of discourse properties. We perform CA on the subset of data containing originals only. In the first step, the dataset is labelled with text IDs only (e.g. EO_001, GO_010, etc.).

In Table 2, we present the Eigenvalues calculated for each dimension to assess how well our

\footnote{We use pvclust() package available in the R environment (version 3.0.2; (Team, 2013)).}
The second dimension (the y-axis) clearly indicates language-independent differences in genres:

Figure 1: Variation of discourse phenomena across languages

Figure 2: Variation of discourse phenomena across genres

Table 2: Contribution of dimensions for variation across languages

Table 3: Contribution of dimensions for variation across genres

We plot the results in a two-dimensional graph in Figure 1, representing the first two dimensions, which explain 67.60% (cumulative value) of the data inertia. The second dimension although covering only 20.50% is also important for our analysis if we want to explain more than 50% of the data variation. The rest of inertia remain unexplained with the two-dimensional representation.

Concerning dimension 1 (47.10% of inertia), we see a clear distinction between English and German texts (along the x-axis on the left and on the right from zero respectively). So, the distinction along this dimension reflects language contrasts in the use of particular discourse features, i.e. different types of discourse relations via connectives for German, and coreference via demonstrative pronouns, modal meaning of volition and causal logico-semantic relations for English. The assumption is that the second dimension indicates distinction between genres available in our dataset, which is not seen in the data labelled with text IDs only.

For the sake of the visualisation of results, we perform the same analyses labelling our dataset with genres, and also reducing it to subcorpora corresponding to different genres and languages (e.g. EO_ESS containing all texts of English political essays, etc.), see the resulting plot in Figure 2.

This time, we achieve a cumulative value of 82%, with the first dimension covering over 60% of the data variance, see Table 3.
tourism, essays and popular-scientific texts grouping together below zero (with additives, modality and general nouns as features), and fiction, political speeches and letters to shareholders above zero. The features of instruction manuals seem to be language-dependent, as the English and the German INS subcorpora are positioned on the opposite axis sides. Fictional texts of both languages are positioned at the edge of the genre axis, with personal pronouns contributing to this grouping, which coincides with the results obtained by Kunz and Lapshinova-Koltunski (2015) and Kunz et al. (2015) showing that fiction is best distinguished from the other genres for both languages with supervised classification techniques.

Automatic clustering deliver similar results, see Figure 3, with the exception of English fictional texts, which are classified along with the German fictional texts into the cluster of German subcorpora.

Figure 3: Classification of English and German subcorpora

### 4.2 Originals and translations

In the next step, we include translated texts into our analysis. The translation data is labelled with HU and MT, indicating manual or automatic method of translation, whereas digits indicate translation variants. Thus, MT1 and MT2 are produced with two different SMT systems, and HU1 and HU2 were produced by two different groups of translators. The results of the bootstrap resampling\(^5\) suggests two classes in our data, illustrated in Figure 4.

![Cluster dendrogram with AU/BP values (%)](image)

Figure 4: Classification of originals and translations

As seen from the graph, our dataset is clustered into originals (on the right side) and translations (on the left side), which is apparently the most prominent difference in this data. This coincides with the statements of the theory of translationese, see (Gellerstam, 1986) or (Baker, 1993), that translations have their specific feature differing them from the source texts and comparable originals in the target language. A number of studies have shown that these features can be used to automatically discriminate between translated and non-translated texts, such as (Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011). Our results show that this discrimination is also possible with discourse features, which means that translations differ from originals also in these properties.

The only exceptions in our results are manually produced translations of political speeches (HU2-SP) and instruction manuals (HU2-INS) classified together with political speeches and letters to shareholders originally written in German. Most of the smaller clusters within the bigger ‘non-translated’ class are grouped rather according to languages than genres, e.g. political essays, tourism texts, manuals and popular-scientific arti-

\(^5\) We achieve a good classification performance with an average error rate of 0.06.
tourism, essays and popular-scientific texts grouping together below zero (with additives, modality and general nouns as features), and fiction, political speeches and letters to shareholders above zero. The features of instruction manuals seem to be language-dependent, as the English and the German INS subcorpora are positioned on the opposite axis sides. Fictional texts of both languages are positioned at the edge of the genre axis, with personal pronouns contributing to this grouping, which coincides with the results obtained by Kunz and Lapshinova-Koltunski (2015) and Kunz et al. (2015) showing that fiction is best distinguished from the other genres for both languages with supervised classification techniques.

Automatic clustering deliver similar results, see Figure 3, with the exception of English fictional texts, which are classified along with the German fictional texts into the cluster of German subcorpora.

Figure 3: Classification of English and German subcorpora

4.2 Originals and translations

In the next step, we include translated texts into our analysis. The translation data is labelled with HU and MT, indicating manual or automatic method of translation, whereas digits indicate translation variants. Thus, MT1 and MT2 are produced with two different SMT systems, and HU1 and HU2 were produced by two different groups of translators. The results of the bootstrap resampling suggests two classes in our data, illustrated in Figure 4.

Figure 4: Classification of originals and translations

As seen from the graph, our dataset is clustered into originals (on the right side) and translations (on the left side), which is apparently the most prominent difference in this data. This coincides with the statements of the theory of translationese, see (Gellerstam, 1986) or (Baker, 1993), that translations have their specific feature differing them from the source texts and comparable originals in the target language. A number of studies have shown that these features can be used to automatically discriminate between translated and non-translated texts, such as (Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011). Our results show that this discrimination is also possible with discourse features, which means that translations differ from originals also in these properties.

The only exceptions in our results are manually produced translations of political speeches (HU2-SP) and instruction manuals (HU2-INS) classified together with political speeches and letters to shareholders originally written in German. Most of the smaller clusters within the bigger ‘non-translated’ class are grouped rather according to languages than genres, e.g. political essays, tourism texts, manuals and popular-scientific arti-

---

5. We achieve a good classification performance with an average error rate of 0.06.
Next, we want to prove if the observed difference between originals and translations is dependent on the source or the target language (which would indicate the phenomenon of shin- ing through or normalisation). For this reason, we perform two classification experiments applying the same clustering technique and including German translation data and their English sources in the first experiment (Figure 5), and the same German translations together with German comparable non-translated texts in the second (Figure 6). The results show that in both cases, the data is separated into translations and originals, with the same two subcorpora as exceptions. So, no shin- ing through/normalisation effect can be detected.

Figure 5: German translations and non-translated English source texts

Figure 6: German translations and comparable German non-translated texts

4.3 Human and machine translations

Finally, we perform classification on the data subset containing translations only. The resulting den- drogram in Figure 7 reveals four heterogeneous classes of translations, all containing both man- ually and automatically produced outputs. The two human translations that were classified with the non-translated data in previous experiments in Section 4.2 form a cluster on their own. This is the only cluster containing one type of translations in the whole data subset. The other three clusters consist of a mixture of human and machine trans- lations. They presumably form genre-sensitive

Figure 7: Human and machine translations

On the one hand, this suggests that genre is more prominent than translation method, i.e. there are more differences between various genres than between human and machine translations in the data under analysis, if discourse properties are concerned. On the other hand, the results may also indicate that discourse features are more in- formative in genre classification than in the di-
tinction into human vs. machine. Similar results were shown by Zampieri and Lapshinova-Koltunski (2015) who were able to achieve better results in the classification between genres than between translation methods, operating with delexicalised n-grams and using supervised classification techniques. Therefore, we claim that the distributions of the discourse features under analysis are genre-dependent, which coincides with the results of the previous analyses within a number of multilingual genre studies.

As seen in the analyses above (see Figures 4, 5, 6 and 7), political speeches and letters to shareholders are always clustered together in translated data. Similar observations were also made in (Lapshinova-Koltunski, inpress) for a different set of features. According to Neumann (2013), these two registers seem to be closer in English than in German, and so, their commonalities in our translation data might indicate the influence of the source texts. However, CA performed on German and English originals reveal that these register are similar not only within each language, but also cross-lingually, as they are situated on the same level of the y-axis, see Figure 2. As a result, translations also reveal these similarities.

5 Conclusion and Discussion

We have demonstrated an example of a corpus-based analysis of discourse properties in a multilingual dataset which contains both translated and non-translated texts, using exploratory and automatic clustering techniques. The results show that discourse-related features vary depending on the languages and genres involved. Languages, even such closely related ones as English and German, have different preferences in the usage of discourse properties, which are also prone to interlingual variation in terms of genres. This knowledge on contrasts will be valuable not only for contrastive linguistics and translation studies, but also for natural language processing including statistical MT, as it is available in form of frequency-based information and can be used for language models. The observed variation of discourse properties is also influenced by the nature of the texts (translated vs. non-translated). Both human and machine translations have constellations of discourse properties different from those of their underlying originals, and from comparable non-translated texts in the target language.

Comparing machine-translated texts with those translated by humans, we stated that genre-membership of translations determines more prominent differences between them than the methods they were translated with (manual vs. automatic). This points to the fact that machine translations resemble rather human translations than non-translated texts in both the source and the target languages, if discourse features are considered. On the one hand, this confirms the hypothesis of levelling out indicating that individual translated texts are more alike than individual original texts, in both source and target languages. On the other hand, our results conform to those obtained by Rabinovich and Wintner (2015) who show that multi-genre data is more difficult to be classified with translationese (translation-specific) features.

Furthermore, the results seem to contradict the findings in (Guzman et al., 2014), which used discourse information to develop automatic MT evaluation metrics. However, we believe that the differences in the outcome are caused by the nature of the dataset: translations in the present study originate from multiple genres, whereas Guzman et al. (2014) use news texts only. Intralingual variation in both English and German imply that if a model is applicable for a certain genre in one language, it is not necessarily applicable to a different genre of the same language, as the distributions of the underlying phenomena differ (sometimes) tremendously.

The contrasts between translated and non-translated texts suggest that we need more research on how to incorporate discourse-based language models induced from comparable and not parallel data. In this way, we might achieve a closer approximation of machine translation to non-translated texts in a target language. This is relevant not only for the development of machine translation systems but also for their evaluation, as the similarities between a reference and an MT output might be confounding in the quality judgement, if discourse phenomena are concerned. In the future, experiments could be planned that apply the present results for the development and evaluation of MT. Moreover, it would be interesting to learn if the differences between translated and original text affect perception of the quality of the text, for which experiments involving human judgements are required.

---

6 Variation in individual translators is not considered.
References


Sofia, Bulgaria, August. Association for Computational Linguistics.


Investigating Genre and Method Variation in Translation Using Text Classification

Marcos Zampieri¹,² and Ekaterina Lapshinova-Koltunski¹

¹Saarland University, Germany
²German Research Center for Artificial Intelligence (DFKI)

Abstract. In this paper, we propose the use of automatic text classification methods to analyse variation in English-German translations from both a quantitative and a qualitative perspective. The experiments described in this paper are carried out in two steps. We trained classifiers to 1) discriminate between different genres (fiction, political essays, etc.); and 2) identify the translation method (machine vs. human). Using semi-delexicalized models (excluding all nouns), we report results of up to 60.5% F-measure in distinguishing human and machine translations and 45.4% in discriminating between seven different genres. More than the classification performance itself, we argue that text classification methods can level out discriminative features of different variables (genres and translation methods) that thus enabling researchers to investigate in more detail the properties of each of them.

Keywords: human and machine translation, text classification, genres

1 Introduction

Text classification is an important area of research in Natural Language Processing (NLP) and it has been applied in a wide range of tasks such as spam detection [1] and temporal text classification [2]. From a purely engineering perspective, researchers are interested in how well classification methods can distinguish between two or more classes and what kind of features and algorithms deliver the best performance in each task. In recent work [3, 4], however, state-of-the-art text classification methods were proposed to investigate language variation across corpora. These methods were successfully applied in the identification of languages, varieties and dialects, as well as genres.

The present study is an attempt to use the same techniques for the identification of translation varieties – translations which differ in genres, e.g. essays, fiction, or methods, i.e. human and machine. We train classifiers to distinguish translated texts according to either their genre or method of translation, using the VARTRA corpus [5], a collection of English to German translations. More than the classification results per se, we use (semi-)delexicalized representations aiming to reduce topical bias and, therefore, levelling out interesting linguistic features that can be further used in linguistic analysis and NLP applications.
2 Related Work and Theoretical Background

Genre-specific variation of translation is related to studies within register and genre theory, e.g. [6], [7], which analyse contextual variation of languages. In lexico-grammatical terms, this variation is reflected in the distribution of linguistic patterns, i.e. subject/objects, evaluative patterns, negation, modal verbs, discourse phenomena (e.g. coreference or discourse markers).

Multilingual genre analysis is concerned with the distribution of such lexico-grammatical patterns not only across genres but also across languages, comparing the settings specific for the languages under analysis, e.g. [7] on English, Nukulaelae Tuvaluan, Korean and Somali, [8] and [9] on English and German. Moreover, the latter two also consider genres in translations. Applying a quantitative approach, Neumann (2013) [9] analyses an extensive set of features and shows to what degree translations are adapted to the requirements of different genres. Other scholars [10–13], also integrate register analysis in translation studies. However, they either do not account for distributions of these features, or analyse individual texts only. De Sutter et al. (2012) [14] and Delaere & De Sutter (2013) [15] in their analysis of translated Dutch also pay attention to genre variation, but concentrate on lexical features only.

Whereas attention is paid to genre settings in human translation analysis, they have not yet been considered much in machine translation. There exist some studies in the area of SMT evaluation, e.g. errors in translation of new domains [16]. However, the error types concern the lexical level only, as the authors operate solely with the notion of domain and not genre. Domains represent only one of the genre parameters and reflect what a text is about, i.e. its topic, and further settings are thus ignored. Although some NLP studies, e.g. those employing web resources, do argue for the importance of genre conventions, see e.g. Santini et al. (2010) [17], genre remains out of the focus of machine translation. In the studies on adding in-domain bilingual data to the training material of SMT systems [18] or on application of in-domain comparable corpora [19], again, only the notion of domain is taken into consideration.

Studies involving translation methods mostly focus on translation error analysis, and human translation serves usually as a reference in MT evaluation tasks. Some of them do consider linguistic properties, or linguistically-motivated errors [20, 21]. The latter one includes style errors, which is partly related to genre.

To our knowledge, the only study investigating differences between human and machine translation is Volansky et al. (2011) [22]. The authors analyse human and machine translations, as well as comparable non-translated texts. They use a range of features based on the theory of translationese (see [23] or [24]) expecting that the features specific for human translations can also be used to identify machine translation. Some of the translationese features were investigated using NLP techniques [25–27] similar to the ones we propose in this paper. What is most important for our study, however, is the claim by Volansky et al. (2011) [22] that some features of human translations coincide with those of machine-translated texts, whereas other features are diversifying between these two translation methods.
3 Methods

3.1 Data

For the purpose of our study we use VARTRA [5], a corpus of multiple translations from English into German. These translations were produced by: (1) human professionals (PT1), (2) human student translators (PT2), (3) a rule-based MT system (RBMT), (4) a statistical MT system trained with a large quantity of unknown data (SMT1) and (5) a statistical MT system trained with a small amount of data (SMT2). The genres available in VARTRA are: political essays (ESS), fictional texts (FIC), instruction manuals (INS), popular-scientific articles (POP), letters of share-holders (SHA), prepared political speeches (SPE), and touristic leaflets (TOU). Each subcorpus represents a translation variety, a translation setting which differs from all others in both method and genre (e.g. PT1-ESS or PT2-FIC, etc.).

Before classification was carried out, we split the corpus into sentences (of size between 12 and 24 tokens). This created 6,200 instances. The data was then split into a training (80%) and a test (20%) set.

The features used in different experiments include bag-of-words (bow), word bigrams, word trigrams and word 4-grams. The novelty of our approach is that we substitute all nouns with placeholders in some of the experiments. This results in what we call a semi-delexicalized text representation, which lies between fully delexicalized representations [3] and the classical bag-of-words or n-gram language models. Previous studies [4, 28] show that named entities significantly improve the result of text classification systems, so we decided to use this semi-delexicalized representation to minimize topic variation. The decision was motivated by both our goal of investigating translation variation influenced by both genre and method, and our aim to obtain a robust classification method that could perform well on different corpora.

3.2 Algorithms

We used two algorithms in our experiments. The first is a Naive Bayes (NB) classifier using bag-of-words as features. Naive Bayes classifiers work based on an independence assumption (the presence of a particular feature of a class is not related to the presence of any other feature), which is particularly useful for supervised learning and makes them extremely fast.

The second algorithm is based on a likelihood function calculated over n-gram language models as described by Zampieri and Gebre (2014) [29]. The language models can contain characters and words (e.g. bigrams and trigrams), linguistically motivated features such as parts-of-speech (POS) or morphological categories [4], or (semi-) delexicalized models such as the one we explore here.

4 Results

In this section, we present the results we obtained in different classification experiments. For the evaluation step we used standard NLP metrics such as
4. Zampieri, Lapshinova-Koltunski

Precision, Recall and F-Measure. The linguistic analysis and discussion of the most important differences between both method and genre variation will be presented later in section 4.5.

4.1 Genres and Methods

The first experiment shows why it is important to use semi-delexicalized features in a dataset that represents both dimensions of variation in translation (genre and method). The question posed at this stage is simple: how different are the samples with respect to methods and genres? We use the aforementioned Naive Bayes classifier trained on (non-delexicalized) bag-of-words. In Table 1 we present the results as well as a baseline computed based on the random assignment of all documents to a particular class.

<table>
<thead>
<tr>
<th>Type</th>
<th>Classes Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genres</td>
<td>7</td>
<td>57.4%</td>
<td>57.8%</td>
<td>57.3%</td>
</tr>
<tr>
<td>Methods</td>
<td>5</td>
<td>35.9%</td>
<td>36.2%</td>
<td>35.3%</td>
</tr>
</tbody>
</table>

Table 1. Naive Bayes: Genres and Methods

The results of this preliminary experiment show that the classifier was able to distinguish between the seven translation genres with up to 57.3% F-Measure and between the five translation methods with up to 35.3% f-measure. The method is aided by named entities and content words that are domain specific and, therefore, influence the performance of the classifier. Therefore, we use placeholders to substitute nouns (both named entities and common nouns) to minimize topical bias in the following experiments. At the same time, the results of the present experiment will allow us to compare classification performance of non-delexicalized vs. semi-delexicalized representations.

4.2 Translation Methods

In the next experiment, we take a closer look at the differences between five methods of translation (PT1, PT2, RBMT, SMT1 and SMT2) while minimizing topic influence, i.e. trying to distinguish between them excluding all nouns.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>35.1%</td>
<td>35.9%</td>
<td>34.9%</td>
<td>20.0%</td>
</tr>
<tr>
<td>4</td>
<td>43.2%</td>
<td>44.9%</td>
<td>43.1%</td>
<td>25.0%</td>
</tr>
</tbody>
</table>

Table 2. Naive Bayes: Translation Method

The data contains outputs of two different SMT systems and in this step, we decide to merge them into a unique class of SMT. This was mainly done to answer the question of whether this kind of distinction is meaningful in practical terms, and whether the outputs of SMT1 and SMT2 are significantly different.
The results (presented in Table 2) improved substantially after the grouping. In the five-class setting the f-measure obtained for class SMT1 was the lowest of all 26.4%, whereas the SMT class could obtain the best result in the four-class setting (58.5%). This indicates that the outputs of both systems contain similar features.

4.3 Different Genres: Different Language?

In the next step, we try to automatically distinguish between different genres represented in the dataset. For this experiment, we also use the semi-delexicalized features.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>45.5%</td>
<td>46.1%</td>
<td>45.4%</td>
<td>14.2%</td>
</tr>
</tbody>
</table>

Table 3. Naive Bayes: Genres in Translation

As seen from Table 3, the automatic distinction between the seven genres achieved ca. 45% of both Precision and F-measure. This performance is substantially above the 14.2% baseline which indicates that the genres in VARTRA are essentially different. However, we are also interested in whether the classifier’s performance for genre discrimination was consistent across all translation methods. For this step, we perform genre classification within each translation method (with both SMT outputs in one class) and the result can be seen in figure 1.

![Fig. 1. Genre Distinction Across Method](image_url)

The performance across all seven genres is constant regardless of the translation method applied. For example, instruction manuals (INS) followed by fiction (FIC) are the easiest genre to identify in all four translation methods, whereas
speech (SPE) and essays (ESS) are consistently regarded as the most problematic ones. All the results are significantly higher than the expected baseline accuracy. The nominal values used to generate figure 1 are presented in table 4 on a scale from 0 to 1 with three decimal digits. The baseline we consider is once again the majority class, 14.2% f-measure.

<table>
<thead>
<tr>
<th>Method</th>
<th>ESS</th>
<th>FIC</th>
<th>INS</th>
<th>POP</th>
<th>TOU</th>
<th>SPE</th>
<th>SHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT2</td>
<td>0.399</td>
<td>0.533</td>
<td>0.595</td>
<td>0.372</td>
<td>0.421</td>
<td>0.346</td>
<td>0.536</td>
</tr>
<tr>
<td>PT1</td>
<td>0.314</td>
<td>0.606</td>
<td>0.664</td>
<td>0.456</td>
<td>0.425</td>
<td>0.371</td>
<td>0.507</td>
</tr>
<tr>
<td>RBMT</td>
<td>0.397</td>
<td>0.536</td>
<td>0.632</td>
<td>0.411</td>
<td>0.440</td>
<td>0.320</td>
<td>0.515</td>
</tr>
<tr>
<td>SMT</td>
<td>0.394</td>
<td>0.503</td>
<td>0.630</td>
<td>0.455</td>
<td>0.460</td>
<td>0.408</td>
<td>0.505</td>
</tr>
</tbody>
</table>

Table 4. Genre Distinction Across Method

4.4 Human vs. Machine

In the last experiment, we investigate whether the differences between the four translation methods are weaker than between a less fine-grained classification into human and machine translation. For this step, we unify PT1 and PT2 into one class, and RBMT and SMT1 and SMT2 into the other. We also tested different sets of semi-delexicalized features, i.e. bigrams, trigrams and 4-grams to find out which allow the best classification results, see Table 5. In all three scenarios, the model performs above the expected baseline of 50.0% F-measure. The best performance, however, is obtained for the trigram model (60.5% f-measure and 61.1% precision).

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>bigrams</td>
<td>53.3%</td>
<td>53.3%</td>
<td>53.3%</td>
<td>50.0%</td>
</tr>
<tr>
<td>trigrams</td>
<td>61.1%</td>
<td>60.0%</td>
<td>60.5%</td>
<td>50.0%</td>
</tr>
<tr>
<td>4-grams</td>
<td>55.2%</td>
<td>54.2%</td>
<td>54.7%</td>
<td>50.0%</td>
</tr>
</tbody>
</table>

Table 5. N-grams: Human x Machine

As the amount of training data is not large, from 4-grams onwards the method seems to suffer from data sparsity and as can be expected, performance drops.

4.5 Feature Analysis

This section aims to identify the most informative features from the semi-delexicalized n-grams in our experiments. This step is manual and carried out by looking through the most informative features and thus discriminative for certain genres and methods in our translation data. We evaluated the trigrams, as the performance of trigram models achieved the best results in the classification task. The list of the features specific either to human or machine translation is shown in Table 6. Using the same strategy, we generate a list of features discriminating genre pairs (for the sake of space, we display political essays and fictional texts) in Table 7.
Genre and Method Variation in Translation

<table>
<thead>
<tr>
<th>human</th>
<th>machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>full nominal phrase</td>
<td>full nominal phrase</td>
</tr>
<tr>
<td>(with def./indef. modif.)</td>
<td>(with def./indef./poss. modif.)</td>
</tr>
<tr>
<td>personal reference</td>
<td>personal reference</td>
</tr>
<tr>
<td>(1st pers. plural)</td>
<td>(1st pers. sg)</td>
</tr>
<tr>
<td>extended reference (demonstr.)</td>
<td>extended reference (pers.)</td>
</tr>
<tr>
<td>prepositional phrase</td>
<td>prepositional phrase</td>
</tr>
<tr>
<td>with local meaning</td>
<td>with different meanings</td>
</tr>
<tr>
<td>discourse markers</td>
<td>discourse markers</td>
</tr>
<tr>
<td>with additive meaning</td>
<td>with adversative meaning</td>
</tr>
</tbody>
</table>

Table 6. Features discriminating between human and machine translations

<table>
<thead>
<tr>
<th>ESS</th>
<th>FIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>passive constructions</td>
<td>active verbs</td>
</tr>
<tr>
<td>modal verbs with the meaning of volition and obligation</td>
<td>to-infinitives</td>
</tr>
<tr>
<td>to-infinitives</td>
<td></td>
</tr>
<tr>
<td>prepositional phrase</td>
<td>predicative adjectives</td>
</tr>
<tr>
<td>demonstrative reference</td>
<td>personal reference</td>
</tr>
<tr>
<td>discourse markers</td>
<td>discourse markers</td>
</tr>
<tr>
<td>with additive meaning</td>
<td>with adversative meaning</td>
</tr>
</tbody>
</table>

Table 7. Features discriminating between political essays and fictional texts

Semi-delexicalized trigrams consist of a sequence of words and placeholders, e.g., können PLH PLH, zu erfüllen hat, das PLH, aber, etc. Intuitively, we try to recognize more abstract categories, i.e., modal verbs with the meaning of possibility, infinitive clauses, discourse markers with adversative function for the given trigrams. As seen from the lists, both translation methods have similar discriminating features, i.e., full nominal phrases, coreferring expressions, prepositional phrases and discourse markers. However, the differences between them can be identified on a more fine-grained level: if we take into account morphological preferences and the scope of referring expressions, the meaning of prepositional phrases and discourse markers. All these phenomena seem to be related to participants and structures of textual discourse.

The features that turn out to be specific for genres include verbs and verbal constructions, further types of phrases, and also different types of coreferring expressions and discourse markers. Genre-discriminating features are also, as in case of methods, on a more fine-grained level. However, the level of description is not on morphological, but rather on syntactic level (active vs. passive, prepositional vs. adjectival phrases). Moreover, they describe rather processes than participants of discourse. The last features coincide in both tables (additive vs. adversative construction), which means that they are informative in both genre and method classification.

Our preliminary observations on features coincide with the results of empirical analyses on genres, e.g., those obtained by Neumann (2013) [9]. For instance,
the author point to personal pronouns, predicative adjectives, mental and verbal processes as indicators of narration and casual style which are specific for fictional texts. Political essays, which are characterized as expository texts with rather neutral style, contain relational processes, verbs of declarative mood, frequent nominalisations and almost no personal pronouns.

We believe that we need a more detailed analysis of the resulting features to have firm basis to build upon in our final conclusions on the features. For instance, the definition of mood and tense of verbal phrases, as well as their membership in a certain semantic verb class would contribute to a better specification of genres. Moreover, this step can be automatized with the help of existing morphological tools, taggers and wordnets, which is however, beyond the scope of the present paper.

The resulting lists of features can be beneficial for not only genre classification task but also for machine translation task, as they can help to automatically differentiate between human and machine translation.

5 Conclusion and Outlook

This paper is, to our knowledge, the first attempt to use text classification techniques to discriminate methods and genres in translations and to identify their specific features and relevant systemic differences in a single study. We report results of up to 60.5% F-measure in distinguishing human and machine translations and 45.4% in discriminating between seven different genres.

We used different algorithms and sets of features to study variation in English-German translation data. For that we used not only the classical bag-of-words and n-gram language models but also the use of (semi-)delexicalized representations along with classical bag-of-words and n-gram language models, which helps us to decrease the thematic bias in classification. The aim was both the discrimination of methods and genres per se, and also the identification of relevant systemic differences across genres and methods of translation.

The results of our analysis can find application in both human and machine translation. In the first case, they deliver valuable knowledge on the translation product, which is influenced by the methods used in the process and the context of text production expressed by the genre. In case of machine translation, the results will provide a method to automatically identify genres in translation data thus helping to separate out-of-genre data from a training corpus.

The aforementioned practical applications of the results are part of our future work, which will also include tests with other classification algorithms such as the popular support vector machines [30] used in Petrenz and Webber (2012) for a similar task [31]. We also plan to automate the generation of more abstract categories for the informative features as well as to experiments other kinds of de-lexicalized representations such as the one used by Quinio et al. (2012) [32]. Finally, we would like carry out further and more detailed linguistic analysis.
References


10. Zampieri, Lapshinova-Koltunski


Interaction of Dimensions and their Implications
Exploratory analysis of dimensions influencing variation in translation: the case of text register and translation method

Ekaterina Lapshinova-Koltunski

Abstract

The present study investigates the interplay between two dimensions influencing translation: text register and translation method. This is achieved by a corpus-based analysis which involves extraction of specific linguistic features occurring in multiple translations of the same texts. These translations differ, on the one hand, in registers the texts belong to, and on the other hand, in the translation method applied (human vs. machine translation). Our analysis is based on two frameworks – register theory and corpus-based translation studies – which also serve as sources for the definition of the features under analysis. Our quantitative analysis is supported with statistical methods. Unsupervised techniques are used to trace the degree of variation caused by the two dimensions, and also to identify the dimension having a greater impact on the translations. The results of our analysis shed light on the main factors influencing translation, and also deliver explanations for translation errors. In addition, further factors affecting linguistic features of translations, e.g. the experience involved, are traced in the present analysis. In this way, the study contributes to a better understanding of both translation product and translation process, and provides information which is useful for both improvement and evaluation of translation.

1 Research goals and motivation

In the present study, we analyse the interplay between two dimensions influencing variation in translation: translation methods (human and machine translation) and text registers (e.g. fiction, political speeches). Our starting assumption is that the interplay between translation method and text register is reflected in the lexicogrammar of translated texts. As shown by Neumann (2013), translations are influenced by both language and context of situation (i.e. register a text belongs to). Linguistic features of trans-
lations vary according to these different dimensions. We believe, however, that there are more dimensions at play in translation than has been presumed so far. For example, due to recent developments in translation-oriented language technologies, translations are increasingly produced not only by human translators, but also by machine translation systems. There are also mixed forms of translations, such as computer-aided human translation or post-edited machine translation, with more classes within each subtype (e.g. human translations assisted by different tools, such as translation memories or terminology databases, or produced by experienced vs. inexperienced translators, with rule-based or statistical machine translation systems). Our assumption here is that translation types produced by/results resulting from these different translation methods constitute another possible context of variation for translation, alongside language and register. We call these translation subtypes translation varieties. To date, variation along the third parameter (translation method) has not received much attention in studies devoted to translation. Some studies that address both human and machine translations (cf. Babych and Hartley 2004; Papineni et al. 2002; Popović 2011; White 1994) focus solely on translation error analysis, using human translation as a reference in the evaluation of machine translation outputs. Only a few of them operate with linguistically-motivated categories (e.g. Fishel et al. 2012; Popović 2011), but once again, these categories are used to detect errors in translations. To our knowledge, the only study dealing with the differentiation between human and machine translation is Volansky et al. (2011). They derive the features they analyse from corpus-based translation studies and try to detect those which are common to both translation varieties (e.g. contextual function words, part-of-speech patterns) and those that differentiate them (average sentence length, passive verb ratio, etc.). Their dataset contains newspaper articles only, and hence, cannot be used to reveal variation along the parameter of register.

Our primary interest is in linguistic features of translation varieties, such as active vs. passive verb constructions, preferences for certain functional verb classes, modality meanings, proportion of nominal vs. verbal phrases. Our assumption is that they reflect the interplay of the dimensions described above: language-specific (both source and target language), register-specific, as well as translation method-specific. In this study, we focus on the variation affected by register and translation method only. The reason for that is the dataset we have at our disposal which consists of English-to-German translations only. As shown in various studies (see Teich 2003 or Neumann 2013 among others), variation in translation is realised
by different linguistic phenomena which are situated at various linguistic levels (morpho-syntax, lexis and text). This study will, therefore, investigate quantitative distributions of linguistic features reflected in the lexicogrammar of texts. The comparison of the distribution of these features across translation varieties will provide insights into the interplay of the two dimensions under analysis, which is expected to be traced in their correlation.

We believe that this type of analysis enables general statements based on larger quantities of data instead of individual texts, as in Neumann (2013). Quantitative analysis requires corpus-based methods and involves classification and counting of features, as well as their statistical validation, see McEnery and Wilson (2001) and Biber et al. (1998). Our objective is to shed light on linguistic properties of translated texts produced by both humans and machines with a view to arriving at a better understanding of the translation product, and, to some extent, of the translation process. We also aim at showing the usefulness of the results obtained in this study for the fields of translation evaluation and quality estimation. Although registers have been taken into account in several studies on translation quality evaluation in human translation, they have remained under-researched in the field of machine translation.

2 Related work

As already stated in section 1, translations may vary according to different parameters, e.g. language, register and translation method. The present section examines studies that deal with variation in translation along all three parameters. These studies serve as important sources for the aggregation of linguistic features required for a corpus-based analysis.

2.1 Language variation and translation

Studies on translation variation along the language dimension concentrate on differences between translations and non-translated texts in both source and target languages (Hansen 2003; House 2014; Matthiessen 2001; Steiner 2004; Teich 2003). Most of them are related to the notion of translationese, invoked by Gellerstam (1986) to describe differences between original and translated texts. Baker (1993, 1995) elaborates specific features of translations that are believed to be universal, irrespective of the
source language and holding for any target language. These universals include explicitation, a tendency to spell things out rather than leave them implicit; simplification, a tendency to simplify the language used in translation; normalisation, a tendency to exaggerate features of the target language and to conform to its typical patterns; and levelling out, similarity of individual texts among each other in a set of translations if compared to individual texts in a set of originals. For the latter, we prefer to use the term convergence, which means a relatively higher level of homogeneity of translated texts with regard to their scores of lexical density, sentence length, etc. Another translation feature which is not explicitly analysed by Baker is Toury’s law of interference (Toury 1995). We prefer to use the term shining through, features of the source texts observed in translations, see Teich (2003).

In some recent approaches, machine learning techniques have been applied to the analysis of translationese. Here, text classification is used to identify translation features or to differentiate between originals and translations. For instance, Baroni and Bernardini (2006) analyse the features of Italian translations on the basis of monolingual comparable corpora (translations into Italian from a number of languages, including English, Arabic, French, Spanish and Russian, as well as their comparable non-translated originals) using machine learning techniques for text categorisation. Their work shows that it is possible to distinguish translations from originals automatically. Ilisei et al. (2010) differentiate between Spanish translated from English and non-translated texts with the help of simplification feature measured with average sentence length, proportion of simple and complex sentences, lexical richness, etc. Another example is Kurokawa et al. (2009) which investigates the possibility of automatically distinguishing whether a piece of text is an original text or a translation using word and part-of-speech n-grams. They are especially interested in the implication of their work for machine translation performance, but do not consider the identification of human- vs. machine-translated texts (see also Lembersky et al. 2012).

However, none of the above-mentioned studies compares registers, although Baroni and Bernardini (2006) do point out the importance of text registers. They acknowledge that the use of a comparable corpus representing one specific register is a drawback of their experiment.
2.2 Register variation and translation

Studies within register and genre theory, e.g. Quirk et al. (1985), Halliday and Hasan (1989), Biber (1995), analyse contextual variation in languages. In their terms, languages vary with respect to usage context. These contexts influence the distribution of particular lexico-grammatical patterns which manifest language registers. The canonical view is that situations can be characterised by the parameters of field, tenor and mode of discourse. Field of discourse relates to processes and participants (e.g. Actor, Goal, Medium), as well as circumstantial (Time, Place, Manner, etc.) and is realised in lexico-grammar in lexis and colligation (e.g. argument structure). Tenor of discourse relates to roles and attitudes of participants, author-reader relationship, and is reflected in stance expressions or modality. Mode of discourse relates to the role of the language in the interaction and is linguistically reflected at the grammatical level in Theme-Rheme constellations, as well as cohesive relations at the textual level. In other words, the contextual parameters of registers correspond to sets of specific lexico-grammatical features, and variation across different registers can be seen in the distribution of these features which are expressed in lexico-grammatical patterns.

Multilingual register studies concern linguistic variation, i.e. the distribution of lexico-grammatical features, not only across registers but also across languages, comparing the settings specific for the languages under analysis, e.g. Biber (1995) on English, Nukulaelae Tuvaluan, Korean and Somali, Hansen-Schirra et al. (2012) and Neumann (2013) on English and German. In addition, the latter two also consider register analysis in translations. Some other scholars, e.g. House (1997) and Steiner (1996, 1998, 2004), also integrate register analysis in translation studies. However, they either do not account for distributions of these features, or analyse individual texts only. Likewise, De Sutter et al. (2012) and Delaere and De Sutter (2013), in their analysis of translated Dutch, pay attention to register variation, but concentrate on lexical features only.

Applying a quantitative approach, Neumann (2013) analyses an extensive set of features and shows the degree to which translations are adapted to the requirements of different registers, thereby detecting a further dimension to the study of translation properties and showing how both register and language typology are at work.
2.3 Variation in translation method

As previously mentioned (see section 1), studies involving both human and machine translation mostly focus on translation error analysis, i.e. automatic error detection. Human translation usually serves as a reference for the comparison of different machine translation outputs. However, some of them do consider linguistic properties, or linguistically-motivated errors, see for instance Popović (2011), Popović and Burchardt (2011) or Fishel et al. (2012). The latter operate with such features as missing words (content vs. grammatical), incorrect words (incorrect disambiguation, wrong lexical choice, etc.), wrong word order, etc. This classification also includes style errors, which is partly related to register. Yet, this error type is analysed on the level of words only, and thus, further register settings involved, e.g. part-of-speech classes, passive vs. active verb construction are not taken into account. So, none of these studies provides a comprehensive descriptive analysis of specific features of different translation methods.

El-Haj et al. (2014) compare translation style and consistency in human and machine translations of Camus’s novel *The Stranger*, which was translated from French into English and Arabic. They use readability as a proxy for style, and then measure how it varies within and between translations. Their work, however, has evaluative and not descriptive character, as they aim at quality evaluation of both human and machine translation.

To our knowledge, the only study describing linguistic differences between human and machine translation is Volansky et al. (2011). The authors analyse human and machine translations, as well as comparable non-translated texts. They use a range of features based on the theory of translationese (simplification, explicitation, normalisation and shining through), expecting that the features specific to human translations can also be used to identify machine translation. The authors claim that there are features of human translations which coincide with those of machine-translated texts, whereas other features differ across these two translation methods (see section 1 above), and the combination of their methodology and translationese features can reveal similarities and dissimilarities between human and machine translations.
3 Linguistic features, data and methodology

3.1 Features under analysis

For our analysis, we selected a set of features derived from the studies on language, register and translation method variation described in section 2. These features represent lexico-grammatical patterns of more abstract concepts, e.g. textual cohesion expressed via pronominal or nominal reference, evaluative patterns expressed via certain syntactic constructions. The selected features were chosen because they reflect linguistic characteristics of all texts under analysis, are content-independent (do not contain terminology or keywords), and are easy to interpret, thereby yielding insights into the differences between the variables under analysis (cf. Volansky et al. 2011). As a result, some of the features analysed in the studies mentioned above were excluded from our analysis. For instance, no token n-grams were used here, as they are rather content-dependent, and reflect domains. In addition, as we have multiple translations of the same texts in our data, this kind of feature is not suitable for our analysis. We also used groupings of nominal and verbal phrases (based on chunk annotations) rather than part-of-speech n-grams, as they are easier to interpret than n-grams.

The set of selected features used in our analysis is outlined in Table 1. The first column shows the lexico-grammatical patterns that we extracted for the quantitative analysis. The second column presents the correspondence of these patterns to the context parameters (field, tenor and mode of discourse) (see section 2.2). The third column links the lexico-grammatical patterns with the translation features outlined in section 2.1.

Content words and their proportion to the total number of words in a text (row 1) represent lexical density. This corresponds to the mode parameter in register theory and simplification in translationese studies (lexical richness of translations). The number of nominal and verbal parts-of-speech, as well as their groupings into nominal and verbal phrases or chunks (row 2) reflect participants in the field parameter, shining through and normalisation (as languages use different grammatical structures, which is reflected in translations, and English tends to be more verbal than German; see Steiner 2012). For the same reasons, field and shining through / normalisation can also be analysed via the distribution of nominalisations (ung-nominalisations in row 3). Reference expressed either in nominal phrases or in pronouns (row 4) reflects textual cohesion in the parameter of mode. From the point of view of translationese studies, this feature can...
point to explicitation, as pronouns are less explicit than nouns or nominal phrases. Moreover, preferences for personal or demonstrative pronouns in different languages (in our case English and German) can be reflected in shining through and normalisation. The distribution of abstract or general nouns and their comparison to other nouns (row 5) gives information about lexical choices (parameter of field) and preferences for more concrete or abstract words in translations (explicitation). Conjunctions (including both grammatical conjuncts such as \textit{und} – \textit{and}, \textit{aber} – \textit{but} and multiword expressions like \textit{aus diesem Grund} – \textit{that is why}) for which we analyse distributions of logico-semantic relations (row 6), belong to the parameter of mode as they express cohesion, and at the same time to the explicitation feature, as they explicitly mark relations in discourse.

Table 1. Features under analysis.

<table>
<thead>
<tr>
<th></th>
<th>lexico-grammatical patterns</th>
<th>register analysis</th>
<th>Translationese studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>content vs. total words</td>
<td>mode</td>
<td>simplification</td>
</tr>
<tr>
<td>2</td>
<td>nominal vs. verbal parts-of-speech and phrases (np.chunk, vp.chunk)</td>
<td>field</td>
<td>shining through / normalisation</td>
</tr>
<tr>
<td>3</td>
<td>\textit{ung}-nominalisation (ungnom)</td>
<td>field</td>
<td>shining through / normalisation</td>
</tr>
<tr>
<td>4</td>
<td>nominal (all.np) vs. pronominal (pronnp) and demonstrative vs. personal reference (perspron, dempron)</td>
<td>mode</td>
<td>explicitation, shining through / normalisation</td>
</tr>
<tr>
<td>5</td>
<td>abstract or general nouns (gen.-nouns) vs. all other nouns</td>
<td>field</td>
<td>explicitation</td>
</tr>
<tr>
<td>6</td>
<td>logico-semantic relations: additive, adversative, causal, temporal, modal</td>
<td>mode</td>
<td>explicitation</td>
</tr>
<tr>
<td>7</td>
<td>modality: obligation, permission, volition</td>
<td>tenor</td>
<td>shining through / normalisation</td>
</tr>
<tr>
<td>8</td>
<td>evaluation patterns</td>
<td>tenor</td>
<td>shining through / normalisation</td>
</tr>
</tbody>
</table>

Modal verbs, e.g. \textit{können} – \textit{can}, \textit{müssen} – \textit{must} (row 7) express modality, i.e. the parameter of tenor, are grouped according to different meanings, and also reveal contrasts in languages, as described by Teich (2003) and König and Gast (2012) for differences between English and German. That is why they their distribution in translation also reflects normalisation and shining through. Similarly, these phenomena (tenor and shining through /
Exploratory analysis of dimensions influencing variation in translation

normalisation) are reflected in evaluation patterns (e.g. *es ist interessant/wichtig zu wissen... - it is interesting/import to know*, row 8) which may have different distributions in different languages.

Information on the structural properties of the features, including examples illustrating subcategories, is presented in section 3.3 below.

3.2 Corpus data

To our knowledge, the only corpus resource suitable for our research agenda is VARTRA-SMALL (cf. Lapshinova-Koltunski 2013). This corpus contains different translation varieties, the texts of which are translated from English into German. The translations included in VARTRA-SMALL were produced with the following methods: by 1) professional humans (PHT, Professional Human Translation) and 2) student translators (SHT, Student Human Translation); as well as with MT systems: 3) a rule-based MT system (RBMT, Rule-Based Machine Translation) and two statistical MT systems – 4) Google Statistical Machine Translation (GSMT) and 5) Moses Statistical Machine Translation (MSMT). The dataset contains multiple translations of the same texts, which vary both in translation method and text register, and thereby represent different translation varieties (as defined above).

Translations by professionals (PHT) were exported from the already existing CroCo corpus (Hansen-Schirra et al. 2012). The SHT variant was produced by student translators with at least a BA degree, who have little or no experience in translating. Translators in the SHT production were assisted with different translation memories (available in the OPUS collection at http://opus.lingfil.uu.se), used with the help of Across, a computer-aided translation tool which can be integrated into the usual work environment of a translator. We did not time the translation tasks, and did not take into account other settings of the translation process, e.g. time spent on terminology search. The translation task was not part of any examination, and the students could freely make decisions on the use of additional reference resources.

The rule-based machine translation variant was produced with SYS-TRAN, whereas for statistical machine translation we used two systems – Google Translate (GSMT), and the in-house Moses-based system (MSMT). MSMT was trained with EUROPARL, a parallel corpus containing texts from the proceedings of the European parliament, cf. Koehn (2005). This training set is one of the largest existing datasets which are freely available.
Moreover, this corpus is used in most studies on machine translation. The decision to include two SMT systems is justified by the fact that the first one (GSMT) is trained with enormous data Google has at its disposal, whereas MSMT is trained with a parallel corpus, which is smaller than the data available at Google, and has a register restriction (as it contains political speeches only). In this way, we have two translation varieties displaying shortages in experience or data: inexperienced translators in SHT and insufficient training data in MSMT.

Each translation subcorpus contains translations of the same texts which cover seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to share-holders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The corpus offers translations which vary not only in the method by which they were produced, but also in the register to which they belong. It should be noted that some of these registers represent a continuum between written and spoken dimensions, i.e. SPEECH, which is written-to-be-spoken, and FICTION, which, because of the dialogues it contains, is on the border between spoken and written texts. The dataset contains both frequently machine-translated texts, e.g. political speeches, and those which are usually not translated with MT systems, such as fiction. All translation variants in VARTRA-SMALL comprise ca. 600,000 tokens. All subcorpora are tokenised, lemmatised, and tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations in VARTRA-SMALL were obtained with Tree Tagger (Schmid 1994).

The subcorpora are encoded in the CWB format (CWB, 2010) and can be queried with the help of CQP regular expressions, the syntax of which is described in Evert (2005).

3.3 Feature extraction

As mentioned in 3.2, VARTRA can be queried with CQP, which allows the definition of language patterns in the form of regular expressions based on string, part-of-speech and chunk tags, as well as further constraints. Table 2 outlines a number of examples for the queries used in our corpus. For instance, query 1 is used to differentiate between personal and demonstrative reference. Here, we simply search for personal or demonstrative pronouns and make use of the part-of-speech annotation in our corpus. In the second query, we also add lexical information, reducing our search to
items which are tagged as nouns and end with -\textit{ung}. This query is used to extract German \textit{ung}-nominalisations. The third query is even more restricted, as we reduce our search to certain lexical items, such as modal verbs. We utilise this restriction to classify between different modal meanings. The classification of different logico-semantic relations expressed via conjunctions is achieved with the help of query 4. Here, we do not use part-of-speech annotation, as we are interested not only in grammatical conjunctions, like \textit{und} (and), \textit{oder} (or), \textit{deswegen} (therefore), but also in multiword expressions such as \textit{darüber hinaus} (in addition) and \textit{aufgrund dessen} (that is why).

Manually compiled lexical lists are used to extract these items (ranging from ca. 80 to ca. 100 types per list). The lists of conjunctions expressing logico-semantic relations were derived from Lapshinova and Kunz (2014), who describe a procedure to semi-automatically annotate these relations in a multilingual corpus.

Table 2. Queries for feature extraction.

<table>
<thead>
<tr>
<th>feature category</th>
<th>CQP query</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 personal</td>
<td>[pos=&quot;PP.*&quot;]</td>
</tr>
<tr>
<td>2 demonstrative</td>
<td>[pos=&quot;PD.*&quot;]</td>
</tr>
<tr>
<td>3 \textit{ung}-nominalisation</td>
<td>[pos=&quot;NN.*&quot;&amp;lemma=&quot;).<em>ung.</em>&quot;]</td>
</tr>
<tr>
<td>4 obligation</td>
<td>[pos=&quot;VM.*&quot;&amp;lemma=&quot;müssen</td>
</tr>
<tr>
<td>5 permission</td>
<td>[pos=&quot;VM.*&quot;&amp;lemma=&quot;können</td>
</tr>
<tr>
<td>6 volition</td>
<td>[pos=&quot;VM.*&quot;&amp;lemma=&quot;wollen</td>
</tr>
<tr>
<td>4 additive</td>
<td>$additive-conjunction</td>
</tr>
<tr>
<td>5 adversative</td>
<td>$adversative-conjunction</td>
</tr>
<tr>
<td>4 causal</td>
<td>$causal-conjunction</td>
</tr>
<tr>
<td>5 temporal</td>
<td>$temporal-conjunction</td>
</tr>
<tr>
<td>5 modal</td>
<td>$modal-conjunction</td>
</tr>
<tr>
<td>5 abstract nouns</td>
<td>[pos=&quot;NN.*&quot;&amp;lemma=$abstract_nouns&quot;]</td>
</tr>
</tbody>
</table>
| 6 evaluation patterns | "es|Es[pos="VAFIN"]|posließ|[$.$,]{0,3} [pos="AD.*"] | [?"dass|ß|"s]
| 6 evaluation patterns | "es|Es[pos="VAFIN"]|posließ|[$.$,]{0,3} [pos="AD.*"] | [?"zu|wenn|für"]
| 6 evaluation patterns | "(A|a)m" [pos="AD.*"\&word=".*ste.*"] |

In some cases, e.g. for the extraction of abstract nouns, we also add a part-of-speech restriction, as shown in query 5. For evaluation patterns (derived from pattern grammar by Francis and Hunston 2000), we need more morpho-syntactic restrictions. These include sequences of parts-of-speech and lexical elements to extract seven evaluation patterns (see Table 3). For
the time being, we do not classify them according to their evaluative meaning, simply considering the amount of evaluation in translation varieties.

Table 3. Evaluation patterns.

<table>
<thead>
<tr>
<th>evaluation pattern</th>
<th>German example</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>es+Verb BE+ADJ+,+dass</td>
<td>es ist erforderlich, dass</td>
<td>it is required that</td>
</tr>
<tr>
<td>Verb BE+es+ADJ+,+dass</td>
<td>ist es erforderlich, dass</td>
<td>is it required that</td>
</tr>
<tr>
<td>es+Verb BE+ADJ+,+zu/wenn/für</td>
<td>es ist besser, wenn</td>
<td>it is better if</td>
</tr>
<tr>
<td>Verb BE+es+ADJ+,+zu/wenn/für</td>
<td>ist es besser, wenn</td>
<td>is it better if</td>
</tr>
<tr>
<td>Verb machen+es+ADJ</td>
<td>machen es schwerer</td>
<td>to make it more difficult</td>
</tr>
<tr>
<td>es+ADJ+Verb machen</td>
<td>es schwerer machen</td>
<td>to make it more difficult</td>
</tr>
<tr>
<td>am+ADJsuperlative</td>
<td>am wichtigsten ist dabei</td>
<td>the most important thing here</td>
</tr>
</tbody>
</table>

CQP facilities allow us to count the extraction results and sort them along the texts, registers and varieties they occur in. The extracted frequencies of these features are saved in a matrix for further validation in R (version 3.0.2; R Core Team 2013).

3.4 Statistical analysis

For our analysis, we use an unsupervised technique – hierarchical cluster analysis (HCA), see Baayen (2008) and Everitt et al. (2001). With the help of this technique, we can discover differences and similarities between translation varieties (subcorpora) under analysis.

Unsupervised data analysis, which is sometimes called knowledge discovery (cf. Murphy 2012), allows us to discover ‘interesting structures’ in the data. In our case, we are looking for structures in the form of translation clusters which are formed according to different dimensions, i.e. translation method and register. Whereas in a supervised case, we would define the clusters and the features specific to them, in an unsupervised case, we are free to choose the number of clusters as we like. Applying this technique, we hope to trace the interplay of variation dimensions in our data, to see which of them has a greater impact on the clustering of translations, and also to discover further structures in our data.
In hierarchical cluster analysis, a set of dissimilarities for the \( n \) objects being clustered is used. These dissimilarities are calculated by the Euclidean distance, i.e. distance between datasets, in our case between the subcorpora under analysis. Euclidean distance is one of the most straightforward and generally accepted ways of computing distances between objects in a multi-dimensional space.

We employed the complete-linkage method to perform clustering, as it yields compact clusters of approximately equal diameter. According to this method, the similarity between two clusters is the “worst-case” similarity between any pair of subcorpora in the two clusters. This means that at each step, the two clusters separated by the shortest distance are combined. The ‘shortest distance’ between two clusters is based on all pair-wise distances between the elements of both clusters, so that the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved.

Complete-linkage clustering can avoid chaining (clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very far apart from each other) (cf. Everitt et al. 2001).

The results of hierarchical clusters are represented graphically in a dendrogram, which is a branching diagram that represents the relationships of similarity among a group of entities. Its leaves present the variables (subcorpora), and the branches or clades – the clusters. The arrangement of the clades tells us which leaves are most similar to each other. The height of the branch points indicates how similar or different they are from each other: the greater the height, the greater the difference. Highly correlated clusters are nearer the bottom (also leftmost or outermost, depending on the representation) of the dendrogram.

4 Analyses, results and interpretation

This section presents the results of the analysis which we performed in several steps. First, we analyse intra-dimensional variation, i.e. variation along the dimension of translation method and along the dimension of register (4.1). Second, we combine both dimensions and analyse variation in-
fluenced by both dimensions (4.2). In section 4.3, we analyse the features contributing to the resulting variation.

4.1 Intra-dimensional variation

a) Variation across translation methods

First, the variation across translation methods is analysed, for which we have five dependent variables: PHT, SHT, RBMT, GSMT and MSMT. Distance measures are calculated for each variable on the basis of the total number of occurrences of the features. The subcorpora, each representing the lowest (most left) leaves of the dendrogram (see Figure 1) are joined up to the top (on the right side of the graph). The dendrogram is then read from the bottom up (from left to right in our case), identifying in this way, which subcorpora, and hence which translation varieties, are most similar to each other.

Figure 1. Variation across translation methods.

The first node to join together is that of PHT and RBMT. The connection between them is the closest link to the bottom of the diagram, which means that the distance between them is the smallest. Together, they join with GSMT. This indicates that every node within this cluster is more similar to every other node within this cluster than to any nodes that join at a higher
level (SHT and MSMT in this case). We move further and join this node with SHT, then, moving further to MSMT. The length of the horizontal lines indicates the degree of difference between branches.

The greatest differences (the longest lines between the last two clusters) are observed for SHT and MSMT, which vary strongly from other subcorpora. We assume that the variance of SHT and MSMT from the rest can be explained by another dimension of variation that comes into play—experience (which includes both degree of experience of human translators and the amount of training data in a statistical machine translation system) involved in the translation process. Differences in this dimension were pointed out in several studies, e.g. by Göpferich and Jääskeläinen (2009) and Carl and Buch-Kromann (2010) for human translation or Estrella et al. (2007) and Koehn and Haddow (2012) for machine translation. Although none of them combine human and machine translation in their analyses, some parallels can be observed in both translation varieties. For instance, Göpferich and Jääskeläinen (2009) state that with increasing translation competence, translators focus on larger translation units, just as in (phrase-based) machine translation, large training data sets make it possible to learn longer phrases (see Koehn 2010).

In our data, both SHT and MSMT reveal a certain lack of experience: the former is produced by novice translators, whereas the latter is trained with a restricted set of data. We assume that human translators and statistical MT systems behave in a related fashion in terms of their performance. That is the degree of experience in human translation and the amount of data in machine translation contribute to similar outcomes in both translation varieties. Translations by novice translators are similar to those produced with a statistical system trained with a small amount of data, whereas texts translated by professional translators may be comparable to those produced with SMT systems trained with larger data. However, this trend will need to be further tested in future experiments.

b) Variation across registers

We also analyse the variation across registers of all translated texts (not differentiating between methods). In this case, we have seven dependent variables: ESSAY, FICTION, INSTR, POPSCI, SHARE, SPEECH and TOU. The dendrogram representing distances and clusters of these variables is shown in Figure 2.
Moving from left to right in Figure 2, we observe the following groups of registers: 1) ESSAY and TOU; 2) FICTION; 3) INSTR and POPSCI and 4) SHARE and SPEECH. The next two nodes join 1) with 2), and 3) with 4), although FICTION varies more strongly from ESSAY and TOU than INSTR and POPSCI from SHARE and SPEECH. The greatest difference lies between two bigger groups of registers, which we can observe at the very right.

![Diagram of Figure 2. Variation across registers.](image)

Similarities between ESSAY and TOU in German non-translated texts have been observed by Neumann (2013) for almost all sub-dimensions of field, tenor and mode (see Neumann 2013: 313, Table 60). Similarly, Dwersy et al. (2014) found a clear tendency of translations to normalise their
Exploratory analysis of dimensions influencing variation in translation

features in order to adapt to target language conventions. The authors used the same dataset as that employed by Neumann (2013). Our results show that the same tendency can be observed not only for professional human translation (as in the case of the two studies mentioned here), but also for machine translation. Neumann (2013) also demonstrates the individuality of FICTION in both English and German. She points out that translations in the FICTION register do not exhibit any deviation from the originals (both English and German), which can probably be explained by the fact that the register features of fictional texts coincide in both languages. As a result, this register is distinctive from the other registers under analysis. The features of SHARE and SPEECH in our translations coincide, and we observe the phenomena of shining through and normalisation here, which means that the features of both source and target languages interact in translations of these registers. In Neumann's description of register profiles, SHARE and SPEECH would coincide if we take into account both languages, see profiles of the English and German original registers in Neumann (2013: 309 and 2013: 313) for details. At the same time, the similarities between INSTR and POPSCI in our data rather contradict Neumann's definition of register profiles. In her description, these two registers coincide only in mode of discourse, if both languages are considered. If compared monolingually, that is, within each language, INSTR and POPSCI have similarities in English only, and only in the parameter of tenor. However, the author also states that register profiles of translations might deviate from those of originals. We suspect that in the case of INSTR and POPSCI we observe this kind of deviation, which explains the unexpected clustering of these registers in our data.

Comparing the distance indication on the scale in the two figures, we find that the difference between register clusters is greater than that between translation methods. On this basis, it can be concluded that variation along the register dimension is greater in our translation data than variation along the translation method dimension. To test this, we need to combine both dimensions in one single analysis.

4.2 Inter-dimensional variation in translation varieties

To analyse the interplay between translation methods and registers, and to find out which of these two dimensions causes greater variation in our data, we calculated the distances between registers of all translations that underlie the hierarchical clustering. In this case, we have 35 dependent
variables, for which distances are calculated. The results are represented in Figure 3.

The dendrogram clearly reveals two very distinct groups: the bottom group seems to consist of two more distinct clusters, while the clustering of most classes in the upper group is more levelled out.

If we start to generate the tree from the outermost nodes (or leaves), we observe a clear predominance of register features for the clustering of fictional texts, tourism leaflets and political essays. The only exception here is the set of tourism texts translated by students, which varies more from other translations of TOU, though ultimately it joins the cluster formed by TOU and ESSAY.
It is worth noting that SHT-FICTION is also outstanding in the corresponding class. These results are in line with our observations regarding groups of registers in section 4.1 above, where we show that FICTION builds a class of its own. The tendency to group according to experience (which we observed within the intra-dimensional variation across translation methods in 4.1) is detected on the smallest nodes of fictional and tourism texts: PHT shows more similarities with RBMT and GSMT. However, in the register of political essays, both varieties of human translations are clustered together before they are clustered with the varieties of the MT systems. On the whole, the results for this group confirm the observations.
that variation along the register dimension is more prominent in translations than along the translation method dimension.

The clusters comprising TOU, ESSAY and FICTION on the one hand, and INSTR, POPSCI, SHARE and SPEECH on the other, reveal different variation behaviour. The latter does not demonstrate any prominence of either register or method, as the two bigger classes within this group are built up by a mixture of registers: one of instruction manuals and popular-scientific articles, and the other made up by SHARE and SPEECH texts. The dimension of experience is not present in this group of translations. However, we observe a tendency for translations to form human vs. machine clusters. In some cases, we even observe a more fine-grained method-driven clustering: Moses-generated translations of INSTR and POSPCI, as well as SHARE and SPEECH are grouped, which means that these translations do not diversify in terms of register, and at the same time, the resulting translation varieties deviate more strongly from other translations.

In general, groups in the INSTR-POPSCI-SHARE-SPEECH cluster are more heterogeneous and diverse, and contrary to the TOU-ESSAY-FICTION cluster, it is rather difficult to identify which of the two dimensions predominate in this dataset. Both registers and methods are less pronounced and distinct in the second cluster, and the new dimension of variation which we discovered in section 4.1 is not observed here at all.

As we are not able to define which of the two dimensions under analysis is more prominent in the case of the second group, we tested how many clusters we need to describe this dataset, excluding ESSAY, FICTION and TOU from the analysis. The underlying assumption here is that if a resulting cluster contains more variables representing translation method, it shows the prominence of the register dimension, while the prominence of the method dimension is demonstrated by a greater number of register variables.

The maximum number of clusters that we can have should not exceed five (as we have four registers and five translation methods), and accordingly, we cut the tree into five clusters (see Table 4).

Table 4. Cluster membership in a five-cluster tree.

<table>
<thead>
<tr>
<th></th>
<th>SHT</th>
<th>PHT</th>
<th>RBMT</th>
<th>GSMT</th>
<th>MSMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>POPSCI</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>SHARE</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>SPEECH</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Cluster 1 contains two registers and four translation methods, and thereby reflects the register dimension. Cluster 2 is represented by three registers and four translation methods, being rather on the borderline of the two dimensions. Clusters 3 and 4 are both built up by two registers and two translation methods. Interestingly, they contain the same register variables (SHARE and SPEECH), opposing in this way human translation vs. statistical machine translation (difference in the dimension of translation method). The last cluster (5) includes MSMT only, and also reflects translation method.

Summarising these observations, we can say that the greatest differences are still observed for the register dimension, and not that of translation method, as variation along the latter is observed rather for smaller groups of variables. However, we admit that register groupings are different across human vs. machine translation. In general, machine translations show a lower number of distinctive registers than human translations. This finding can be explained by the fact that register features have not been taken into account in machine translation to date. In studies on machine translation, authors mostly operate with the notion of domain, which covers the lexical level only, cf. Lapshinova and Pal (2014). Therefore, on the level of register settings, the translations produced by MT systems do not vary as much as those produced by humans who are aware of register constraints. This observation does not tie in with the findings in section 4.1 above, in which the clustering of registers in all translations together were taken into account. This means that if a translation is analysed in terms of register settings, we need to take into account the method with which this translation was produced. For instance, in order to judge the quality of a translation regarding its correspondence to the register standards in the target language, the production method involved needs to be considered.

4.3 Features involved in the cluster formation

To analyse the influencing factors, i.e. the features contributing to the classification of the translation varieties in our data, we need to consider the numeric data. Table 6 presents the original feature set for each of the seven clusters. We chose this number of clusters according to the number of registers in our data as they seem to be more prominent than methods, although some of them are mixed, see Table 5.

Comparing the figures across clusters, we can characterise each cluster according to the set of features specific to it. For instance, cluster 1 (politic-
Ekaterina Lapshinova-Koltunski

...al essays) is characterised by an average distribution of nominal and verbal parts-of-speech and phrases, a relatively low number of conjunctive relations. The amount of pronominal reference here is also lower than average. General nouns, *ung*-nominalisations, as well as modal verbs expressing obligation are more frequent here than on average. Modality, especially that with the meaning of obligation, is one of the indicators of argumentative goal orientation (a part of the context parameter of field in register theory, see section 2.1). Argumentative texts contain significantly more modality than texts pursuing other goals.

Table 5. Main clusters in the analysed data.

<table>
<thead>
<tr>
<th>clusters</th>
<th>translation varieties</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>political essays</td>
</tr>
<tr>
<td>2</td>
<td>fictional texts</td>
</tr>
<tr>
<td>3</td>
<td>instruction manuals and popular-scientific texts except SHT-POPSCI</td>
</tr>
<tr>
<td>4</td>
<td>letters-to-shareholders and political speeches except SHT- and PHT-SHARE</td>
</tr>
<tr>
<td>5</td>
<td>human translations of letters-to-shareholders and SHT- SPEECH</td>
</tr>
<tr>
<td>6</td>
<td>student translations of tourism texts</td>
</tr>
<tr>
<td>7</td>
<td>tourism leaflets except SHT-TOU</td>
</tr>
</tbody>
</table>

Cluster 1 has a high number of modal verbs of obligation, if compared to other clusters (cluster 5 only has a higher number of them). The verbs are used rather in their meaning of personal obligation, than logical necessity, see example (1).


‘First, we must balance the rising production by providing a clean and efficient energy consumption in the foreground. Second, we must expand our international relationships with the consumer and producer countries. Third, we must expand our energy sources and
At the same time, the translations of cluster 1 also demonstrate features of goal type exposition, which are reflected in lexico-grammar by the low number of personal pronouns and a higher number of nouns, including general nouns and ungrammaticalisations. These characteristics certainly correspond with the observation made by Neumann (2013: 184).

The main distinctive features of cluster 2 include first a prevalence of pronominal reference, which is expressed with personal pronouns in most cases, see example (2). This usually characterises spoken language, and as we know, FICTION is on the borderline between written and spoken language as it contains conversations.

Table 6. Features contributing to cluster definition.

<table>
<thead>
<tr>
<th>cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>content.words</td>
<td>7037</td>
<td>4407</td>
<td>9469</td>
<td>10351.5</td>
<td>11098</td>
<td>6633</td>
<td>8568.5</td>
</tr>
<tr>
<td>np.chunk</td>
<td>4651</td>
<td>3384</td>
<td>6274</td>
<td>7039</td>
<td>7332</td>
<td>4115</td>
<td>5506</td>
</tr>
<tr>
<td>vp.chunk</td>
<td>1712</td>
<td>1570</td>
<td>2344</td>
<td>2593.5</td>
<td>2908</td>
<td>1381</td>
<td>1495</td>
</tr>
<tr>
<td>nominal</td>
<td>8885</td>
<td>5679</td>
<td>11483</td>
<td>13239</td>
<td>13772</td>
<td>7993</td>
<td>10535.5</td>
</tr>
<tr>
<td>verbal</td>
<td>4197</td>
<td>3485</td>
<td>5857</td>
<td>6396.5</td>
<td>7124</td>
<td>3361</td>
<td>4177</td>
</tr>
<tr>
<td>additive</td>
<td>726</td>
<td>622</td>
<td>717</td>
<td>1041.5</td>
<td>1169</td>
<td>752</td>
<td>880</td>
</tr>
<tr>
<td>adversative</td>
<td>313</td>
<td>250</td>
<td>502</td>
<td>413.5</td>
<td>453</td>
<td>232</td>
<td>274</td>
</tr>
<tr>
<td>causal</td>
<td>115</td>
<td>169</td>
<td>271</td>
<td>133</td>
<td>179</td>
<td>65</td>
<td>79</td>
</tr>
<tr>
<td>temporal</td>
<td>384</td>
<td>332</td>
<td>614</td>
<td>565.5</td>
<td>628</td>
<td>213</td>
<td>307.5</td>
</tr>
<tr>
<td>modal</td>
<td>96</td>
<td>146</td>
<td>248</td>
<td>130.5</td>
<td>159</td>
<td>57</td>
<td>89</td>
</tr>
<tr>
<td>pronnp</td>
<td>39</td>
<td>154</td>
<td>82</td>
<td>66.5</td>
<td>62</td>
<td>35</td>
<td>21</td>
</tr>
<tr>
<td>all.np</td>
<td>3399</td>
<td>2717</td>
<td>4602</td>
<td>5121</td>
<td>5140</td>
<td>2799</td>
<td>3969</td>
</tr>
<tr>
<td>gen.nouns</td>
<td>121</td>
<td>52</td>
<td>107</td>
<td>138.5</td>
<td>138</td>
<td>37</td>
<td>48.5</td>
</tr>
<tr>
<td>all nouns</td>
<td>3464</td>
<td>1852</td>
<td>4816</td>
<td>5291.5</td>
<td>5701</td>
<td>3944</td>
<td>4417</td>
</tr>
<tr>
<td>obligation</td>
<td>72</td>
<td>32</td>
<td>46</td>
<td>71.5</td>
<td>76</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>permission</td>
<td>76</td>
<td>38</td>
<td>174</td>
<td>96</td>
<td>162</td>
<td>81</td>
<td>64.5</td>
</tr>
<tr>
<td>volition</td>
<td>18</td>
<td>35</td>
<td>32</td>
<td>21.5</td>
<td>30</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>evaluation</td>
<td>11</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>ungrammatical</td>
<td>685</td>
<td>94</td>
<td>422</td>
<td>786</td>
<td>756</td>
<td>134</td>
<td>210.5</td>
</tr>
<tr>
<td>perspron</td>
<td>521</td>
<td>1127</td>
<td>781</td>
<td>1140.5</td>
<td>1084</td>
<td>288</td>
<td>407.5</td>
</tr>
<tr>
<td>dempron</td>
<td>122</td>
<td>102</td>
<td>175</td>
<td>262.5</td>
<td>252</td>
<td>43</td>
<td>68</td>
</tr>
</tbody>
</table>

(2) *Er glaubte es nicht einmal ansatzweise, aber er wollte, dass ich es glaubte oder mir Gedanken darüber machen würde. Und er könnte sich dann über mich lustig machen (SHT-FICTION).*

‘He did not even start to believe *it*, but he wanted *me* to believe *it* or think about it. And then he could laugh about *me* [...]’.

225
Moreover, pronouns also correspond to one of the characteristics of narrative texts. Further characteristics include description of events in clausal structures rather than in nominal structures, which leads to a lower amount of nominalisations in this cluster. Again, the features characterising cluster 2 correspond to those defined for FICTION, see Steiner (2012) and Neu mann (2013).

Cluster 7 contains all translations of tourism texts except the one produced by inexperienced translators. We assume that the main difference is in the proportion of nominal and pronominal reference in translation variants of this register. SHT-TOU seems to contain more pro-forms than the other translation varieties, see examples in (3).

(3a) Nordirlands Bevölkerung liebte schon immer die Natur. Die Menschen aus Ulster sind keine Stubenhocker! Manche trödeln an den 'Loughs' (Seen oder Meeresarmen) herum, andere verbringen ihre Freizeit mit Angeln oder Bootfahren, wieder andere machen Ausflüge mit der Familie in die Berge oder in die Waldparks speziell am Wochenende (PHT-TOU).

‘Northern Ireland’s population has always loved nature. The people of Ulster are no couch potato! Some dawdle on the 'Loughs' (lakes or estuaries) around, others spend their leisure time with fishing or boating, others make trips with the family in the mountains or in the forest park especially on weekends.’


‘The cultural heritage of Northern Ireland consists largely of its landscape. The people of Ulster like to stay outdoors. In their spare time they stroll along the coast, and on the weekends they take a family trip to the mountains.’

(3c) Das Erbe von Nordirland ist in großem Maße ländlich. Ulster-Leute sind Leute im Freien. Sie verbringen ihre Freizeit, die um die Küste oder das Gehen auf Familienexpeditionen zu den Bergen an den Wochenenden pottering ist (RBMT-TOU).

‘The heritage of Northern Ireland is largely rural. Ulster people are people in outdoors. They spend their leisure time, which is pottering
around the coast or going on family expeditions to the mountains on weekends.’

‘The heritage of Northern Ireland is characterized largely as rural. Ulster people are outdoor people. They spend their free time pottering around the coast or go on family expeditions to the mountains on weekends.’

‘The heritage of Northern Ireland is largely rural. Ulster people are people of outdoors. They their free time pottering around the coast or family built up to the mountains on weekends.’

SHT-TOU forms its own cluster (nr. 6), which means that it is distinctive from the other translation varieties. However, we are not able to claim that the relation between nominal and pronominal reference is the only indicator of this variation, and we need to test the interplay of further features in this subcorpus to be able to test the differences of SHT-TOU from other translations of tourism texts.

The heterogeneous cluster 3 (containing instruction manuals and popular-scientific texts) is characterised by conjunctive relations and modal verbs, which indicates the argumentative goal. Interestingly, previous studies did not reveal commonalities between these two registers. In Neumann (2013), instruction manuals appear to be very distinct from the other registers under analysis.

SHARE and SPEECH, which compose the fourth cluster, have a number of commonalities. According to Neumann (2013), these two registers seem to be closer in English than in German, which might indicate the influence of the source texts on our translations.

Cluster 5 is the second smallest cluster and contains the human-translated SHARE texts and political speeches translated by novice translators. The figures in Table 6 show that this cluster is characterised by a large amount of both nominal and verbal classes, general nouns and nominalisations, as well as additive and temporal conjunctive relations. These are the same features which also characterise cluster 4, and thus the other transla-
tions of the same two registers. However, translations in cluster 5 reveal a greater amount of conjunctive relations and modality meanings than those in 4.

Table 7. Differences between translation methods of SHARE.

<table>
<thead>
<tr>
<th></th>
<th>SHT</th>
<th>RBMT</th>
<th>GSMT</th>
<th>MSMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHT</td>
<td>0.007501</td>
<td>4.099e-13</td>
<td>3.053e-10</td>
<td>2.536e-11</td>
</tr>
<tr>
<td>SHT</td>
<td>&lt; 2.2e-16</td>
<td>&lt; 2.2e-16</td>
<td>&lt; 2.2e-16</td>
<td></td>
</tr>
<tr>
<td>RBMT</td>
<td></td>
<td>0.2752</td>
<td>0.5023</td>
<td></td>
</tr>
<tr>
<td>GSMT</td>
<td></td>
<td></td>
<td>0.6733</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Differences between translation methods of SPEECH.

<table>
<thead>
<tr>
<th></th>
<th>SHT</th>
<th>RBMT</th>
<th>GSMT</th>
<th>MSMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHT</td>
<td>0.001616</td>
<td>0.3293</td>
<td>0.3293</td>
<td>0.3293</td>
</tr>
<tr>
<td>SHT</td>
<td>3.974e-05</td>
<td>3.974e-05</td>
<td>3.974e-05</td>
<td></td>
</tr>
<tr>
<td>RBMT</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GSMT</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

We calculate p-values (calculated with the Chi-squared test) to see the differences between translation varieties in terms of modal verb distribution. Considering the p-values for modal verbs expressing permission in SHARE and SPEECH (Tables 7 and 8), we clearly see that SHT and PHT differ from machine translation in SHARE (also differing from each other), whereas SHT differs from all other translation varieties in SPEECH (PHT differs from SHT only). It is interesting to note that whereas in human translations of SHARE modal verbs express both possibility and obligation (expressed with können and dürfen) with possibility/ability prevailing, machine-translated SHARE contains possibility verbs only.

Overall, the features contributing to our cluster formation correspond to the features specific to the registers involved, as described by Neumann (2013) or Steiner (2012), with some exceptions. From this, it follows that the dimension of register has more influence on variation in translation than that of translation method, at least in the dataset analysed here. The features contributing to the generation of the resulting clusters can shed light on the linguistic properties of different translations. At the same time, the exceptions observed, i.e. emerging mixed classes, point to further issues that we
need to analyse more closely. This requires an investigation of the features contributing to such mixed class formation. We believe that these features can point to translation problems caused by textual differences between the source and the target language. To test this, we need to include original data (of both source and target language) into our analysis.

Moreover, the observed differences between human and machine translations in register classification show that the linguistic features of translations are influenced by the methods involved in the translation process. These differences result partly from the fact that register settings are not paid attention in the development of machine translation systems.

5 Conclusion and future work

The present study analysed the interplay between two dimensions influencing the linguistic settings of translations: register and translation method. We applied unsupervised analysis techniques with the aim of discovering new structures in our translation data. Our assumption was that translations in our dataset would cluster according to either registers the texts belong to or translation methods involved in the translation process, indicating in this way the prominence of one of the two dimensions.

Our results show that both dimensions are present in the groupings of our subcorpora. For some of them, e.g. all translations of fictional and tourism texts, as well as political essays, we clearly see the prevalence of the register dimension. Interestingly, variation along translation method can be detected within register-specific clusters. Moreover, we observe another dimension of variation – that of translation experience. By contrast, translations of the other texts (instruction manuals, popular-scientific texts, political speeches and letters to shareholders) are classified into more fine-grained clusters, for which individual tendencies are observed, influenced by either one of the two dimensions. We observe here smaller clusters of human vs. machine translations, as well as mixed register clusters. The predominance of one or the other dimension is observed on lower nodes. The existence of mixed classes in our data indicates the need to have a closer look at them and their linguistic properties. The emergence of these classes might indicate the existence of phenomena causing translation problems, e.g. ambiguities. Differences in their resolution result in the independent variation of translations that we observe in our data. This kind of analysis,
however, needs to take non-translated texts alongside translated texts into consideration.

We also think that we need to perform a more detailed analysis of the dimension of translation experience, as the amount of experience of a translator and the amount of data involved in the training of an SMT system seem to have a similar influence on the outcome of the translation process.

Overall, the results of our analysis deepen our knowledge of the linguistic properties of translated texts, which, in turn, furthers our understanding of variation processes in translation. Knowledge of the variation caused by translation methods also provides us with information on method-specific features, which can facilitate their improvement, especially in the area of machine translation. In this way, the results of our analysis can be used for enhancement of MT systems, as well as evaluation and quality estimation of both human and machine translation.

In our future work, we plan to include supervised techniques of analysis, which will allow us to detect distinctive features of each translation variety. We believe that a combination of unsupervised and supervised techniques, as elaborated in Diwersy et al. (2014), is needed to detect linguistic properties on a more fine-grained level, as hierarchical clustering allows us to observe tendencies, as well as differences and similarities between certain subcorpora. We also need to include an intra-lingual analysis of the same features in English and German non-translated texts, as we believe that they also have a great impact on translation varieties. We will then be able trace the variation along the additional dimension, that of language, which was not taken into account in the present analysis. In this case, we expect to observe different degrees of shining through and normalisation in the translation varieties under analysis.

References


Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Gill Francis, Mona Baker & Elena Tognini-Bon-
Exploratory analysis of dimensions influencing variation in translation


Exploratory analysis of dimensions influencing variation in translation 31


32  Ekaterina Lapshinova-Koltunski


Notes
i http://www.my-across.net/
Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification

Anonymous NAACL submission

Abstract
This paper introduces information density and machine translation quality estimation inspired features to automatically detect and classify human translated texts. We investigate two settings: discriminating between translations and comparable originally authored texts, and distinguishing two levels of translation professionalism. Our framework is based on delexicalised sentence-level dense feature vector representations combined with a supervised machine learning approach. The results show state-of-the-art performance for translationese detection with information density and quality estimation based features, while results on translation expertise classification are mixed.

1 Introduction
Translations, regardless of the method they were produced with, are different from their source texts and from originally authored comparable texts in the target language. This has been confirmed by many linguistic studies on translation properties commonly called translationese (Gellerstam, 1986). These studies show that translations tend to share a set of lexical, syntactic and/or textual features distinguishing them from non-translated texts. As most of these features can be measured quantitatively, we are able to automatically distinguish translations from originals (Baroni and Bernardini, 2006; Ozdowska and Way, 2009; Kurokawa et al., 2009). This is useful for Statistical Machine Translation (SMT), as language and translation models can be improved if the translation direction and status of the data (translation or original) is known (Lembersky, 2013).

Research on translationese has recently focused on exploring features capturing aspects of translationese such as simplification, explicitation, convergence, normalisation and shining-through (Volansky, 2012; Ilisei, 2012). Here we extend this work as follows: (i) we investigate the impact of information density and surprisal features, (ii) we explore the use of features used in machine translation quality estimation (Blatz et al., 2003; Specia et al., 2010), (iii) we explore classification between originally authored text and trainee and professional translation, as well as between professional and trainee translation. In order to avoid biasing classification by topic content, throughout our experiments we use fully delexicalised features, resulting in dense vector representations (rather than sparse vectors, where the size of the vectors can be up to and in fact exceed the size of the vocabulary). We show that information theory as well as translation quality estimation inspired features achieve state-of-the-art results in original vs. translation classification.

Languages provide speakers with a large number of possibilities of how they may encode messages. These include the choice of phonemes, words, syntactic structures, as well as arranging sentences in discourse. Speakers’ decisions regarding these choices are influenced by diverse factors: cognitive processing limitations can impact variation in linguistic encoding across all linguistic levels. Text production conditions, including monolingual vs.
multilingual settings, can influence this variation: in translation, choices can be shaped by aspects of both the source and the target language.

Contrastive studies have shown that information density is distributed differently in English and German (Doherty, 2006; Fabricius-Hansen, 1996). These contrasts may impact translation, and in case of source language shining through\(^1\), we would expect to observe differences between translations and comparable originals in terms of information density. Additionally, translations are often more specialised and more conventionalised than originals (excluding translation of fictional texts). In this paper we investigate whether and to what extent information density based features are useful in translation classification.

Quality estimation (QE) (Blatz et al., 2004; Ueffing and Ney, 2005) is the attempt to learn models that predict translation quality without access to a reference translation at prediction time. Translation is always a process of transforming a source into a target text. This process is prone to error. In this paper we explore whether and to what extent the extensive research on QE can be brought to bear on the problem of translation vs. originals classification, and in particular the discrimination between novice and professional translation output.

Below we explore the ability of our features to distinguish between 1) non-translated texts and translations by professionals, 2) non-translated texts and translations by translator trainees, and 3) the two translation varieties that diverge in the degree of translation experience. We report results in terms of accuracy and f-score, and provide a feature analysis in order to understand the role of the information density and QE inspired features in the task.

The paper is organised as follows: related work is presented in Section 2. The experimental setup is detailed in Section 3, followed by the results and analysis in Section 4. Finally, conclusion and future work are provided in Section 5.

2 Related Work

We briefly review previous work on translationese, information density, machine translation quality estimation and studies on human translation expertise.

2.1 Translationese

A number of corpus-based studies on translation have shown that it is possible to automatically predict whether a text is an original or a translation (Baroni and Bernardini, 2006; Koppel and Ordan, 2011). These approaches are based on the concept of translationese – a term coined by Gellerstam (1986). The idea is that translations exhibit properties which distinguish them from original texts, both the source texts of the translation and comparable texts originally authored in the target language. Baker (1993; 1995) claimed these properties to be universal, i.e. (source) language-independent, emphasising general effects of the process of translation.

However, translationese includes features involving both source and target language. Most linguistic studies distinguish explicitation – a tendency to spell things out rather than leave them implicit and implicitation (the opposite effect), simplification – a tendency to simplify the language used in translation, normalisation – a tendency to exaggerate features of the target language and to conform to its typical patterns, levelling out or convergence – a relatively higher level of homogeneity of translated texts compared to non-translated ones, and interference or shining through (e.g. Teich (2003)). While simple lexicalised features including word tokens and character n-grams can produce near perfect classification results (Avner et al., 2014), a significant amount of work has gone into devising features that can capture presumed linguistic aspects of translationese (Volansky, 2012).

Rabinovich et al. (2015) explore unsupervised discrimination of translations based on principal components analysis for dimensionality reduction followed by a clustering step. The method is robust to unbalanced and heterogeneous datasets, which may be useful to handle mixed domain, genre and source of data, a common situation when training language and translation models.

Automatic classification of original vs. translated texts has applications in machine translation,
especially in studies showing the impact of the nature (original vs. translation) of the text in translation and language models used in SMT. Kurokawa et al. (2009) show that taking directionality into account when training an English-to-French phrase-based SMT system leads to improved translation performance. Ozdowska & Way (2009) analyse the same language pair and demonstrate that the nature of the original source language has an impact on the quality of SMT output. Lembersky et al. (2012) show that BLEU scores can be improved by language models compiled from translated texts and not from comparable originally authored ones.

2.2 Information Density

In a natural communication situation, speakers tend to exploit variations in their linguistic encoding – modulating the order, density and specificity of their expressions to avoid informational peaks and troughs that may result in inefficient communication. This is often referred to as the uniform information density hypothesis (Frank and Jaeger, 2008). The information conveyed by an expression can be quantified by its surprisal, a measure of how predictable an expression is given its context. Simplification and explicitation may impact the average information density measured on translated texts compared to comparable originally authored ones in the same language. Source language interference should result in peaks of measured surprisal values in translated texts, while the information density may remain uniform in originals.

According to Hale (2001), a surprisal model allows the estimation of the probability of a parse tree given a sentence prefix. Levy (2008) showed that a lexical-based surprisal measure can be obtained by computing the negative log probability of a word given its preceding context: $S = -\log P(w_{k+1}|w_1 \ldots w_k)$. Following Demberg et al. (2013), we estimate surprisal in three ways, at the word, part-of-speech and syntax levels, based on $n$-gram language models and language models trained on unlexicalized part-of-speech sequences and flattened syntactic trees. Note that all resulting feature vectors do not represent lexical information but information theoretic surprisal measures.

2.3 Quality Estimation

Machine translation QE is the process of estimating how accurate a translation is through characteristic features of the source and target texts, and (possibly) also the translation engine, with a supervised machine learning setting to estimate quality scores. QE can be applied at the word, sentence and document level (Gandrabur and Foster, 2003; Ueffing et al., 2003; Blatz et al., 2003; Scarton and Specia, 2014).

Many different delexicalised dense features have been explored in previous work on QE, including language and topic models, $n$-best lists, etc. (Quirk, 2004; Ueffing and Ney, 2004; Specia and Gimenez, 2010; Rubino et al., 2013a). It has been shown that the performance of a supervised classifier to distinguish between originals and translations is correlated with the quality of the translated texts (Aharoni et al., 2014): low quality translation, containing grammatical and syntactic errors, as well as incorrect lexical choices, are robust indicators of automatic translations. To the best of our knowledge, there are no empirical studies on the level of professional expertise in the translation process and its correlation with the performance of a translationese classifier.

2.4 Translator Experience

Jääskeläinen (1997) describes translational behaviour of professionals and non-professionals who perform translation from English into Finnish. Carl and Buch-Kromann (2010) apply psycholinguistic methods in their analysis. They present a study of translation phases and processes for student and professional translators, relating translators’ eye movements and keystrokes to the quality of the translations produced. They show that the translation behaviour of novice and professional translators differs with respect to how they use the translation phases. Englund Dimitrova (2005) develops a combined process and product analysis and compares translators with different levels of translation experience, but concentrates only on cohesive explicitness.

Most of these works are rather process-oriented than product-oriented, which means that features of translated texts are rarely taken into account.
However, some of the findings are valuable for the analysis of translated texts. For instance, Göpferich & Jääskeläinen (2009) find that with increasing translation competence, translators focus on larger translation units, which can impact the choice of linguistic encoding translators use.

### 3 Experimental Setup

Our experiments are designed to investigate underexplored topics focusing on (i) information theoretic and (ii) machine translation QE features in translation classification. We use dense vector representations with fully delexicalised features and investigate three hypotheses:

1. originals & professional translations should be close in terms of quality and thus more difficult to separate automatically,

2. originals & student translations should be distant in terms of quality and thus easier to classify,

3. professional & student translations should both contain translationese features and thus may be very difficult to differentiate.

#### 3.1 Supervised Classification

In order to train a classifier and predict binary labels on unseen data, we use a dense vector sentence-level representation associated with a class \((x_i, y_i), i = 1, \ldots, l\) (l is the number of training instances) with \(x_i \in \mathbb{R}^n\) (n is the size of a dense vector) and \(y_i \in \{-1, 1\}\). We train classification models with a support vector machine SVM (the C-SVC implementation in LIBSVM (Chang and Lin, 2011)) as a quadratic optimization problem:

\[
\begin{align*}
\min_{\omega, b, \xi} & \quad \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{l} \xi_i, \\
\text{subject to} & \quad y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.
\end{align*}
\]

\(\phi\) is a kernel function and allows the projection of training data to a higher dimensional space. We use the radial basis function (RBF) kernel, as it produced the best empirical results compared to linear and polynomial kernels. We predict the class for unseen instances \(x\) as follows:

\[f(x) = \text{sgn}(\omega^T \phi(x) + b).\]

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Token (M)</th>
<th>Sentence (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl Originals</td>
<td>4.1</td>
<td>155.5</td>
</tr>
<tr>
<td>Literature Originals</td>
<td>1.3</td>
<td>48.1</td>
</tr>
<tr>
<td>Literature Translations</td>
<td>1.4</td>
<td>45.8</td>
</tr>
<tr>
<td>Politics Originals</td>
<td>0.2</td>
<td>9.7</td>
</tr>
<tr>
<td>Politics Translations</td>
<td>0.2</td>
<td>8.7</td>
</tr>
</tbody>
</table>

**Table 1:** Details of the corpora used to train language and \(n\)-gram frequency models for originally authored texts and translations.

Two hyper-parameters have to be set for C-SVC with the RBF kernel: the regularisation parameter (or penalty) \(C\) and the kernel parameter \(\gamma\). We use grid-search to find optimal values, performing a 5-fold cross-validation on the training data. To avoid over-fitting, we use a held-out development set to evaluate the models obtained.

#### 3.2 Datasets

The datasets used in our experiments are separated into two subsets: corpora used to extract features and corpora used to train, tune and test our classifiers. The former are taken from the publicly available bilingual English-German parallel corpora consisting of parliamentary proceedings, literary works and political commentary, compiled by (Rabinovich et al., 2015). These corpora are used individually to train language models and compute \(n\)-gram frequency distributions. Basic corpus statistics are presented in Table 1. The latter ones are composed of German texts, taken from the VARTRA corpora (Lapshinova-Koltunski, 2013), which were either originally written in German (originals) or translated from English (translations). Originals and translations belong to the same genres and registers and can be considered comparable. They include a mixture of literary, tourism and popular-scientific texts, instruction manuals, commercial letters and political essays and speeches. The VARTRA translations are split in two sets: one produced by professional translators, and one produced by translator trainees. Details are presented in Table 2. We extract balanced subsets of training, tuning and testing data containing three, one and two thousands sentences, respectively, of each type.
3.3 Feature Sets

For classification, input text is represented as a set of feature vectors. The features capture aspects of information density and translation QE. Throughout we use unlexicalized lower-dimensional dense vectors rather than high-dimensional lexicalized sparse vectors to minimize the input of specific content on classification results. We extract a total of 778 features\(^2\) and separate them into four subsets corresponding to broad but distinct characteristics of original and translated sentences: surface and distortion features are related to QE, surprisal and complexity features are inspired by information theory.

**Surface Features** - 13 surface features based on meta representations of sentences’ lexical form. Features include sentence and average word length, the number of word tokens and number of punctuation marks. Three case-based features capture the number of upper-cased letters and words, and a binary feature indicates whether a sentence starts with an upper-case character. Another binary value encodes whether the sentence ends with a period. Two features are obtained from the ratio between the number of upper-cased and lower-cased letters, the number of punctuation marks and the length of the sentence. Finally two features capture the number of periods merged with words and words with mixed-case characters.

**Surprisal Features** - 225 features based on the surprisal measure presented in Section 2.2 are extracted using language models trained on words, delexicalised part-of-speech and flattened syntactic trees. The language models are trained on individual\(^3\) corpora presented in Table 1. We extract \(n\)-gram \((n \in \{1; 5\})\) log-probabilities and perplexities, with and without the tags indicating the beginning and ending of sentences, using the SRILM toolkit (Stolcke et al., 2011).

**Complexity Features** - 315 features based on \(n\)-gram frequencies, indicating how frequent the lexical choices, part-of-speech and flattened syntactic sequences present in the text to be classified are. As for the surprisal features, we use the same originally authored and translated texts individually to extract \(n\)-grams frequency quartiles. We extract the percentage of \(n\)-grams \((n \in \{1; 5\})\) occurring in each quartile. Frequency percentages are averaged at the sentence level, leading to 4 features per sentence (one per quartile) given a value of \(n\), for each corpus used to define the frequency quartiles. This approach allows us to avoid encoding raw \(n\)-gram features and keep a dense vector representation (Blatz et al., 2003).

**Distortion Features** - 225 features based on the possible distortion in lexical, part-of-speech and syntactic structures observed between originals and translations, as well as between different levels of translation experience. These features are extracted the same way as the surprisal features, but based on language models trained on sentence-level reversed text. The backward language model features are popular in translation quality estimation studies and show interesting results (Duchateau et al., 2002; Rubino et al., 2013b).

3.4 Preprocessing and Tools

All data used in our experiments are sentence-split, lower-cased and tokenised using the CORENLP toolkit (Manning et al., 2014). The part-of-speech tags and syntactic trees required to extract some features are obtained with the same set of tools. For parsing, we use the probabilistic context-free grammar model trained on the Negra corpus (Brants et al., 2003) and described in (Rafferty and Manning, 2008), before flattening the trees. Both part-of-speech and flattened syntactic trees are then

\(^2\)Too many to list in the paper, a complete list is provided with the additional material submitted.

\(^3\)Originally authored texts and translations are used separately in order to model their characteristics.
delexicalised in order to remove all surface forms from the representations.

4 Results and Analysis

Below we provide details on discriminating between originally authored texts and translations, followed by the prediction of translation experience comparing professional translators and students. Finally, we evaluate features individually to determine which ones are the most useful for a particular task.

4.1 Original vs Translated Texts

Two sets of experiments are conducted to discriminate between originals and professional translations (Table 3) and originals and student translations (Table 4). For each classification task, we evaluate feature groups on the test set containing 4,000 unseen sentences balanced over two classes, reporting overall accuracy, and also precision, recall and f-score. Finally, a classification model is trained and evaluated combining all features.

### Table 3: Accuracy, precision, recall and F-measure obtained on the originals versus professional translations classification task.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acc (%)</th>
<th>Originals</th>
<th>Professional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>54.7</td>
<td>0.54 0.64 0.58</td>
<td>0.56 0.46 0.50</td>
</tr>
<tr>
<td>Surprisal</td>
<td>69.2*</td>
<td>0.66 0.77 0.71</td>
<td>0.73 0.61 0.66</td>
</tr>
<tr>
<td>Complexity</td>
<td>65.3</td>
<td>0.63 0.73 0.68</td>
<td>0.68 0.57 0.62</td>
</tr>
<tr>
<td>Distortion</td>
<td>70.0*</td>
<td>0.66 0.81 0.73</td>
<td>0.75 0.59 0.66</td>
</tr>
<tr>
<td>All</td>
<td>66.5</td>
<td>0.64 0.74 0.69</td>
<td>0.70 0.59 0.64</td>
</tr>
</tbody>
</table>

The classification of originals and student translations shows that the combination of the four feature types leads to the most accurate classifier, followed by the distortion and the surprisal sets (with equivalent accuracy results at \( p < 0.05 \)). The two latter feature sets are the best performing ones overall based on the two classification tasks. Comparing the two tasks, originally authored texts are closer to professional translations and more distant to student translations, which validates two of our hypotheses listed in Section 3.

### Table 4: Accuracy, precision, recall and F-measure obtained on the originals versus student translations classification task. Best results in bold and statistically significant winner marked with * (\( p < 0.05 \)).

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acc (%)</th>
<th>Originals</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>57.8</td>
<td>0.58 0.58 0.58</td>
<td>0.58 0.58 0.58</td>
</tr>
<tr>
<td>Surprisal</td>
<td>69.7*</td>
<td>0.69 0.72 0.70</td>
<td>0.71 0.67 0.69</td>
</tr>
<tr>
<td>Complexity</td>
<td>65.4</td>
<td>0.62 0.81 0.70</td>
<td>0.73 0.49 0.59</td>
</tr>
<tr>
<td>Distortion</td>
<td>70.8*</td>
<td>0.69 0.75 0.72</td>
<td>0.73 0.66 0.69</td>
</tr>
<tr>
<td>All</td>
<td>71.1*</td>
<td>0.69 0.76 0.72</td>
<td>0.73 0.66 0.69</td>
</tr>
</tbody>
</table>

The classification of originals and student translations reaches a maximum accuracy of 70.0% using the distortion feature set with surprisal a close second at 69.2%. The difference is not statistically significant (bootstrap resampling at \( p < 0.05 \)). They outperform the other types of features, as well as the combination of all feature types. Per class evaluation shows a similar trend with the best performing feature sets. The results show that originals and professional translations exhibit differences in terms of sequences of words, part-of-speech and syntactic tags which are captured by language model based features.
choices of the human translators as captured by our features.

Table 5: Accuracy, precision recall and F-measure obtained on the professional versus student translations classification task. Best results in bold and statistically significant winner marked with * (p < 0.05).

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acc (%)</th>
<th>Professional</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>54.5</td>
<td>0.56 0.43 0.48</td>
<td>0.54 0.66 0.59</td>
</tr>
<tr>
<td>Surprisal</td>
<td>55.7</td>
<td>0.57 0.48 0.52</td>
<td>0.55 0.64 0.59</td>
</tr>
<tr>
<td>Complexity</td>
<td>56.0</td>
<td>0.56 0.55 0.56</td>
<td>0.56 0.57 0.56</td>
</tr>
<tr>
<td>Distortion</td>
<td>57.7</td>
<td>0.58 0.55 0.56</td>
<td>0.57 0.60 0.59</td>
</tr>
<tr>
<td>All</td>
<td><strong>58.7</strong></td>
<td>0.59 0.57 <strong>0.58</strong></td>
<td>0.58 0.61 0.59</td>
</tr>
</tbody>
</table>

4.3 3-way Classification

Table 6 shows the confusion matrix obtained with the classifier trained on the combination of the four feature sets. This classifier reaches third position overall in terms of accuracy, behind distortion and surprisal sets with first and second positions, respectively. This ranking of classifiers trained on different feature sets follows the trend observed in the originals versus professional translations binary classification task.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Originals</th>
<th>Professional</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>1318</td>
<td>656</td>
<td>544</td>
</tr>
<tr>
<td>Originals</td>
<td>276</td>
<td>699</td>
<td>491</td>
</tr>
<tr>
<td>Professional</td>
<td>406</td>
<td>645</td>
<td>965</td>
</tr>
</tbody>
</table>

The training method chosen for the multi-class problem is the one against one, where individual classifiers are first trained on each binary classification task before being combined to form the final multi-class classifier. (Hsu and Lin, 2002) The results indicate that our feature sets distinguish originally authored texts from professional and student translations (first line of the matrix), while the professional translations are more difficult to separate from the two other types of text. Also, student translations have characteristics differing from originals and professional translations, which can be captured with our feature sets (last line of the matrix).

4.4 Feature Performance

Evaluating the performance of our feature sets is done by calculating the discriminative power of each feature individually which allows us to rank features according to their correlation with a class given a classification task. We follow the “f-score” measure (1) as proposed by Chen (2006):

\[
F(i) = \frac{1}{n+1} \sum_{k=1}^{n} (\bar{x}^{(+)}_{k,i} - \bar{x}^{(-)}_{k,i})^2 + \frac{1}{n-1} \sum_{k=1}^{n} (\bar{x}^{(-)}_{k,i} - \bar{x}^{(-)}_{i})^2
\]

with training vectors \(x_k\) and \(k = 1, \ldots, m\), binary classes \(n_+\) and \(n_-\) for positive and negative instances, \(\bar{x}_i\), \(\bar{x}^{(+)}_i\), \(\bar{x}^{(-)}_i\) the average of the \(i\)th feature of the whole, positive and negative instances, and \(\bar{x}^{(+)}_{k,i}\) and \(\bar{x}^{(-)}_{k,i}\) the \(i\)th feature of the \(k\)th positive or negative instance. The measure indicates how discriminative a feature is given a binary classification task. A drawback of the F-score is that it does not take into account possible features complementarity.

We report the top 10 features ranked by this method on the three binary classification tasks in Table 7. Complexity features, based on \(n\)-gram frequencies, are the most discriminative for the three tasks. The length of the \(n\)-grams varies between 1 and 3 words, indicating that longer sequences are not reliable features. The most useful resource appears to be the corpus of Political texts translated into German and used to learn the \(n\)-gram frequency distribution. The classification task involving originals and professional translations shows the usefulness of surprisal inspired features, four of them being included in the top 10. However, meta features based on the lexical surprisal measure do not appear in the top 10, and only those calculated on delexicalized flat syntactic trees are discriminative for the originals versus professional translations task.

The results obtained on individual feature discriminative power do not reflect the ones obtained using features grouped by types. Individually, features indicating complexity based on \(n\)-gram frequencies are ranked highest. However, none of the distortion features appear in the discriminative ranking while this feature type reaches the highest accuracy scores on the three binary classification tasks.
5 Conclusion

This paper presented a first step in using information density, and especially surprisal and complexity inspired features, as well as features used in translation quality estimation, as indicators of translationese for originally authored and translated text classification. We focused on separating originals and translations produced by humans with different levels of expertise and showed that translationese features based on information density and quality estimation are useful indicators of whether a text was translated or originally produced.

At the same time our experiments show that the best performing features are based on a set of quality estimation inspired distortion indicators, extracted from backward language models trained on originally authored and translated texts, followed by surprisal and complexity features. When evaluated individually according to the “f-score” measure (Chen and Lin, 2006), the most discriminative features are the complexity ones, extracted from n-gram frequency quartiles, followed by surprisal based on delexicalized syntactic parse trees.

The features used in our experiments are extracted at the word-level. As future work, we plan to extend our feature sets to information theoretic aspects of character-level indicators, such as character n-grams and language models. This approach would allow to capture sub-word information density indicators, such as morphological information (Avner et al., 2014).

References


Alexander Ehud Avner, Noam Ordan, and Shuly Wintner. 2014. Identifying translationese at the word and sub-word level. In Proceedings of ACL.


Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. 2003. Syntactic annotation of a german newspaper...
Rubino et al. (2016)  

**NAACL 2016 Submission ***. Confidential review copy. DO NOT DISTRIBUTE.**

Raphael Rubino, Jennifer Foster, Rasoul Samed Zadeh
Kaljahi, Johann Roturier, and Fred Hollowood. 2013b.
Estimating the quality of translated user-generated
Carolina Scarton and Lucia Specia. 2014.
Document-level translation quality estimation:
exploring discourse and pseudo-references. In
Proceedings of EAMT.
Lucia Specia and Jesús Gimenez. 2010. Combining
Confidence Estimation and Reference-Based Metrics
for Segment Level MT Evaluation. In Proceedings of
AMTA.
Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010.
Machine translation evaluation versus quality
Andreas Stolcke, Jing Zheng, Wen Wang, and Victor
Elke Teich. 2003. Cross-linguistic variation in system
and text: A methodology for the investigation of
translations and comparable texts, volume 5. Walter
de Gruyter.
Rules and Confidence Measures for Statistical
Machine Translation. Proceedings of Advances in
Natural Language Processing, pages 70–81.
confidence estimation for machine translation using
phrase-based translation models. In Proceedings of
EMNLP, pages 763–770.
Nicola Ueffing, Klaus Macherey, and Hermann Ney.
Vered Volansky. 2012. The features of translationese.
Master’s thesis, University of Haifa.
Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach

Ekaterina Lapshinova-Koltunski and Mihaela Vela
Saarland University
A2.2 University Campus
D-66123 Saarbrücken
{e.lapshinova,m.vela}@mx.uni-saarland.de

Abstract

In this paper, we apply text classification techniques to prove how well translated texts obey linguistic conventions of the target language measured in terms of registers, which are characterised by particular distributions of lexico-grammatical features according to a given contextual configuration. The classifiers are trained on German original data and tested on comparable English-to-German translations. Our main goal is to see if both human and machine translations comply with the non-translated target originals. The results of the present analysis provide evidence for our assumption that the usage of parallel corpora in machine translation should be treated with caution, as human translations might be prone to errors.

1 Introduction: Motivation and Goals

In the present paper, we demonstrate that both manually and automatically translated texts differ from original texts in terms of register, i.e. language variation according to context (Halliday and Hasan, 1989; Quirk et al., 1985). Similar observations were made in other studies, such as those by Gellerstam (1986), Baker (1995) and Teich (2003), who show that translations tend to share a set of lexical, syntactic and/or textual features. Several studies, including (Ozdowska and Way, 2009; Baroni and Bernardini, 2006; Kurokawa et al., 2009) and (Lembersky et al., 2012), employ computational techniques to investigate these differences quantitatively, mainly applying text classification methods.

Our main aim is to show that human translations, which are extensively deployed as data for both training and evaluation of statistical machine translation (SMT), do not necessarily obey the conventions of the target language. We define these conventions as register profiles on the basis of comparable data in the form of original, non-translated texts in the target language. These register-specific profiles are based on quantitative distributions of features characterising certain registers derived from theories described in Section 2.1 below. The non-translated data set and the corresponding register-specific features are used to train classifiers, for which we apply two different classification methods (see Section 3.4). The resulting classes serve as approximation for the standards of the target language. For the test data, we use multiple translations of the same texts produced by both humans and machines. The results of this analysis provide evidence for our assumption that we should treat the application of human translations in multilingual technologies, especially SMT (for instance, its evaluation), with caution. Our results show that there is a need for new technologies which would allow a machine-translated text to be a closer approximation to the original text in terms of its register. However, we are not aiming to provide solutions for this problem in the paper, but rather to show the importance of registers for both human and machine translation.

2 Related Work

2.1 Main notions within register theory

Studies related to register theory, e.g. by Quirk et al. (1985), Halliday and Hasan (1989) or Biber (1995), are concerned with contextual variation of languages, and state that languages vary with respect to usage context within and across languages. For example, languages may vary according to the activity of the involved participants or the relationship between speaker and addressee(s). These parameters correspond to the variables of (1) field, (2) tenor and (3) mode de-
fined in the framework of systemic functional linguistics (SFL), which describes language variation according to situational contexts; see, for instance, studies by Halliday and Hasan (1989) and Halliday (2004). These variables are associated with the corresponding lexi-co-grammatical features. Field of discourse is realised in term patterns or functional verb classes, such as activity (approach, supply, etc.), communication (answer, inform, suggest, etc.) and others. Tenor is realised in modality expressed by modal verbs (can, may, must, etc.) or stance expressions (used by speakers to convey personal attitude to the given information, e.g. adverbs like actually, certainly, amazingly, importantly). And mode is realised in information structure and textual cohesion, e.g. coreference via personal (she, he, it) and demonstrative (this, that) pronouns. Thus, differences between registers can be identified through the analysis of occurrence of lexi-co-grammatical features in these registers; see Biber’s studies on linguistic variation (Biber, 1988; Biber, 1995; Biber et al., 1999).

The field of discourse also includes experiential domain realised in the lexis. This corresponds to the notion of domain used in the machine translation community. However, it also includes colligation (morpho-syntactic preferences of words), in which grammatical categories are involved. Thus, domain is just one of the parameter features a register can have.

2.2 Register in translation

Whereas attention is paid to register settings in human translation as described by House (2014), Steiner (2004), Hansen-Schirra et al. (2012), Kruger and van Rooy (2012), De Sutter et al. (2012), Delaere and De Sutter (2013) and Neumann (2013), registers have not yet been considered much in machine translation. There are some studies in the area of SMT evaluation, e.g. those dealing with the errors in translation of new domains (Irvine et al., 2013). However, the error types concern the lexical level only, as the authors operate solely with the notion of domain (field of discourse) and not register (which includes more parameters, see Section 2.1 above). Domains reflect what a text is about, its topic. So, consideration of domain alone would classify news reporting on certain political topics together with political speeches discussing the same topics, although they belong to different registers. We expect that texts from the latter (political speeches) translated with a system trained on the former (news) would be lacking in persuasiveness, argumentation and other characteristics reflected in their lexi-co-grammatical features, for instance, imperative verbal constructions used to change the addressee’s opinion, or interrogatives as a rhetorical means. The similarity in domains would cover only the lexical level, in most cases terminology, ignoring the lexi-co-grammatical patterns specific for the given register (see the discussion on domain vs. register in (Lapshinova-Koltunski and Pal, 2014)). More recently, Zampieri and Lapshinova-Koltunski (2015) and Lapshinova-Koltunski (inpress) have shown the dominance of register-specific features of translated texts over translation-method-specific ones. Although some NLP studies, for example, those employing web resources, do argue for the importance of register conventions, see (Santini et al., 2010) among others, register remain out of the focus of machine translation. One of the few works addressing the relevance of register features for machine translation is (Petrenz, 2014), in which the author uses text features to build cross-lingual register classifiers.

2.3 The impact of target and source texts in translation quality

If languages differ in their register settings (Hansen-Schirra et al., 2012; Neumann, 2013), the register profiles of the source and the target are also different. In his work on translation quality, Steiner (2004) applies ‘the guiding norms’ for evaluation derived from both the target language and the register properties of the source. In MT evaluation, various methods and metrics of evaluation commonly rely on reference translations, which means that the relation between machine-translated texts and human translations is considered. We believe that we cannot judge the quality of a translation by merely comparing a source and a (reference) translation. Quality assessment also requires consideration of the target language conventions, i.e. those derived from comparable texts (belonging to the same registers) in a target language.

Some recent corpus-based studies on translation (Baroni and Bernardini, 2006; Koppel and Ordan, 2011) have shown that it is possible to automatically predict whether a text is an original or a
translation. Furthermore, automatic classification of original vs. translated texts found application in machine translation, especially in studies showing the impact of the nature (original vs. translation) of the text in translation and language models used in SMT. Kurokawa et al. (2009) show that for an English-to-French MT system, a translation model trained on an English-to-French data performs better than one trained on French-to-English translations. However, the ‘better performance’ of an SMT system is measured by BLEU scores (Papineni et al., 2002), indicating to which extend an SMT output comply with a reference, which is a translation itself. Inspired by Kurokawa et al. (2009)’s work, Lembersky et al. (2012) show that the BLEU score can be improved if they apply language models compiled from translated texts and not non-translated ones. They also show that language models trained on translated texts fit better to reference translations in terms of perplexity. In fact, this confirms the claim that machine translations comply more with translated rather than with non-translated texts produced by humans. It results in the improvement of the BLEU score, but not necessary leading to a better quality of machine translation. Several studies have confirmed the fact that BLEU scores should be treated carefully, see (Callison-Burch et al., 2006; Vela et al., 2014a; Vela et al., 2014b).

3 Methodology and Resources

3.1 Research questions

Following the assumption that translated language should normalise the linguistic features (like those described in 2.1 above) in order to adapt them to target language conventions, we use a classification method (using German original data for training, and translations for testing) to prove if register settings in translations correspond to those of the comparable originals. It is not our intention to directly measure the differences between originals and translations in the same language. This has been a common practice in numerous corpus-based translation studies that concentrate mostly on features in isolation, not paying much attention to their correlation: see Section 2.3 above.

Instead, we want to investigate if the register-related differences modelled for non-translated texts also apply for translation, and if they are sensitive to the variation according to the translation method involved. In fact, we model register classes for German non-translated texts, and test them on German translations from English source texts which are comparable to German non-translated ones in terms of registers. We expect that for some types of translations (e.g. human vs. machine), registers are identified more easily than for the others. We measure the accuracy scores (precision, recall and f-measure) which are class-specific numbers obtained for various sets of data: see details in Section 3.4.

Our classification analysis is structured according to the following questions: (1) Do translations from English into German correspond to German originals in their register settings? (2) Which translation can be classified best in terms of register? (3) Is there any difference between human (PT1 and PT2) and machine translations (RBMT and SMT), if register settings are concerned?

3.2 Feature selection

The input for the classifiers represents a set of features derived from register studies described in Section 2.1 above. These features constitute lexico-grammatical patterns of more abstract concepts, i.e. textual cohesion expressed via pronominal coreference or other cohesive devices, evaluative patterns (e.g. it is interesting/important that) and others. Several studies (Biber et al., 1999; Neumann, 2013), successfully employed these features for cross-lingual register analysis, showing that they reflect intra-lingual linguistic variation. In our previous work, see (Lapshinova-Koltunski, inpress), we applied a similar set of features to analyse register variation in translation.

Register features should reflect linguistic characteristics of all texts under analysis, be content-independent (do not contain terminology or keywords), be easy to interpret yielding insights on the differences between variables under analysis. So, we use groupings of nominal and verbal phrases instead of part-of-speech n-grams, as they are easier to interpret as n-grams. The set of selected features for the present analysis is outlined in Table 1. The first column denotes the extracted and analysed patterns, the second represents the corresponding linguistic features, and the third denotes the three context parameters according to register theory as previously described in Section 2.1.

The number of nominal and verbal parts-of-speech, chunks and nominalisations (ung-
nominalisations) reflect participants and processes in the field parameter. The distribution of abstract or general nouns and their comparison to other nouns gives information on the vocabulary (parameter of field). Modal verbs grouped according to different meanings defined by Biber et al. (1999), and evaluation patterns express modality and evaluation, i.e. the parameter of tenor. Content words and their proportion to the total number of word in a text represent lexical density, which is an indicator of the parameter of mode. Conjunctions, for which we analyse distributions of logico-semantic relations, belong to the parameter of mode as they serve as discourse-structuring elements. Reference, expressed either in nominal phrases or in pronouns, reflects textual cohesion (mode). Overall, we define 21 features1 representing subtypes of the categories given in Table 1.

3.3 Corpus resources

German non-translated texts (GO=German originals) used as training data for classifiers are extracted from CroCo (Hansen-Schirra et al., 2012), a corpus of both parallel and comparable texts in English and German. The dataset contains 108 texts which cover seven registers: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to share-holders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The decision to include this wide range of registers is justified by the need for heterogeneous data for our experiment. Therefore, the dataset contains both frequently machine-translated texts, e.g. SPEECH, ESSAY and INSTR, and those, which are commonly not translated with MT systems, such as FICTION or POPSCI. The number of texts per register in GO comprises approximately 36 thousand tokens.

The translation data set is smaller (50 texts) and contains multiple German translations (both human and machine) of the same English texts, see (Lapshinova-Koltunski, 2013). Translations vary in (1) translator expertise, which differentiate them into professional (PT1), and novice (PT2) translations; and in (2) translation tools, which include rule-based (RBMT) and statistical machine translation (SMT). PT1 was exported from the above mentioned corpus CroCo (Hansen-Schirra et al., 2012), which contains not only GO but also comparable German translations from English originals covering the same registers as in GO. PT2 was produced by trainee translators with at least BA degree, who have little experience in translation. All of them produced translations using different translation memories (available via OPUS2) with the help of Across3, a computer-aided translation tool which can be integrated into the usual work environment of a translator. The rule-based machine translation variant was produced with SYSTRAN64 (Systran, 2001), whereas for statistical machine translation, a Moses-based system was used which was trained with EUROPARL, a parallel corpus containing texts from the proceedings of the European parliament (Koehn, 2005). Every translation subcorpus has the same number of texts, as the data represent multiple translations of the same texts.

To extract the occurrences of register features described in 3.2, we annotate all subcorpora with information on token, lemma, part-of-speech (pos), syntactic chunks and sentence boundaries using Tree Tagger (Schmid, 1994). The features are then defined as linguistic patterns in form of the Corpus Query Processor regular expressions (Evert and Hardie, 2011), available within the CWB tools (CWB, 2010). As the procedures to annotate and to extract features are fully automatic, we expect them to influence some of the results, e.g. lexical density, which is entirely based on the pos categories assigned by Tree Tagger. So, the erroneous output of the tagger could also affect the results on the features. However, a gold-standard corpus is needed to evaluate the performance of the feature extraction, which is beyond the goals of the present work.

3.4 Classification methods

For our classification task, we train two different models by using two different classifiers on German original data. The applied techniques include (1) \( k \)-nearest-neighbors (KNN), a non-parametric method, and (2) support vector machines (SVM) with a linear kernel, a supervised method, both commonly used in text classification.

---

1Note that we select 18 only for the final classification, see details in Section 3.4.
When using KNN, the input consists of the K closest training examples in the feature space, and the output is a class membership. This method is instance-based, where each instance is compared with existing ones using a distance metric, and the distance-weighted average of the closest neighbours is used to assign a class to the new instance (Witten et al., 2011).

For our experiments we have to determine the final number for K and the most appropriate number of features used in the classification, for which the Monte Carlo cross-validation method is used (as this method provides a less variable, but more biased estimate). Having the most significant features in the set, we calculate the distribution of errors by cross-validating 10 pairs of training-validation sets and choosing K\(^5\) and the tuple (numberOfFeatures=17, K=11) is selected for our classification analysis. The classification is then performed on the translation (test) data, using the knn package (Ripley, 1996; Venables and Ripley, 2002).

Because the features that we select for classification have different measurement scales in our data, both the training and the test data are standardised using Formula 1 below.

\[
x_s = \frac{x - \text{Min}}{\text{Max} - \text{Min}}
\]

Applied to our corpus, the classification algorithm is supposed to store all available cases in GO (108 data points) and classify new cases in translation data (50 data points) based on a distance function measure, for which Euclidean distance is used.

\(^5\)with in an interval between 3 and 19

When using SVM models (Vapnik and Chervonenkis, 1974), the learning algorithm tries to find the optimal boundary between classes by maximising the distance to the nearest training data of each class. Given labelled training data, the algorithm outputs an optimal hyperplane which categorises new instances. One of the reasons why SVM are used often is their robustness towards overfitting as well as their ability to map to a high-dimensional space.

We apply SVM on the same data set as for KNN, meaning that the same standardised training (108 data points) and test (50 data points) sets, as well as the same features were selected. We also apply the same procedures, training the SVM classifier on the German originals and testing the resulting model on the German translations.

First, both classifiers are tested in the 10-fold cross-validation step (Section 4.1). Judging the performance scores in terms of precision, recall and f-measure, we decide on classes (registers) used to answer the research questions formulated in Section 3.1. As already mentioned above, these scores are class-specific and indicate the results of automatic assignment of register labels to certain non-translated texts. In case of precision, we measure the class agreement of the data with the positive labels given by the classifier. For example, there are ten German fictional texts in our data. If the classifier assigns FICTION labels to ten texts only, and all of them really belong to FICTION, then we will achieve the precision of 100%. With recall, we measure, if all translations of a certain register were assigned to the register class they should belong to. So, if we have ten fictional texts, we would have the highest recall if all of them are assigned with the FICTION label. F-measure combines both precision and recall, and is under-

---

<table>
<thead>
<tr>
<th>pattern</th>
<th>feature</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal and verbal chunks</td>
<td>participants and processes</td>
<td>field</td>
</tr>
<tr>
<td>ung-nominalisations and general nouns</td>
<td>vocabulary and style</td>
<td></td>
</tr>
<tr>
<td>modals with the meanings of permission, obligation, volition</td>
<td>modality</td>
<td>tenor</td>
</tr>
<tr>
<td>evaluative patterns</td>
<td>evaluation</td>
<td></td>
</tr>
<tr>
<td>content vs. functional words</td>
<td>lexical density</td>
<td></td>
</tr>
<tr>
<td>additive, adversative, causal, temporal, modal conjunctive relations</td>
<td>logico-semantic relations</td>
<td>mode</td>
</tr>
<tr>
<td>3rd person personal and demonstrative pronouns</td>
<td>cohesion via reference</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Features under analysis
stood as the harmonic mean of both. For the tests on translation data, we select registers for which we could achieve at least 60% of f-measure.

Next, we apply the classifiers on the translation data, which is split into different variables according to the posed research questions in Section 3.1, i.e. all translation variants or human vs. machine. As in the previous step, we also analyse the scores for precision, recall and f-measure, as our assumption is that these values would indicate if German translated texts correspond with their register settings to the non-translated German. Hence, the higher the values, the better a translation correspond to comparable originals.

4 Classification analysis

4.1 Classifier performance

In the first step, we validate the performance of our classifiers trained on German originals with the selected set of features. As we don’t have comparable data in German at hand to test the classifier, we perform 10-fold cross-validation for both KNN and SVM classifiers. The results of the cross-validation are presented in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KNN</td>
<td>SVM</td>
<td>KNN</td>
</tr>
<tr>
<td>ESSAY</td>
<td>0.43</td>
<td>0.64</td>
<td>0.70</td>
</tr>
<tr>
<td>FICTION</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>INSTR</td>
<td>1.00</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>POPSCI</td>
<td>0.75</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>SHARE</td>
<td>0.67</td>
<td>0.71</td>
<td>0.36</td>
</tr>
<tr>
<td>SPEECH</td>
<td>0.54</td>
<td>0.89</td>
<td>0.39</td>
</tr>
<tr>
<td>TOU</td>
<td>0.76</td>
<td>0.53</td>
<td>0.73</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.74</td>
<td>0.81</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 2: Classification results for GO per register

The best results are shown for fictional texts, popular-scientific texts and instruction manuals, for which the resulting f-measure amounts between 80-100%. SPEECH and SHARE reveal the lowest scores, and thus, are excluded from further analysis.

4.2 Question 1: Translations and register

Table 3 provides an overview of the f-measure values representing basically the diagonal of the confusion matrix of all classes (registers) under analysis, for the four different translation methods and two different classifiers. The table reveals that our classification algorithms perform differently depending on the register.

The best results are achieved for FICTION with both classification methods (lower performance is observed for PT2 with KNN and RBMT with SVM), where we observe f-measures up to 100%. This means that translations of English fictional texts best match the standards of German fiction. The worst results are observed for translations of political essays and popular-scientific texts, where missing correspondence with originals is observed for machine-translated texts in terms of SVM. The KNN values, although better, achieve the maximum of 53% for RBMT-POPSCI.

Misclassification results are observed for every class, varying in the translation method involved.

The classification results with both classifiers do not demonstrate the same results, e.g. SVM performs better for FICTION and INSTR, whereas KNN’s best performance is observed for ESSAY, POPSCI and TOU. Therefore, we cannot claim that certain registers are generally more difficult to be identified in translated data than others, as the performance of the classifiers vary depending not only on the register but also the translation method involved.

4.3 Question 2: The best performance

To answer the second question, we compare the average values (for all classes) for precision, recall and f-measure for each translation variant in our data, as shown in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KNN</td>
<td>SVM</td>
<td>KNN</td>
</tr>
<tr>
<td>PT1</td>
<td>0.56</td>
<td>0.49</td>
<td>0.71</td>
</tr>
<tr>
<td>PT2</td>
<td>0.53</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>RBMT</td>
<td>0.43</td>
<td>0.24</td>
<td>0.61</td>
</tr>
<tr>
<td>SMT</td>
<td>0.50</td>
<td>0.32</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 4: Average values for the classification per translation variant

Ranking translations according to the calculated values, we observe the best performance of translations by humans with both classifiers. The differences between the KNN and SVM results are caused by the differences in the approach to learning: for KNN, all K neighbours influence the classification, whereas the SVM classifier draws a line...
to separate the data points. Significance analysis confirms that the KNN results are similar for all translation varieties, as no significant difference can be observed (p-value of 0.99). This means that all translation variants correspond to comparable originals in a similar way. By contrast, the SVM values reveal variation, as the calculated p-value equals 0.03 (which is below the significance level of 0.05). Thus, we see that PT2 comply more with the register settings of the target language.

4.4 Question 3: Human vs. machine

In the following step, we compare the values for human and machine translations, analysing them per class (register). The results (see Table 5) show that both human and machine translations perform similarly, although both classifiers perform better on human translations (with the average f-measures of 0.58 vs. 0.48 for KNN and 0.52 vs. 0.33 for SVM). Our significance tests show that the results for HU vs. MT differ in terms of SVM (p-value of 1.59e-11), and is similar in terms of KNN (p-value of 0.08).

A more detailed analysis of the calculated values (presented in Figure 1) reveals much variation across registers in the results. Human translation performs better for certain registers only, i.e. ESSAY and POPSCI (both with KNN and SVM). The results for FICTION, INSTR and TOU vary depending on the classifier used. Table 6 indicates which translation method performed better for the given registers depending on the classifier used.

<table>
<thead>
<tr>
<th>register</th>
<th>KNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>HU</td>
<td>HU</td>
</tr>
<tr>
<td>FICTION</td>
<td>MT</td>
<td>HU</td>
</tr>
<tr>
<td>INSTR</td>
<td>HU</td>
<td>MT</td>
</tr>
<tr>
<td>POPSCI</td>
<td>HU</td>
<td>HU</td>
</tr>
<tr>
<td>TOU</td>
<td>MT</td>
<td>HU</td>
</tr>
</tbody>
</table>

Table 6: Performance for human and machine translation across registers

Table 3: F-measure scores for classification per translation variant and register

5 Discussion and Outlook

We have shown that translations can be classified according to register features corresponding to the target language conventions. In case of a good classification performance, translations seem to adapt these conventions. However, we also observed misclassification cases, e.g. for tourism texts or those of political essays. We suppose that the reason for this lies in the nature of translated texts which differ from comparable originals. MT systems trained with such human translations result in the same kind of non-correspondence with the register standards of the target language. This might explain the similarities in our classification results for both humans and machines. While human translation characteristics in MT are often considered to be beneficial as they can improve the BLEU scores, we believe that the application of human translation as a reference should be treated with caution. There is a need for a closer approximation of the MT outputs to the original texts in terms of register, which are possible in form of high-level language models capturing register profiles in a target language. One of the ideas here is the application of such profiles (see as conventions of the target language) to rank translated texts, which might serve as basis for new techniques of MT evaluation. However, their implementation, as well as exploitation of such profiles for MT development, need a thorough elaboration of features, which is beyond the aims of the present study. In the area of MT development, we suggest that techniques such as document-wide decoding used for other discourse phenomena in Hardmeier et al. (2012) could be promising in the improvement of register profiles in machine-translated texts.

We believe that the knowledge on the discriminative features resulting from our classification can be beneficial for natural language processing, as they indicate register-specific differences of language means. For example, Petrenz and Webber (2011) show that within a newspaper corpus, the occurrence of the word states as a verb...
Table 5: Evaluation of classification results per human and machine translation

|                  | Precision | | | | Recall | | | | F-Measure | | | |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|                  | HU KNN    | HU SVM    | MT KNN    | MT SVM    | HU KNN    | HU SVM    | MT KNN    | MT SVM    | HU KNN    | HU SVM    | MT KNN    | MT SVM    |
| ESSAY            | 0.53      | 0.67      | 0.53      | 0.00      | 0.54      | 0.12      | 0.45      | 0.00      | 0.53      | 0.20      | 0.49      | 0.00      |
| FICTION          | 0.68      | 0.88      | 0.75      | 0.80      | 1.00      | 1.00      | 1.00      | 0.83      | 0.80      | 0.93      | 0.86      | 0.78      |
| INSTR            | 0.42      | 0.28      | 0.25      | 0.40      | 0.65      | 1.00      | 0.25      | 1.00      | 0.51      | 0.44      | 0.25      | 0.57      |
| POPSCI           | 0.80      | 0.88      | 0.44      | 0.00      | 0.50      | 0.33      | 0.63      | 0.00      | 0.61      | 0.44      | 0.52      | 0.00      |
| TOU              | 0.32      | 0.24      | 0.37      | 0.19      | 0.75      | 0.80      | 0.70      | 0.90      | 0.45      | 0.37      | 0.48      | 0.31      |
| AVERAGE          | 0.55      | 0.59      | 0.47      | 0.28      | 0.69      | 0.65      | 0.61      | 0.35      | 0.58      | 0.48      | 0.52      | 0.33      |

Figure 1: Evaluation of classification results per human and machine translation

is higher in letters than in editorials, and the cues on such specific features correlating with registers may impact system performance. The knowledge from confusion matrices can thus be useful for the decision if we can use an MT system trained on texts of one register and translate texts of another register which was commonly classified as the first one in our experiments. Experiments of this kind are part of our future work, which will also include inspection of the feature weights resulting from classification. The higher the weight of a feature, the more distinctive it is for a class, regardless of its positive or negative sign. A feature ranking will help us to determine the relative discriminatory force of certain features specific for a particular register, as described by (Teich et al., 2015) in their work on register diversification in scientific writing.

We also need to have a closer look at the features contributing to misclassification, as they might also serve as translation error indicators. For this, human assessments of quality is required, which involves manual evaluation of our translation data. The manual effort would also allow us to evaluate the performance of the automatic feature extraction, which might be erroneous, as stated in Section 3.3.

6 Acknowledgement

We thank Elke Teich, Erich Steiner and all anonymous reviewers for their constructive comments. We also gratefully acknowledge the help of Heike Przybyl in preparing the final version of this article. All remaining errors and misconceptions are our own.

References


Register-Based Machine Translation Evaluation
with Text Classification Techniques

Mihaela Vela  
Saarland University, Saarbrücken, Germany  
m.vela@mx.uni-saarland.de

Ekaterina Lapshinova-Koltunski  
Saarland University, Saarbrücken, Germany  
e.lapshinova@mx.uni-saarland.de

Abstract

This paper presents a novel approach to machine translation evaluation by combining register features – characterised by particular distributions of lexico-grammatical features – with text classification techniques. The goal of this method is to compare machine translation output with comparable originals in the same language, as well as with human reference translations. The degree of similarity – in terms of register features – between machine translations and originals, and machine translations and reference translations is measured by applying two text classification methods trained on 1) originals and 2) reference translations, and tested on machine translations. The results from the experiments prove our assumption that machine translations share register features rather with human translations than with non-translated texts produced by humans. This confirms that registers are one of the most important factors that should be integrated into register-based machine translation evaluation.

1 Introduction: Motivation and Goals

The state-of-the-art in evaluating machine translation (MT) nowadays is to measure lexical and eventually syntactic and semantic overlap between a machine translation (called hypothesis translation) and a human-produced reference translation. In this paper, we present a new approach to evaluation, integrating the knowledge on register, i.e. language variation according to context, as defined by Halliday and Hasan (1989) and Quirk et al. (1985). The difference in terms of register between original and translated texts has been shown by several studies (Hansen-Schirra et al., 2012; Kruger and van Rooy, 2012; Neumann, 2013), proving that translations tend to share a set of lexical, syntactic and/or textual features. More recent investigations by Baroni and Bernardini (2006), Kurokawa et al. (2009) and Lembersky et al. (2012) applied text classification methods to automatically identify these differences.

The aim of the research presented here is twofold: 1) to show that machine translations and the corresponding reference translations are related to each other in terms of register-specific features and as a consequence of this 2) to show that hypothesis translations and human translations share more than the lexical surface. The novel idea introduced here is the notion of register-specific features which relate reference and hypothesis translations, and therefore have implications for MT (evaluation).

We measure the “closeness” between comparable non-translated originals and machine translations, as well as between human and machine translations by applying two different classification methods. The classification is performed on the basis of extracted register-specific features for two data sets. First, we use original non-translated texts as training data and ma-
chine translations as test data. In a second step, human translations are used as training data and machine translations as test data. Our assumption is that in terms of register specificity quantified in the corresponding features, MT output is closer to the corresponding reference translations than to the comparable non-translated originals. We base our hypothesis on the fact that b) translations tend to normalise towards target language conventions and that a) machine translations will adapt more to these conventions than to source texts (Diwersy et al., 2014).

The remainder of the paper is structured as follows. In Section 2, we present related work from the areas of machine translation evaluation and register theory. In Section 3 we present our research questions, describe the selected features and resources used for the experiments, as well as the applied methods. Section 4 demonstrates the results of our analysis, and in Section 5, we discuss the outcome and give an outlook on further analyses.

2 Related Work

2.1 Machine Translation Evaluation

State-of-the-art MT evaluation applies automatic language-independent metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) in order to compare MT output (hypothesis translation) with one or more human translations (reference translations). Several studies (Callison-Burch et al., 2006; Vela et al., 2014a,b) have confirmed the fact that BLEU scores should be treated carefully, thus advancing the development of new metrics. New evaluation metrics such as METEOR (Denkowski and Lavie, 2014), Asiya (González et al., 2014) and VERTa (Comelles and Atserias, 2014), are incorporating lexical, syntactic and semantic information into their scores, whereas metrics like BEER (Stanojević and Sima’an, 2014), ReVal (Gupta et al., 2015) and COMET (Vela and Tan, 2015) use machine learning approaches for MT evaluation. The accuracy of the newly introduced evaluation methods is usually proven by human evaluation inputs, more specifically by measuring the correlation of the automatically provided scores with human judgements. Human evaluation is realised by ranking MT outputs (Bojar et al., 2013, 2014; Vela and van Genabith, 2015). In addition, post-editing, which is mainly used for measuring productivity (Guerberof, 2009; Zampieri and Vela, 2014), is also a valid human evaluation method.

2.2 Main Notions within Register Theory

Studies related to register theory (Quirk et al., 1985; Halliday and Hasan, 1989; Biber, 1995) are concerned with contextual variation of languages, and state that languages vary with respect to usage context within and across languages. For example, languages may vary according to the activity of the participants involved or the relationship between speaker and addressee(s). These parameters correspond to the variables of (1) field, (2) tenor and (3) mode defined in the framework of systemic functional linguistics (SFL), which describes language variation according to situational contexts; see, for instance, Halliday and Hasan (1989) and Halliday (2004). These variables are associated with the corresponding lexicogrammatical features, e.g. field of discourse is realised in term patterns or functional verb classes, e.g. activity (approach, supply, etc.), communication (answer, inform, suggest, etc.) and others; tenor is realised in modality expressed e.g. by modal verbs (can, may, must, etc.) or stance expressions (used by speakers to convey personal attitude to the given information, e.g. adverbs like actually, certainly, amazingly, importantly, etc.); and mode is realised in information structure and textual cohesion, e.g. coreference via personal (she, he, it) and demonstrative (this, that) pronouns. Thus, differences between registers can be identified through the analysis of occurrence of lexicogrammatical features in these registers; see Biber’s studies on linguistic variation, e.g. Biber (1988), Biber (1995) or Biber et al. (1999). The field of discourse also includes experiential domain realised in the lexis. This corresponds to the notion of domain used in the machine
translation community. However, it also includes colligation (morpho-syntactic preferences of words), in which grammatical categories are involved. Thus, domain is just one of the parameter features a register can have.

2.3 Register in Translation

Several studies in systemic functional linguistics are concerned with register settings in human translation (Steiner, 2004; Hansen-Schirra et al., 2012; De Sutter et al., 2012; Neumann, 2013; House, 2014) and their application into translation practice (Vela and Hansen-Schirra, 2006; Vela et al., 2007). To our knowledge, the machine translation (including its evaluation) community has not yet taken into consideration the notion of register, at least according to the definition in the present paper. Studies in the field of MT concerned with translation errors of new domains are covering only the lexical level (Irvine et al., 2013), as the authors operate solely with the notion of domain (field of discourse) and not register (which includes more parameters, as described in Section 2.2 above). Research on adding in-domain bilingual data to the training material of SMT systems (Eck et al., 2004; Wu et al., 2008) or on application of in-domain comparable corpora (Laranjeira et al., 2014; Irvine and Callison-Burch, 2014) consider the notion of domain. However, further register features are mostly ignored.

Domains reflect what a text is about, i.e. its topic. So, consideration of domain alone would classify news reporting on certain political topics together with political speeches discussing the same topics, although they belong to different registers. We expect that texts from the latter (political speeches) translated with a system trained on the former (news) would be lacking in persuasiveness, argumentation and other characteristics reflected in their lexico-grammatical features, e.g. imperative verbal constructions used to change the addressee’s opinion, or interrogatives as a rhetorical means, etc. The similarity in domains would cover only the lexical level, in most cases terminology, ignoring the lexico-grammatical patterns specific for the given register, as shown by Lapshinova-Koltunski and Pal (2014) in their discussion on domain vs. register. Although some NLP studies employing web resources are arguing for the importance of register conventions, as by Santini et al. (2010), register remains out of the focus of machine translation. One of the few works addressing the relevance of register features for machine translation is Petrenz (2014), in which the author uses text features to build cross-lingual register classifiers.

3 Methodology and Resources

3.1 Research Questions

Following the assumption that translated language should normalise the linguistic features (like those described in Section 2.2 above) in order to adapt them to target language conventions, we use two different classification methods, KNN and SVM, to prove that in terms of register settings 1) machine translations correspond to human reference translations to a greater extent than 2) to comparable original non-translated texts in the same language. This requires a two-fold experiment design. In the first experiment, we use German original data for training and German machine translations for testing. Based on classification accuracy we can determine the “closeness” of machine translations to comparable non-translated texts in the same language. For the second experiment, we use a different data set applying the same classification methods. Human reference translations are used as training data and machine translations as test data. In this way we can observe the relation between machine translations and comparable non-translated texts in the same language, as well as between machine translations and reference translations.

We also aim at answering the following questions:
(i) Do German machine translations correspond to comparable German non-translated originals?

(ii) Are German machine translations closer to human reference translations than to comparable original German texts?

(iii) What are the main parameters influencing the classification outcome?

Our assumption is that machine translations will comply more with register standards of human-produced translated texts rather than with non-translated texts written by humans, as it was shown by Lapshinova-Koltunski and Vela (2015).

3.2 Feature Selection

For our analysis, we select a set of features derived from register studies described in Section 2.2 above. These features represent lexicogrammatical patterns of more abstract concepts, i.e. textual cohesion expressed via pronominal coreference or other cohesive devices, evaluative patterns (e.g. it is interesting/important that, etc.) and others. The selected features reflect linguistic characteristics of all texts under analysis, are content-independent (do not contain terminology or keywords), and are easy to interpret yielding insights on the differences between the analysed variables. For instance, we use groupings of specific types of phrases (e.g. nominal, verbal, etc.) instead of part-of-speech n-grams, as they are easier to interpret as n-grams. The set of selected features for our analysis is outlined in Table 1. The first column denotes the extracted and analysed patterns, the second represents the corresponding linguistic features, and the third denotes the three context parameters according to register theory as previously described in Section 2.2.

The number of nominal and verbal parts-of-speech, chunks and nominalisations (ung-nominalisations) reflects participants and processes in the field parameter. The distribution of abstract or general nouns and their comparison to other nouns gives information on the vocabulary (parameter of field). Modal verbs grouped according to different meanings (Biber et al., 1999), and evaluation patterns express modality and evaluation, i.e. the parameter of tenor. Content words and their proportion to the total number of words in a text represent lexical density, which is an indicator of the parameter of mode. Conjunctions, for which we analyse distributions of logico-semantic relations, belong to the parameter of mode as they serve as discourse-structuring elements. Reference expressed either in nominal phrases or in pronouns reflects textual cohesion (mode). Overall, we define 21 features representing subtypes of the categories given in Table 1.

<table>
<thead>
<tr>
<th>pattern</th>
<th>feature</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal and verbal chunks</td>
<td>participants and processes</td>
<td>field</td>
</tr>
<tr>
<td>ung-nominalisations and general nouns</td>
<td>vocabulary and style</td>
<td>field</td>
</tr>
<tr>
<td>modals with the meanings of permission, obligation, volition</td>
<td>modality</td>
<td>tenor</td>
</tr>
<tr>
<td>evaluative patterns</td>
<td>evaluation</td>
<td>tenor</td>
</tr>
<tr>
<td>content vs. functional words</td>
<td>lexical density</td>
<td>mode</td>
</tr>
<tr>
<td>additive, adversative, causal, temporal, modal conjunctive relations</td>
<td>logico-semantic relations</td>
<td>mode</td>
</tr>
<tr>
<td>3rd person personal and demonstrative pronouns</td>
<td>cohesion via reference</td>
<td>mode</td>
</tr>
</tbody>
</table>

Table 1: Features under analysis
However, for the final interpretation, a reduced number of features is used, which results from the validation step described in Section 3.4.1 below.

3.3 Data

In the first experimental setting, the training data set consists of German non-translated texts (GO=German originals) extracted from CroCo (Hansen-Schirra et al., 2012), a corpus of both parallel and comparable texts in English and German. The dataset contains 108 texts which cover seven registers: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The decision to include this wide range of registers is justified by the need for heterogeneous data for our experiment. Therefore, the dataset contains both frequently machine-translated texts, e.g. SPEECH, ESSAY and INSTR, and those, which are commonly not translated with MT systems, such as FICTION or POPSCI. The total number of tokens in GO is 252711.

The corresponding test data set is smaller and includes 50 texts translated from English into German with a rule-based (RBMT) and a statistical (SMT) machine translation system. The rule-based machine translations were produced with the rule-based system SYSTRAN6 (Systran, 2001). The statistical machine translations were produced with a Moses-based system\(^1\), trained with EUROPARL (Koehn, 2005), a parallel corpus containing texts from the proceedings of the European parliament. The total number of tokens in RBMT and SMT comprise 127865 and 124462 respectively. Both variants contain translations of the same texts belonging to the same registers as in the originals (training data).

In the second setting, English to German human translations (HT) – extracted from the CroCo corpus – are used for training, and comprise 100 texts (262655 tokens). For testing we use the same data as in the previous setting – machine translated texts (RBMT and SMT). Both training and test data are translations of the same source texts, which corresponds a common setting in MT evaluation. Table 2 gives an overview of the number of texts, sentences and tokens in both experiment settings.

<table>
<thead>
<tr>
<th>Setting 1</th>
<th>Setting 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>Texts</td>
<td>GO</td>
</tr>
<tr>
<td>Sentences</td>
<td>15736</td>
</tr>
<tr>
<td>Tokens</td>
<td>252711</td>
</tr>
</tbody>
</table>

Table 2: Statistics for the data sets used in experiments

To extract the occurrences of register features described in Section 3.2, we annotate both training and test sets with information on token, lemma, part-of-speech, syntactic chunks and sentence boundaries using Tree Tagger (Schmid, 1994). The availability of these annotation levels in both corpora allows us to analyse certain lexico-grammatical patterns (see Section 3.2) required for register-sensitive analysis of translation. The features are then defined as linguistic patterns and modelled as regular expressions for the Corpus Query Processor (Evert, 2005), available within the CWB tools (CWB, 2010).

3.4 Classification Approaches

For our classification task, we train two models by using two different classifiers for each experiment setting: \textit{k-nearest-neighbors} (KNN), a non-parametric method, and \textit{support vector}

\(^1\)We did not perform tuning for the SMT system.
machines (SVM) with a linear kernel, a supervised method, both commonly used in text classification. The models are then tested on German translations, and in this way, we obtain classification performance scores for:

- KNN and SVM, having German originals as training data and machine translations (English-German) as test data;
- KNN and SVM, having human translations (English-German) and again machine translations (English-German) as test data\(^2\).

The performance scores are then judged in terms of precision, recall and f-measure. These scores are class-specific and indicate the results of automatic assignment of register labels to certain machine-translated texts. In case of precision, we measure the class agreement of the data with the positive labels given by the classifier. For example, there are ten German fictional texts in our data. If the classifier assigns FICTION labels to ten texts only, and all of them really belong to FICTION, then we will achieve the precision of 100%. With recall, we measure if all translations of a certain register were assigned to the register class they should belong to. So, if we have ten fictional texts, we would have the highest recall if all of them are assigned with the FICTION label. F-measure combines both precision and recall, and is understood as the harmonic mean of both.

3.4.1 K-Nearest Neighbors (KNN)

When using KNN, the input consists of the K closest training examples in the feature space\(^3\), and the output is a class membership. This method is instance-based, where each instance is compared with existing ones using a distance metric, and the distance-weighted average of the closest neighbours is used to assign a class to the new instance, see (Aha et al., 1991; Witten et al., 2011).

For our experiments we have to determine the final number for K and the most appropriate number of features used for classification. By measuring the distribution of errors during training (by performing 10-folds cross-validation) we determined the best K=11\(^4\) and the final number of features for our setting. For our classification analysis we work with the tuple \((\text{numberOfFeatures}=17, K=11)\), performing classification on the German translation (test) data by using the knn library in Weka (Hall et al., 2009). The final list of features include:

- total words – words per text
- content words – content (lexical) words
- NP chunks – nominal chunks
- VP chunks – verbal chunks
- chunks – chunks per text
- nominal – nominal part-of-speech categories
- verbal – verbal part-of-speech categories
- adversative – adversative conjunctive relations
- causal – causal conjunctive relations
- temporal – temporal conjunctive relations
- modal – modal conjunctive relations
- ung-nominalisations – nominalisations formed with \(\text{ung-}\) suffix
- pron – personal pronouns
- dempron – demonstrative pronouns
- pronnp – nominal phrases filled with pronouns
- gnouns – general nouns
- modals denoting permission

\(^2\)Note that human and machine translations are translation variants of the same English source text.

\(^3\)In our experiments, the features are quantified by their frequency in the corresponding data set.

\(^4\)The final value for K was chosen from an interval between 3 and 19
3.4.2 Support Vector Machines (SVM)

When using SVM models (Vapnik and Chervonenkis, 1974), the learning algorithm tries to find the optimal boundary between classes by maximising the distance to the nearest training data of each class. Given German labelled training data, the algorithm outputs an optimal hyperplane which categorises new instances, here German translations. We use the same list of features for the classification with SVM as in the classification with KNN. We perform SVM classification with a 10-fold cross-validation.

The cross-validation in the training phase has shown that registers SPEECH and SHARE show low accuracy, especially in the first experiment setting (when German original data is used). For this reason, we exclude these registers from further analyses.

3.5 Experimental Setup

In the first setting, both classifiers are supposed to store all cases from German originals (108 data points) with the corresponding register labels available. New cases from test data, which are machine translations in this case (50+50 data points), are then classified, i.e. assigned register labels. Classification is performed on the basis of distance function measure, for which Euclidean distance is used. The results of automatic assignment are indicated with scores (precision, recall and f-measure), which are used to measure the class agreement of the data with the positive labels given by the classifiers.

In the second setting, both classifiers store all labelled cases available in human translations (100 data points). The trained model is then applied on the same machine translations as in the first setting (50+50 data points). And again, we use precision, recall and f-measure to judge the class agreement of the data with the positive labels given by the classifiers.

4 Classification Results

As already mentioned above, the results of both classification algorithms are analysed in terms of precision, recall and f-measure. In case of precision, we measure the class agreement of the data with the positive labels given by the classifier. Our assumption is that precision values would indicate if the test data correspond to the training data in terms of the register settings. Hence, in the first experiment setting, the higher the precision, the better a machine translation corresponds to comparable originals, whereas in the second setting, the higher the precision, the better a machine translation corresponds to human translations.

4.1 Setting 1: Originals vs. Machine Translations

<table>
<thead>
<tr>
<th></th>
<th>ESSAY</th>
<th>FICTION</th>
<th>INSTR</th>
<th>POPSCI</th>
<th>TOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.36</td>
<td>0.86</td>
<td>0.17</td>
<td>0.53</td>
<td>0.50</td>
</tr>
<tr>
<td>SVM</td>
<td>0.00</td>
<td>0.75</td>
<td>0.55</td>
<td>0.00</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 3: F-measure scores for classification per machine translation variant and register in the first setting

Table 3 provides the confusion matrix for the five registers and the two MT outputs split on the two classification methods. Concerning both the classification method and the MT system, we notice that FICTION is the register performing best, achieving an f-measure of 100% for the combination SMT with SVM (SMT-SVM). By ranking the registers according to f-measure

---

5 One of the reasons why SVM are often used is their robustness towards overfitting, as well as their ability to map to a high-dimensional space.
we observe that ESSAY performs well for the classification with KNN but fails with SVM. FICTION performs also very well achieving an f-measure of 100% for the combination SMT-SVM even. INSTR and TOU perform similar in terms of f-measure for both KNN and SVM, the lowest f-measure values being measured for POPSCI. POPSCI fails for the combinations SMT-KNN and RBMT-SVM, and even for KNN, the values for the f-measure can be interpreted as fair (0.53% for RBMT and 0.46% for SMT). These results imply that overall, FICTION performs best for both translation types and both classifiers, which indicates that translated fictional texts obey best to original fictional texts in terms of register settings.

Obviously, the data used for developing/training MT systems as well as the classification method play an important role. Registers like ESSAY (political essays) and INSTR (manuals) are usually used for training SMT systems, whereas registers like FICTION (fictional texts) and TOU (tourism leaflets) are less likely used for training MT systems. The more surprising are here the results for FICTION. The big difference in the results for both classifiers can be explained in the distinction between these two classification techniques. KNN uses all training data (predefined neighbours) in classification, while for SVM the maximised distance (margin) to the nearest example of each class plays a crucial role (all non-support vectors being discarded), thus influencing the difference in the classification outcome.

The overview of the performance of both MT systems in Table 4 (split on register and classification method) reveals that SMT performs better than RBMT for ESSAY, FICTION and INSTR. The classification failure of the RBMT and SMT produced POPSCI translations for SVM indicates that SVM is not the appropriate classification method for this kind of texts. The fact that we observe contradictory results with both classifiers for ESSAY and POPSCI prevents us to claim that certain registers are generally more difficult to be identified in translated data than others.

<table>
<thead>
<tr>
<th>register</th>
<th>KNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>SMT</td>
<td>RBMT=SMT</td>
</tr>
<tr>
<td>FICTION</td>
<td>RBMT=SMT</td>
<td>SMT</td>
</tr>
<tr>
<td>INSTR</td>
<td>SMT</td>
<td>SMT</td>
</tr>
<tr>
<td>POPSCI</td>
<td>RBMT</td>
<td>RBMT=SMT</td>
</tr>
<tr>
<td>TOU</td>
<td>RBMT</td>
<td>RBMT</td>
</tr>
</tbody>
</table>

Table 4: Performance of KNN and SVM across registers based on f-measure in the first setting

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN SVM</td>
<td>KNN SVM</td>
<td>KNN SVM</td>
</tr>
<tr>
<td>RBMT</td>
<td>0.43 0.24</td>
<td>0.61 0.56</td>
</tr>
<tr>
<td>SMT</td>
<td>0.50 0.32</td>
<td>0.61 0.53</td>
</tr>
</tbody>
</table>

Table 5: Classification accuracy per machine translation variant in the first setting

In the last step, we analyse the performance of the MT systems disregarding registers, see Table 5. We notice that results do not differ much in the MT systems under analysis, which means that register agreement between non-translated and machine-translated texts is not dependent on the method involved in translation. This is proven by Pearson’s chi-square test, which confirms our observation, as for both the KNN and the SVM results, p-value is higher than 0.05 (0.15 and 0.74 respectively).
4.2 Setting 2: Reference Translations vs. Machine Translations

The same experiments and analysis steps as in Section 4.1 are performed for the second setting, where human reference translations are used as training data and hypothesis translations as test data. However, we observe a strong improvement in the results presented in Table 6. The registers ESSAY and FICTION achieve the best performance, showing an f-measure of up to 100%. Over all, f-measure remains over 50% for all registers, which means that classifiers perform also well for TOU, POPSCI and INSTR. These results can, in fact, serve as indicators that human and machine translations have more similarities, sharing register features.

<table>
<thead>
<tr>
<th>Register</th>
<th>RBMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>FICTION</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>INSTR</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>POPSCI</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>TOU</td>
<td>0.57</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 6: F-measure scores for classification per machine translation variant and register in the second setting

In this setting, ESSAY matches register features of human translations best, leading to the assumption that the texts used for developing and training MT systems play a key role. In contrast to the results in the first setting, the difference between the results of the two classification methods for ESSAY is minor. The lowest value is scored by INSTR with an f-measure of 0.54 for RBMT-KNN.

Table 7 demonstrates that, different from the results in the first experiment in Section 4.1, RBMT performs better than SMT when compared to human reference translations, with some exceptions. RBMT-INSTR and RBMT-TOU are better classified than the same registers translated with the SMT system, if the results from KNN are taken into account.

<table>
<thead>
<tr>
<th>Register</th>
<th>KNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>RBMT=SMT</td>
<td>RBMT=SMT</td>
</tr>
<tr>
<td>FICTION</td>
<td>RBMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>INSTR</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>POPSCI</td>
<td>RBMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>TOU</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
</tbody>
</table>

Table 7: Performance of the classification methods across registers based on the f-measure in the second setting

Similar tendencies are observed if we average the scores for registers.

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>SVM</td>
<td>KNN</td>
</tr>
<tr>
<td>RBMT</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>SMT</td>
<td>0.82</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 8: Classification accuracy per translation variant in the second setting

We observe in Table 8 that, in terms of MT system, RBMT performs almost always better than SMT, and in terms of classification method applied, SVM performs always better than KNN. However, significance test show that the difference between both machine translation systems is not significant, as the computed p-values exceed 0.05 for both KNN and SVM scores.
(0.78 and 0.97 respectively), confirming not only our assumption that reference and hypothesis translations are similar if register features are considered, but also that both machine-translated variants are similar as well.

5 Conclusion

The results of the presented experiments have proven our assumption that machine translations share their register features rather with human produced translations than with human produced non-translated texts, regardless the method involved in translation. In Figure 1, we illustrate this graphically, unifying the results for both MT systems and presenting them separately for each classification method. The graphs shows that f-measure scores for the classification of reference translations (HT) against machine translations (MT), marked here with triangles, are higher than for the classification of originals (GO) against machine translations (MT).

Figure 1: Evaluation of classification results per classification method

Setting1: Original vs. Machine Translation The results of the first experiment show that register settings of most German machine translations do not comply with the register settings of non-translated German, this being shown also by Lapshinova-Koltunski and Vela (2015). This is especially valid for ESSAY and POPSCI, which show low performance in most classification scenarios, not adapting the target language register conventions. The good performance for FICTION is an indicator that fictional texts adapt to the conventions of the target language. However, as known from Neumann (2013), German and English fictional texts share a lot of features, which might also influence the performance of the classifiers. We suppose that the influence of the source language register conventions might also have an impact on the outcome. To prove this, we would need to perform additional experiments, in which English source texts should be used as training data.

Furthermore, as the performed significance tests have shown that the difference between RBMT and SMT is not meaningful, our suggestions apply regardless the translation method involved. In case of SMT, the results are apparently influenced by the training data used, as classification performs better for registers which are commonly used for SMT training, i.e. like

---

6Here, we mean “shining through” of the source language as defined by Teich (2003).
ESSAY and INSTR. Off-the-shelf RBMT systems, like the one used here, are developed to cover a more general aspect of language which also essentially complicate the adaptation of register features.

**Setting 2: Reference vs. Machine Translation** The results of the second experiment presented in Section 4 have proven our assumption that the overlap between hypothesis and reference translations is higher than between hypothesis translations and comparable non-translated texts. On the one hand, this corresponds to our intuition in Lapshinova-Koltunski and Vela (2015), where we show that both human and machine translations do not correspond to comparable German originals, suggesting that both machine and manual translations should have more in common. In fact, this is also shown by Lembersky et al. (2012), who demonstrate that the BLEU score can be improved if they apply language models compiled from translated texts and not non-translated ones. They also show that language models trained on translated texts fit better to reference translations in terms of perplexity. In fact, this confirms our claim that machine translations comply more with translated rather than with non-translated texts produced by humans. This results in the improvement of the BLEU score, but not necessary leading to a better quality of machine translation.

Following the results from both experimental setting, we argue that register features should be integrated into MT evaluation process, as an additional layer to the already existing automatic metrics. As future work, we would like to test this hypothesis by combining and correlating the results presented here with state-of-the-art evaluation metrics.

**References**


268


