Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation

Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation

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

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

Abstract

In this paper, we analyse alignment discrepancies for discourse structures in English-German parallel data – sentence pairs, in which discourse structures in target or source texts have no alignment in the corresponding parallel sentences. The discourse-related structures are designed in form of linguistic patterns based on the information delivered by automatic part-of-speech and dependency annotation. In addition to alignment errors (existing structures left unaligned), these alignment discrepancies can be caused by language contrasts or through the phenomena of explicitation and implicitation in the translation process. We propose a new approach including new type of resources for corpus-based language contrast analysis and apply it to study and classify the contrasts found in our English-German parallel corpus. As unaligned discourse structures may also result in the loss of discourse information in the MT training data, we hope to deliver information in support of discourse-aware machine translation (MT).

1 Introduction

All human languages provide means to create coherence and cohesion in texts, but the precise structures used to achieve this vary even across closely related languages. In this paper, we introduce an automatic method to extract examples of cross-linguistically divergent discourse structures from a corpus of parallel text, creating a new type of resource that is useful for the discovery and description of discourse-related language contrasts. This type of analysis is useful from the point of view of contrastive linguistics, and it can also provide researchers interested in discourse-level machine translation (MT) with a collection of data to guide their intuitions about how text-level phenomena are affected in translation. Our method is strongly data-driven; it enables a bottom-up approach to linguistic analysis that starts from individual occurrences of cross-linguistic correspondences without being constrained by existing linguistic assumptions and theoretical frameworks.

The data source in our analysis is a sentence- and word-aligned parallel corpus, the same type of resource that is typically used for training MT systems. We begin by defining a set of surface patterns that identify the discourse structures of interest and permit their automatic extraction. We then use the word alignments to establish correspondences between the languages. We particularly focus on those cases where there is a relevant pattern in one language, but the word aligner is unable to find a corresponding structure in the other. Such *alignment discrepancies* can simply be due to alignment errors, but they can also stem from systematic language contrasts (Grishina and Stede, 2015, p. 19–20) or from the phenomena of explicitation and implicitation in the translation process.

Our general goal is to explore these alignment discrepancies and analyse their causes. We use a corpus of English-German translations that we automatically annotate for part-of-speech and dependency information. Alignment discrepancies are detected with the help of sentence and word alignment of the annotated structures. Thus we do not use any manually annotated resources, and linguistic knowledge involved is rather of shallow character. Specific cases of extracted discrepancies represented through linguistic patterns are then manually analysed. We concentrate on English-to-German translations, as although these

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 73–81,
Copenhagen, Denmark, September 8, 2017.

two languages are typologically close, this language pair is still among those that are hard for machine translation.

This paper is structured as follows: in the following section (Section 2), we define the phenomenon under analysis and explain the problem. Section 3 provides information on related works. In Section 4, we describe the data, methods and procedures applied for our analysis. Section 5 presents the results. In Section 6, we discuss the outcome of the study and outline the ideas for future work.

2 Defining the Problem

In this paper, we focus on the analysis of English-German parallel data – aligned sentence pairs, in which discourse-related structures in target or source texts have no alignment in the corresponding parallel sentences. The discourse-related structures we consider are defined as potential elements of coreference chains that can be either personal or demonstrative pronouns. These structures are designed in form of linguistic patterns based on the information delivered by automatic part-of-speech and dependency annotation and include both bare pronouns, as *she* and *it* or *this* and *that*, and determiners modifying nouns – parts of full nominal phrases, as *this system* and *the system* in (1).

(1) *..all these chemicals ultimately boost the activity of the brain's reward system... goosing this system makes us feel good... But new research indicates that chronic drug use induces changes in the structure and function of the system.*

As these are parts of coreference chains, they contribute to the overall coherence and hence carry part of the discourse information in both the source and the target language.

Linguistic means expressing coreference exist in both languages. However, the choice between referring expressions is governed by language-specific constraints. For instance, pronouns and adjectives in German are subject to grammatical gender agreement, whereas in English, only person pronouns have this marking and adjectives (for instance, in nominal ellipsis) are unmarked. Such differences in the realisation give rise to transformation patterns in translation, for instance *he* – *der* in (2), which can be obtained from parallel data on the basis of word-level alignment.

(2) *Then we take this piece of paper and give it to a fellow student and he must make us a drawing out of it. – Dann nehmen wir dieses Blatt Papier und geben es einem Kommilitonen und der muss uns daraus eine Zeichnung machen.*

However, in some cases, these differences may cause alignment discrepancies. For instance, German pronominal adverbs like *damit* in example (3) can function as a referring expression (*damit* refers to an event expressed through the whole preceding clause, but can also establish a conjunctive relation). English does not have a direct equivalent for this form. So, the English translation example from a parallel corpus does not preserve this coreference chain.

(3) *Die demographischen Kurven verraten, dass der Sozialstaat von den Jüngeren nicht mehr zu finanzieren ist. Damit versinkt das Land nicht in einer beinah unvergleichlichen Krise, wie manchmal behauptet wird. – The demographic curves reveal that the welfare state can no longer be financed by the younger members of society. This does not mean that the country is descending into an unparalleled crisis ...*

This would result in an alignment discrepancy attributed to language contrasts.

At the same time, alignment discrepancies can also be attributed to the translation process and the phenomenon of explicitation based on the Explicitation Hypothesis, formulated in its most prominent form by Blum-Kulka (1986), who assumes elements in the target text are expressed more explicitly than in the source text. For example, the full nominal phrase *die Aufgabe (the task)* in the German translation in (4) is lexically more explicit than the demonstrative *that*. Its counterpart is called implicitation.

(4) *You want your employees to do what you ask them to do, and if they've done that, then they can do extra. – Sie erwarten von Ihren Angestellten, dass sie tun worum Sie sie gebeten haben, wenn sie die Aufgabe ausgeführt haben, können sie Zusätzliches tun.*

In a parallel corpus of English-German translations, we use automatic word alignment to extract transformation patterns. Those sentence pairs which contain a discourse structure in either the source or the target sentence and for which no transformation patterns could be extracted are defined as alignment discrepancies.

3 Related Work

The method that we use to extract transformation patterns is similar to coreference annotation projection applied by Postolache et al. (2006) and by Grishina & Stede (2015). Both studies use data manually annotated for coreference relations. In our approach, we use automatic annotations only that allow us to define candidate referring expressions – linguistic expressions that are potential members of a coreference chain (not resolved by a human annotator).

Postolache et al. (2006) mark patterns containing heads of the resulting referring expression in the target language aligned with heads of the source referring expressions. Although they mention the situations when the source head is not aligned with any target word or no words of the source referring expressions are aligned with any target words, they do not consider these cases of alignment discrepancies in their analysis.

Grishina & Stede (2015) apply a direct projection algorithm on parallel data to automatically produce coreference annotations for two target languages without exploiting any linguistic knowledge of the languages. However, they describe a number of projection problems, when a referring expression is present in both source and target text but is not projected correctly. They analyse non-equivalences in translation from a linguistic point of view but could not find enough evidence to characterise them as systematic, as the dataset they use is very limited. However, the cases that they describe can be attributed to language contrasts or the effects of translation process. In our study, we use more data creating a resource that can be used for further systematic description of alignment discrepancies and their sources. We suggest that these sources can be classified into three categories: (1) alignment errors; (2) language contrasts and (3) translation process. A number of studies (Kunz and Steiner, 2012; Kunz and Lapshinova-Koltunski, 2015; Novak and Nedoluzhko, 2015) have shown that although the coreference relation is shared across all languages, they may differ considerably in the range of referring expressions.

The phenomenon of explicitation in translation is often understood to occur when a translation explicitly realises meanings that were implicit in its source text. In terms of discourse phenomena, this would mean that a source text does not contain linguistic markers that trigger some discourse relations, whereas its translation does, as was analysed by Meyer & Webber (2013) or by Becher (2011b), including also the opposite process of implicitation.

In other studies, explicitation is seen if a translated text realises meanings with more explicit means than the source text does. In relation to coreference, some referring expressions can be more explicit than the others, as in example (4) in Section 2 above. For instance, Becher (2011a, p. 98) presents a scale for the explicitness of various referring expressions for the language pair English-German.

Most of these studies start from the description of the expressions existing in the language systems they compare, and analyse the distributions of these categories with corpus-based methods. This can be defined as a top-down procedure – starting from what is given (in theories and grammars) and looking for the contrasts in a huge number of language examples represented in corpus data. In our approach, we perform in a different way – we start with the corpus data and try to detect patterns revealing language contrasts or the phenomena of explicitation/implicitation that we define in form of alignment discrepancies.

4 Resources, Tools and Methods

4.1 Data

Our corpus data consists of talks given at the TED conference[1]. It is taken from the training set of the IWSLT 2015 MT evaluation campaign[2], which in turn uses texts downloaded from the TED web site.

We need to mention that the translations of TED talks are rather subtitle than translations, and consequently, there exist some genre-/register specific transformations in this parallel data. However, the transformations in the TED talks are also interesting, especially because the latter have been frequently used as training data for MT.

[1] http://www.ted.com

[2] https://wit3.fbk.eu/mt.php?release=2015-01

We automatically annotated the corpus data using a pipeline of standard tools. The texts in both languages were preprocessed with Penn Treebank tokeniser and Punkt sentence splitter with the language-specific sentence splitting models bundled with NLTK (Bird et al., 2009). Then, the corpus was tagged with the Marmot tagger (Mueller et al., 2013) for the part-of-speech information and parsed for dependency information with the MATE tools (Bohnet, 2010). The tagger and parser were trained on version 1.0 of the Universal Dependency treebank (Nivre et al., 2015).

Word alignment was performed in both direction with *mgiza*[3] and models 1-HMM-3-4, using the training scripts bundled with the Moses machine translation software[4] and default settings. The alignments were symmetrised with the grow-diag-final-and heuristic (Koehn et al., 2003).

	sentences	**tokens**
English	214,889	3,940,079
German	227,649	3,678,503

Table 1: Corpus size

The total number of parallel segments amounts to 194,370 (see details in Table 1).

4.2 Pattern extraction

Using the part-of-speech and dependency annotations, we compiled lists of discourse-related structures defined in terms of lexico-grammatical patterns (combination of part-of-speech tags and grammatical functions that were produced by the parser) for both English and German texts. While the discourse structures we study may be composed of multiple words, we find that they can often be identified reliably with patterns anchored to single words. We select pronouns and demonstratives (which also include definite articles) only (corresponding to the part-of-speech tags 'DET' and 'PRON').

Then, we extracted parallel patterns from the above described data using the word-level alignment. The patterns are based on $1:N$ word alignments linking the word identified by our pattern (for instance *which DET-nsubj* in example (5)) to 1 or more words in the other language (*dies PRON-dobj* in example (5)). If a word has multiple alignment links, multiple output records were generated, one for each aligned target language word.

(5) `which DET-nsubj → dies PRON-dobj`
Educational researcher Benjamin Bloom, in 1984, posed what's called the 2 sigma problem, which he observed by studying three populations. – 1984 veröffentlichte der Bildungsforscher Benjamin Bloom etwas, das '2-Sigma-Problem' heißt. Er beobachtete dies bei drei Populationen.

The resulting data also contains sentence pairs for which no corresponding structure was found in either the source or the target language. These are the cases of alignment discrepancies in discourse-related structures that we select for our analysis. We count the occurrences of the alignment discrepancy patterns with the aim to answer the following questions: (1) Which are the most frequent ones in English? (2) Which are the most frequent ones in German?

In our corpus, English is always the source and German is the target, but we can search discourse-related patterns in the English sources and see what are the corresponding structures in the German translations and which structures are missing. And in the same way, we can search in the German translations and analyse the aligned English sources. This allows us to discover which discourse phenomena 'get lost' in the translation data due to the missing alignment. We can also measure the amount of these discrepancies – perform quantitative analysis, and analyse the underlying causes of these discrepancies in a qualitative analysis. These might include: (1) language contrasts that include both differences in language system and differences of idiomatic character, e.g. collocation use; (2) translation process phenomena such as explicitation – when a German translated sentence contains a discourse pattern which was not aligned to any discourse structure in the corresponding English source sentence, and implicitation – when the English original sentence contains a marked discourse pattern which was not aligned to any discourse structure in the corresponding translated sentence in German; (3) other possible causes, including errors.

[3] https://github.com/moses-smt/mgiza
[4] http://www.statmt.org/moses/

5 Analyses and Results

5.1 General observations

On the total, we extract 26 patterns (types) of discourse structures marked in the German translations, for which no English alignment was automatically assigned (explicitation candidates). The total number of unaligned cases is around 11% in both language settings.

In the English source sentences, there were 14 discourse patterns, for which the alignment in the corresponding German translations is missing (implicitation candidates). The total number of occurring cases (measured by tokens) is also higher for German (69,851) than for English (57,608), which on the one hand, may be interpreted as an evidence for more explicitation than implicitation phenomena in translation. And on the other hand, it may indicate that German has more discourse-related structures that differ from those available in English.

In Table 2, we provide an overview of the 10 most frequent discourse-related structures that were found in the German translation data, for which no corresponding discourse structures were aligned in the English sources.

freq.abs	pattern	example
29868	DET-det	*der Fall*
18026	PRON-nsubj	*er, sie*
10986	PRON-dobj	*ihn, sie*
3525	PRON-nmod	*sein, ihr*
3383	PRON-det	*diese, einige*
1481	PRON-nsubjpass	*das, dieses*
1439	PRON-iobj	*ihm, ihr*
530	PRON-dep	*daran, dafür*
297	PRON-neg	*kein*
48	PRON-appos	*etwas, alles*

Table 2: Patterns in German with no alignment in the corresponding English data

freq.abs	pattern	example
23145	DET-det	*the things*
19030	PRON-nsubj	*he, they*
6798	PRON-nmod	*his, their*
4341	PRON-dobj	*him, them*
1764	DET-nsubj	*this, that*
990	DET-nmod	*which, that*
650	DET-dobj	*this, that*
516	PRON-iobj	*him, them*
253	DET-neg	*no*
54	PRON-conj	*what*

Table 3: Patterns in the English sentences with no alignment in the corresponding German translations

pattern	EN	DE
DET-det	23145	29868
DET-dobj	650	24
DET-nmod	950	18
DET-nsubj	1764	44
PRON-det	14	3383
PRON-dobj	4341	10986
PRON-iobj	516	1439
PRON-nmod	6798	3525
PRON-nsubj	19030	18026

Table 4: Patterns shared by English and German

Table 3 presents an overview of the 10 most frequent discourse-related structures in the English sources, for which no alignment was found in the corresponding translations into German.

DET-det is the most frequent structure in both languages, followed by PRON-subj and PRON-nmod or PRON-dobj (the ranking of the latter two is different in English and German). Further (less frequent) discourse-related structures vary across languages, with English showing preferences for demonstratives (DET) and German – for personal pronouns (that also include relatives in the universal part-of-speech tagset). If the full lists (with 24 and 14 patterns) is considered, we see that PRON and DET are more evenly distributed (53% PRON vs. 47% DET) in English than in German (57% PRON vs. 43% DET).

It is interesting that eight out of the most frequent structures in the 'English' list are shared (occur in both lists). We outline all the shared patterns (nine in total) along with their frequencies in both English and German in Table 4.

5.2 Observations on particular patterns

In the following, we perform a manual qualitative analysis of the most frequent patterns (DET-det and PRON-nsubj) that are shared by both languages. The information on their categorisation frequencies is derived automatically on the basis of extracted patterns containing word information. For instance, structures like der-DET-det[5] are defined as cases of the definite article use, and the structures like der-PRON-nsubj represent relative pronouns. This manual analysis provides us with the information on possible causes of alignment discrepancies. However, at this stage, we do not provide the information on the distribution of these causes in our data.

[5] *der* is one of the forms of the German definite article

DET-det Most cases (ca. 96%) concern the German translations containing definite articles that may trigger a coreference relation between the noun phrase that contains this article and another noun phrase or a clause, as *die Aufgabe* in example (4) in Section 2 above, and for which no alignment was found in the English sources.

Manual analysis of the data sample shows that the discrepancies are often caused by the variation in article use in the expression of generic reference in both languages: in German, generic meaning is expressed with a definite noun, whereas in the English source, it is expressed with a bare noun (often in plural), see examples *people/die Leute*, *conversations/die...Unterhaltung*, *technology/die Technologie* in (6).

(6) a. *You know, it's just like the hail goes out and people are ready to help. – Es ist einfach so, jemand ruft um Hilfe, und die Leute stehen zur Hilfe bereit.*

b. *And we use conversations with each other to learn how to have conversations with ourselves. – Wir benutzen die gegenseitige Unterhaltung, um zu lernen, wie wir Gespräche untereinander führen.*

c. *We turn to technology to help us feel connected in ways we can comfortably control. – Wir wenden uns der Technologie zu, um uns auf Arten und Weisen verbunden zu fühlen, die wir bequem kontrollieren können.*

Many studies have claimed that there is variation in article use in the expression of generic reference in German (Krifka et al., 1995; Oosterhof, 2004), especially in relation to plural generics. German plural generics can be used both as definite nominal phrases and as bare nouns, whereas definite plurals in English cannot be interpreted generically. However, Barton et al. (2015) provide the only empirical analysis known to us, but concentrate on plural generics only. We believe that our approach creates a good foundation (and resource) for a more detailed quantitative analysis of such cases.

In other cases, the discrepancy between definite constructions in German has a rather idiomatic character, as in example (7).

(7) *But in the process, we set ourselves up to be isolated. – Aber dabei fallen wir der Isolation direkt vor die Füße.*

Some individual sentence pairs revealed the phenomena of explicitation, for instance, *der Fall* ('the case") in example (8) is used in German translation to explicate the information given through the ellipsis of the clause *but it's not cheesy* in English.

(8) *You would expect it to be cheesy, but it 's not. – Man könnte annehmen, dass so etwas kitschig ist, aber dem ist nicht der Fall.*

Most cases of the DET-det structure in the English sources missing alignment in the corresponding German translations are also definite noun phrases (ca. 85%). Manual analysis of a sample reveals that most of these cases are alignment errors. This means that the German translation also contains the corresponding definite nominal phrase which was not automatically aligned to the English article.

The phenomenon of implicitation was represented by individual cases that we observed in the data, e.g. in (9), where the English source is more explicit than the corresponding translation.

(9) *Secondly, there had to be an acceptance that we were not going to be able to use all of this vacant land in the way that we had before and maybe for some time to come. – Zweitens musste es eine Übereinkunft geben, dass wir das gesamte brachliegende Land nicht wie vorher nutzen können würden, vielleicht für längere Zeit nicht.*

PRON-nsubj In the German translations, many PRON-subj structures with no alignment in the corresponding English sources are represented by personal pronouns (ca. 54% out of all cases). Around 46% of these pronouns are 1st and 2nd person pronouns that are used for speaker and addressee reference. In many cases, both the source and the target sentence contain this reference type that was not automatically aligned and thus, an error occurred. Addressee and speaker reference is very common in our dataset, as this is one of the specific features of the register under analysis – public talks by experts (mostly addressed to laypeople).

The remaining structures are 3rd person pronouns, among which we observe some interesting

cases, for instances, differences in the expression of impersonal meaning in English and German, as seen in example (10).

(10) a. *A reaction to the medication the clinic gave me for my depression left me suicidal. – Die Medikamente, die sie mir in der Ambulanz gegen meine Depressionen gaben, führten bei mir zu Selbstmordgedanken.*

b. *People say, "I'll tell you what's wrong with having a conversation... – sagen sie, "Ich sage dir, was verkehrt daran ist...*

They are followed relative pronouns (ca. 31% out of all cases) that introduce a relative clause in the German translations. However, their English sources do not contain any relative clauses and the information is expressed in a different construction, as illustrated in example (11).

(11) a. *A polar bear swimming in the Arctic, by Paul Nicklen. – Ein Eisbär, der in der Arktis schwimmt, aufgenommen von Paul Nicklen.*

b. *Across the generations, I see that people can't get enough of each other... – Über alle Generationen hinweg sehe ich Menschen, die nicht genug voneinander bekommen...*

The English sentence in example (11-a) contains a non-finite *ing*-clause. This clause type has direct equivalents in form of present participle in German *schwimmend* ("swimming"). However, the English *ing*-form is used much more widely than the German present participle (Durrell, 2011, p.281–285). In particular, participial clauses are restricted to formal written registers in German and can sound stilted and they are used much less frequently than clauses with *ing*-forms in English (Durrell, 2011, p.281–285). To our knowledge, there are no corpus-based studies confirming this quantitatively. Königs (2011) provides a number of examples as possibilities of translation equivalents for English *ing*-clauses. However, statistical evidence is missing. We believe that our dataset can be used as a resource for this kind of empirical evidence.

Explicitation examples related to this structure include various way of the source reformulation, as in example (12). Here, a nominal phrase was reformulated into a nominal phrase with a clause containing the exophoric pronoun *es*.

(12) *Clouds are the most egalitarian of nature's displays, because we all have a good, fantastic view of the sky. – Wolken sind die größten Gleichmacher, wenn es um die Schönheit der Natur geht, weil wir alle einen gleich guten Blick auf den Himmel haben.*

50% of the PRON-nsubj structures in the English sources that were not aligned to any structures in German include speaker and addressee references. This discrepancy is a clear indicator of the contrasts in pragmatics and style of speeches in English and German and goes in hand with what was stated by House (2014) who provides several dimensions of such contrasts, e.g. addressee (English) vs. content (German) orientation in texts.

(13) a. *If you have fluid with no wall to surround it and keep pressure up, you have a puddle. – Eine Flüssigkeit ohne eine Wand, die sie umgibt und den Druck aufrechterhält, ist eine Pfütze.*

b. *And if you go there, you say, " Hey, everybody's really healthy." Und wenn man dorthin geht und sagt: "Hey, jeder ist kerngesund."*

In example (13-a), the English *you* does not have any correspondences in the German translation, whereas *you* in example (13-b) is transferred to *man* ("one").

The other 50% of discrepancy cases include the third person pronouns, with *it* being most frequent among other forms (43% out of all 3rd person pronouns and 21% of all the PRON-nsubj structures).

These cases also reveal language contrasts such as differences between certain syntactic constructions in English in German. For instance, the German coordinated clause with a negation *an manchen Tagen nicht* in (14-a) does not require a repetition of the subject, whereas the English clause does.

In example (14-b), *it* introduces a cleft sentence construction. These are frequent in English but

used much less frequently in German, where the topic can be shifted into initial position before the verb (Durrell, 2011, p. 455).

(14) a. *Some days it goes up and some days it doesn't go up. – An manchen Tagen geht er hoch und an manchen Tagen nicht.*

b. *And so it was that day that we decided we needed to build a crisis text hotline. – Und an diesem Tag beschlossen wir, dass wir eine Krisen-SMS-Hotline einrichten mussten.*

6 Discussion and Future Work

To our knowledge, this paper is the first attempt to quantitatively describe alignment discrepancies between English-German discourse-related phenomena from a language contrastive perspective. This approach is novel and can be characterised as data-driven, as we use "bottom-up" procedures instead of theory-driven ones that start from the grammar-based contrasts and then use data to find quantitative evidence. This is a new approach of contrast discovery.

Although we concentrated on a limited number of patterns only and described some particular causes of the discrepancies, we were able to obtain interesting observations, e.g. those on the article use with generics or the use of non-finite constructions and their alternatives in German in English. Although these cases are described in traditional grammars, corpus data shows a different behaviour, especially when spoken data is concerned.

We were not able to provide much evidence for systematic translation-process-driven discrepancies. However, we could see that they are also present in our data. We believe that a more detailed quantitative and qualitative analysis of discrepancy sources would provide more corpus evidence for the variation across the two languages under analysis. Our approach, as well as the parallel dataset created allows for such an analysis.

Moreover, the information on systematic discrepancies could serve the task of alignment improvement. For instance, we observed a great number of cases when a pronoun does not have or need a corresponding element in the parallel sentence. These cases are important for MT model development. Naive models for pronouns often lead to overgeneration of such elements in the target language. Having the information on such cases, we could think of ways of integrating them into the models to avoid the overgeneration.

Our future work will include a more detailed analysis of discrepancy sources. For language contrasts, we will investigate further patterns that are less frequent but not less important. It would be also interesting to look into the patterns that occur either in the English or in the German sentences only. Besides, we will extend our analysis on explicitation using Klaudy's classification of various types of explicitations as a starting point (Klaudy, 2008). Then, we will define a scale for coreferential explicitness based on Kunz's reduced scale of Accessibility (Kunz, 2010, p. 76) and existing analyses of connective explicitation (Denturck, 2012; Zufferey and Cartoni, 2014).

Acknowledgements

Christian Hardmeier was supported by the Swedish Research Council under project 2012-916 *Discourse-Oriented Statistical Machine Translation*. We used computing resources on the Abel cluster, owned by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR), provided through the Nordic Language Processing Laboratory (NLPL).

References

Dagmar Barton, Nadine Kolb, and Tanja Kupisch. 2015. Definite article use with generic reference in german: an empirical study. *Zeitschrift für Sprachwissenschaft* 34:147–173. https://doi.org/10.1515/zfs-2015-0009.

Viktor Becher. 2011a. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis, Universität Hamburg.

Viktor Becher. 2011b. When and why do translators add connectives? a corpus-based study. *Target* 23.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media.

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

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference*

on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, pages 89–97. http://aclweb.org/anthology/C10-1011.

Kathelijne Denturck. 2012. Explicitation vs. implicitation: a bidirectional corpus-based analysis of causal connectives in french and dutch translations. *ACROSS LANGUAGES AND CULTURES* 13(2):211–227. http://dx.doi.org/10.1556/Acr.13.2012.2.5.

Martin Durrell. 2011. *Hammer's German Grammar and Usage*. Routledge, London and New York, 5 edition.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora, Beijing, China*. page 14.

Juliane House. 2014. *Translation Quality Assessment. Past and Present*. Routledge.

Kinga Klaudy. 2008. Explicitation. In Mona Baker and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, Routledge, London & New York, pages 104–108. 2 edition.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton (Canada), pages 48–54.

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

Manfred Krifka, Francis J. Pelletier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. Genericity: An introduction. In Gregory N. Carlson and Francis J. Pelletier, editors, *The generic book*, University of Chicago Press, Chicago, IL, pages 1–124.

K.A. Kunz. 2010. *Variation in English and German Nominal Coreference: A Study of Political Essays*. Saarbrücker Beiträge zur Sprach- und Translationswissenschaft. Peter Lang. https://books.google.de/books?id=F_jmEbmeGnoC.

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

Kerstin Kunz and Erich Steiner. 2012. Towards a comparison of cohesive reference in english and german: System and text. In M. Taboada, S. Doval Suárez, and E. González Álvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*, Equinox, London.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 19–26. http://www.aclweb.org/anthology/W13-3303.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 322–332. http://www.aclweb.org/anthology/D13-1032.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal dependencies 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-1464.

Michael Novak and Anna Nedoluzhko. 2015. Correspondences between czech and english coreferential expressions. *Discours [En ligne]* 16. http://discours.revues.org/9058.

Albert Oosterhof. 2004. In Fred Karlsson, editor, *Proceedings of the 20 th Scandinavian Conference of Linguistics*. University of Helsinki, Helsinki, page 1–22.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Tranferring coreference chains through word alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target* 26(3):361–384.