The Use of Parallel and Comparable Data for Analysis of Abstract Anaphora in German and English

Stefanie Dipper*, Melanie Seiss†, Heike Zinsmeister†

*Ruhr University Bochum
dipper@linguistics.rub.de

†Konstanz University
heike.zinsmeister, melanie.seiss@uni-konstanz.de

Abstract

Parallel corpora — original texts aligned with their translations — are a widely used resource in computational linguistics. Translation studies have shown that translated texts often differ systematically from comparable original texts. Translators tend to be faithful to structures of the original texts, resulting in a "shining through" of the original language preferences in the translated text. Translators also tend to make their translations most comprehensible with the effect that translated texts can be more explicit than their source texts. Motivated by the need to use a parallel resource for cross-linguistic feature induction in abstract anaphora resolution, this paper investigates properties of English and German texts in the Europarl corpus, taking into account both general features such as sentence length as well as task-dependent features such as the distribution of demonstrative noun phrases. The investigation is based on the entire Europarl corpus as well as on a small subset thereof, which has been manually annotated. The results indicate English translated texts are sufficiently "authentic" to be used as training data for anaphora resolution; results for German texts are less conclusive, though.

Keywords: Abstract anaphors, comparable corpora, parallel corpora

1. Introduction

This paper presents a validation of parallel texts in comparison to comparable texts in the Europarl corpus (Koehn, 2005). *Parallel texts* refer to bi-texts in a source language (L_o, the original language) and its translation to a target language (L_t), which have been aligned at the sentence level. *Comparable texts* are texts in different languages or varieties that deal with the same overall topic.

Our domain of application is the resolution of abstract anaphora. We address the question whether translated texts (e.g., translations into English: EN_t) are sufficiently similar to original texts of the same language (EN_o) to be used as empirical evidence for feature induction in this domain, or whether original texts only should be used for this purpose.

Abstract anaphora denote anaphoric relations between some anaphoric expression and an antecedent that refers to an abstract object like an event, fact or proposition (cf. Asher (1993)). In the classical example by Byron (2002), the pronoun *it* (underlined in (1a)) refers to an *event*: the migration of penguins to Fiji. In (1b), the demonstrative pronoun *that* refers to the *fact* that penguins migrate to Fiji in the fall.

(1) a. Each Fall, penguins migrate to Fiji. It happens just before the eggs hatch.

b. Each Fall, penguins migrate to Fiji. That's why I'm going there next month.

In (1), the anaphoric elements are pronouns. In this paper, we mainly consider anaphoric noun phrases, as in (2), which is taken from the Europarl corpus. In this example, the NP *this task* refers to the *activity* (a specific type of event) of investigating the best ways of promoting a system.

(2) The Commission will investigate the best ways of promoting this system across the Community and will involve the European Parliament in this task.

This study is motivated by a larger project of analyzing abstract anaphora (Dipper and Zinsmeister, 2010; Dipper and Zinsmeister, to appear; Dipper et al., 2011; Zinsmeister et al., submitted), which pursues an approach of bootstrapping annotation from German to English and vice versa.

The paper is organized as follows. Section 2 provides the background of our study: we define the core concepts of parallel and comparable data, address findings of translation studies on how translated texts differ from comparable original data, and present related work on the use of parallel and comparable corpora. Section 3 introduces our corpora, and Section 4 presents results from comparing general and anaphora-related properties. Section 5 concludes the paper.

2. Background

2.1. Parallel and comparable data

Our study aims at contributing to the debate whether translated text can be used as (training) data in the same way as original text, as is commonly done in computational linguistics. Before going into the details of our approach, we define the relations that hold between different types of text.

A multilingual corpus, such as the Europarl corpus, often consists of texts in various source languages and translations of them into multiple other languages. In our study, we concentrate on texts in German and English. Fig. 1 shows the four subcorpora that we deal with: DE_o (original German texts), DE_t (German texts translated from English), EN_o (original English texts) and EN_t (English texts translated from German).

$$\begin{tikzcd}[column sep=huge, row sep=huge]
DE_o \arrow[r, leftrightarrow, "parallel"] \arrow[d, leftrightarrow, "comp_{corp}"'] \arrow[dr, leftrightarrow] & EN_t \arrow[d, leftrightarrow, "comp_{corp}"] \arrow[dl, leftrightarrow, "comp_{trans}"'] \\
EN_o \arrow[r, leftrightarrow, "parallel"'] & DE_t
\end{tikzcd}$$

Figure 1: Relations between the four subcorpora *DE*$_o$, *DE*$_t$, *EN*$_o$, and *EN*$_t$: (i) parallel, (ii) comparable in the corpus-linguistic sense (*comp*$_{corp}$), and (iii) comparable in the translation-studies sense (*comp*$_{trans}$)

The subcorpora *DE*$_o$ and *EN*$_t$ (and, similarly, *EN*$_o$ and *DE*$_t$) are *parallel* corpora, i.e. original texts along with their translations.

On the other hand, the subcorpora *DE*$_o$ and *EN*$_o$ (and *DE*$_t$ and *EN*$_t$) are comparable corpora, i.e. corpora in different languages that deal with the same overall topic and are from the same overall register. This notion of comparable corpora is usually used in corpus-linguistic research. Hence, we call this type of relation *comparable*$_{corp}$.

Finally, the subcorpora *DE*$_o$ and *DE*$_t$ (and *EN*$_o$ and *EN*$_t$) are also comparable corpora, in that they represent varieties of the same language. Translation studies usually refer to such corpora as comparable, hence we call this type of relation *comparable*$_{trans}$.

2.2. Translation effects

Using parallel texts for cross-linguistic investigations obviously benefits from the fact that the aligned units convey the same meaning and allow for direct comparison of how this meaning is encoded in the two languages. However, cross-linguistic use of parallel texts also has its limitations due to various factors.

First, translated texts can differ systematically from their source texts due to language-inherent reasons, e.g., see Vinay and Darbelnet (1995), Dorr (1994). Klaudy (2008) lists stylistic preferences and cultural differences as further factors that can result in language-specific differences in translations.

Second, the translation process itself has been shown to have an impact on the translated text, i.e., there can be a translation bias (Baker, 1993; Čulo et al., 2008). The translated text might be affected, e.g., by the *shining through* of source language preferences if the translation is too faithful to the source text, cf. Teich (2003). Another effect is described by the *explicitation hypothesis*, which assumes that translators usually strive to make their translations as comprehensible as possible. As a consequence, the translation might make explicit what was implicit in the source text (Vinay and Darbelnet, 1995; Blum-Kulka, 1986).[1]

For our study, we expect factors of the first type to result in differences between languages, i.e., in parallel and comparable$_{corp}$ corpora. Factors of the second type (shining through and the explicitation hypothesis) should show up as differences between original and translated texts, i.e., in comparable$_{trans}$ texts.

2.3. Related work

The exploitation of parallel corpora in Natural Language Processing has been growing in recent years.[2] One reason is annotation projection for under-resourced languages, in which linguistic annotation is transferred from one language to another, when relevant resources and tools are only available in the former language (e.g., Bentivogli and Pianta (2005)).

There are some studies that compare original and translated texts in the Europarl corpus. For instance, van Halteren (2008) shows that based on word *n-grams* it is possible to identify the source language in Europarl translations with accuracies between 87.2–96.7%. Cartoni et al. (2011) investigate the use of discourse connectives in original and translated French texts from Europarl. They find that translated texts contain significantly more discourse connectives than original texts. Korzen and Gylling (2011) find considerable differences between the sentence lengths of original and translated texts in Italian and Danish texts. These findings suggest that one has to look further into the properties of translated texts before using them as a resource for linguistic feature induction.

Multilingual corpora have been annotated for investigations in (abstract) anaphora resolution in Recasens (2008), Navarretta and Olsen (2008), Navarretta (2008), Pradhan et al. (2007), Weischedel et al. (2010). These projects deal with comparable$_{corp}$ rather than parallel corpora.

Annotation of parallel texts has been performed in Vieira et al. (2002), who use a subcorpus from the parallel MLCC corpus.[3] They investigate demonstrative NPs in French and Portuguese. Results for both languages are similar: demonstrative NPs predominantly have an abstract head noun. In their study, they do not distinguish between original and translated texts.

3. Corpus

We chose texts from the Europarl corpus (Release v3, Koehn (2005)) as the basis of our study, which consists of transcripts of European Parliament debates. Speakers (usually) deliver their contributions ('turns') in their native language, and professional translators provide official translations into the other EU languages.

For this study, we only consider turns by German native speakers (*DE*$_o$) and the corresponding English translations (*EN*$_t$), as well as contributions by English native speakers

[1] For a recent survey and critical assessment of the explicitation hypothesis, see Becher (2011, Ch. 2).

[2] See, for example, the workshops on Annotation and Exploitation of Parallel Corpora (Ahrenberg et al., 2010; Simov et al., 2011).

[3] The MLCC corpus contains written questions asked by members of the European Parliament and the corresponding answers from the European Commission. `http://catalog.elra.info/product_info.php?products_id=764`

(*EN_o*) and their German translations (*DE_t*).[4] The translations have been aligned with their originals on the basis of Europarl's align units. Our corpus consists of 12,800 German original turns with 4.9 M tokens, and 11,500 English original turns with 3.4 M tokens.

3.1. Automatic preprocessing

Automatic preprocessing of the German and English subcorpora included POS tagging and chunking by the TreeTagger (Schmid, 1994), with German and English language models as provided by the official TreeTagger website.[5]

In addition, we automatically marked selected abstract noun chunks, which possibly function as abstract anaphors. Noun chunks that would be selected fulfill two conditions: first, they contain a demonstrative determiner — such noun phrases are usually used anaphorically. Second, the head noun is part of a pre-defined set of so-called *label nouns*.[6] The label nouns are highly inspired by the list of English abstract nouns provided by Francis (1994).

For the English data, the set comprises 211 types of label nouns, which have been extracted in their singular and plural forms. Examples for English label nouns in the Europarl corpus are *report, proposal, agreement, issue, point*, etc. For the German data, the most probable German translations of the English label nouns have been used. This resulted in 452 German types of label nouns for which the inflected forms have been marked in the German data. (For more details on the selection of label nouns, see Zinsmeister et al. (submitted).)

3.2. Manual annotation of a subcorpus

Most of the label nouns described above are unambiguously abstract, with some exceptions such as *area* or *report*. In contrast, pronominal anaphors are (mostly) ambiguous and can refer to entities other than abstract objects. We created a small manually-annotated subcorpus called *Anaphora Corpus* of approximately 100 turns for each language, in which annotators disambiguated pre-marked pronominals and demonstrative label-noun chunks. Manual annotation also provided information about the antecedent of pronominal anaphors, and about the position, function, etc. of the anaphoric element. (For a more detailed description of the manual annotation of pronominal abstract anaphors, see Dipper et al. (2011), for the annotation of label nouns, see Zinsmeister et al. (submitted).)

4. Results

The texts that we investigate are highly similar in the sense that they are all from the same text type, namely parliament debates. They deal with different topics, depending on the agenda of the current session, but our basic assumption is that the choice of topic should have no significant impact on low-level properties such as, e.g., the number of nouns. In this study, we focus on features that are considered relevant for anaphora resolution, and the resolution of abstract anaphora in particular. That is, we mainly investigate the distributions of noun phrases, in particular demonstrative label-noun chunks, which are likely to be abstract anaphors. The investigations are based on the parallel and comparable corpora described in Section 3.

As the study aims at testing the similarity of original and translated texts, the main focus is on differences between comparable$_{trans}$ turns, i.e., we compare *DE_o* with *DE_t*, and *EN_o* with *EN_t*. In addition, we compare the language pairings *DE_o–EN_o* with *DE_o–EN_t*, and *EN_o–DE_o* with *EN_o–DE_t*. This can shed light on translational effects: if, e.g., the pairing *DE_o–EN_t* turns out more similar than the pairing *DE_o–EN_o*, this could be a shining through effect.

The comparison would profit from deep linguistic processing (as has been shown by authorship attribution studies, e.g., Kaster et al. (2005)). Such processing is only provided in the manually annotated *Anaphora Corpus*. For investigating anaphora-related properties of the complete Europarl data, we therefore have to use approximations, such as the ratio of definite or demonstrative noun chunks as rough approximations of anaphoric elements.[7]

4.1. Sentence length

We start by comparing the average sentence length as a highly general measure for similarity. The average sentence length is calculated as the average number of tokens per sentence, as identified in the preprocessing.[8]

Corpus	#Sent	Tok/Sent	95% CI	SD
DE_o	220,609	22.2	22.1 . . 22.3	13.8
DE_t	162,528	24.2	24.1 . . 24.2	13.2
EN_o	147,375	26.6	26.5 . . 26.6	14.1
EN_t	184,579	28.8	28.7 . . 28.9	16.2

Table 1: Average sentence lengths: total number of sentences, average number of tokens per sentence, 95% confidence interval, and standard deviation

On average, English sentences are clearly longer than German sentences, as can be seen in Table 1. This could be, among others, attributed to the different realizations of compound nouns in both languages.

Comparing *DE_o* with *DE_t*, and *EN_o* with *EN_t*, we observe that in both languages, sentences of translated texts tend to

[4] The original language of each turn was determined by means of the language tags provided in the Europarl corpus, and complemented by a lookup in databases listing the members of the EU Parliament along with their nationalities.

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. The German chunker was trained on the Tiger Treebank (Brants et al., 2004), the English chunker on the Penn Treebank (Marcus et al., 1993). The POS tagsets used in these annotations are the STTS tagset for German (Schiller et al., 1999), and the UPenn Treebank tagset for English (Santorini, 1990).

[6] In the biomedical literature, label nouns are referred to as 'sortal nouns', e.g. Castaño et al. (2002).

[7] Fraurud (1992) showed that 60.9% of the definite noun phrases in her sample were first mentions, i.e. not used anaphorically. However, it is clear that definite anaphors are the default: only 8.3% of the indefinite noun phrases were anaphoric.

[8] Punctuation marks have been included.

be longer than sentences of original texts. This could be viewed as an effect of explicitation.

The differences in average sentence length between the subcorpora are all statistically significant, as can be seen from the non-overlapping 95% confidence intervals. However, the confidence intervals are very small. This is an effect of the large sample size, i.e., the large number of sentences that contribute to the mean.

It is a general problem in corpus linguistics that statistical hypothesis testing becomes hard to interpret when the sample size is large. Gries (2005) suggests to employ measures of *effect size*, in particular *Cohen's d* [9] (also called *standarized mean difference*), to quantify the amount of observed differences in means independently of the sample size. Cohen's d ranges from $d = 0$, if no effect is observed, to infinity. An effect size of d between 0.2 to 0.3 is considered *small*, $d = 0.5$ *medium*, and $d = 0.8$ to infinity *large* (Cohen, 1988). Negative polarity means that the second mean in the equation is larger than the first one.

Corpora	$\bar{x}_1$	$\bar{x}_2$	Cohen's d
DE$_o$–DE$_t$	22.2	24.2	–0.14
EN$_o$–EN$_t$	26.6	28.8	–0.15
DE$_o$–EN$_o$	22.2	26.6	–0.31
DE$_o$–EN$_t$	22.2	28.8	–0.44
EN$_o$–DE$_o$	26.6	22.2	0.31
EN$_o$–DE$_t$	26.6	24.2	0.18

Table 2: Sentence lengths; $\bar{x}_1$ are $\bar{x}_2$ are the means of the two corpora (cf. Table 1)

The first two rows in Table 2 show that the effect sizes of the standardized mean difference between *DE$_o$* and *DE$_t$* on the one hand, and *EN$_o$* and *EN$_t$* on the other hand, are small ($d = -0.14, d = -0.15$). These differences can be ignored.

The comparable$_{corp}$ pair *DE$_o$–EN$_o$* shows medium differences between the languages ($d = -0.31$).[10] This confirms the significant results displayed in Table 1.

Turning to the parallel pairs (*DE$_o$–EN$_t$*, and *EN$_o$–DE$_t$*), the medium effect could either become more pronounced than in the comparable$_{corp}$ pair, which could indicate an effect of explicitation, or it could become less pronounced, which would rather indicate an effect of shining through. Interestingly, we find indications for explicitation in *EN$_t$* (d increases from –0.31 to –0.44) and for shining through in *DE$_t$* (d decreases from 0.31 to 0.18).

4.2. Nouns in general

We start our task-related comparison by looking at the distribution of nouns and noun chunks (NC) in general. To compare the frequencies between the different subcorpora, we normalize the observed noun frequencies by token frequencies. In addition, we compare the ratios of nouns per clause, definite NCs per NCs in general and the proportion of demonstrative NCs among definite NCs.[11]

Corpora	Nouns/Tok	Nouns/Cl	Def/NC	Dem/Def
DE$_o$– DE$_t$	**–0.34**	–0.17	**–0.38**	–0.04
EN$_o$–EN$_t$	0.10	–0.01	0.12	–0.03
DE$_o$–EN$_o$	0.55	–1.53	0.72	–0.72
DE$_o$–EN$_t$	0.66	–1.64	0.85	-0.76
EN$_o$–DE$_o$	**–0.55**	1.53	**–0.72**	0.72
EN$_o$–DE$_t$	**–0.91**	1.39	**–1.13**	0.69

Table 4: Cohen's d: nouns normalized by the number of tokens and clauses; definite and demonstrative NCs normalized by the number of NCs. Important effect sizes are given in boldface.

Our null hypothesis is that the number of nouns, noun chunks, definite and demonstrative noun chunks should be similar across comparable$_{trans}$ corpora, i.e., across *DE$_o$* and *DE$_t$*, and across *EN$_o$* and *EN$_t$*, respectively. Statistical significance tests (Welch's t-test) reject the null hypothesis in most cases, as is shown in Table 3.[12]

However, as we have seen above (Section 4.1), such significance scores are hard to interpret with large sample sizes. Hence, we again turn to Cohen's d as an alternative measure. Table 4 depicts the effect sizes of the comparisons.

Looking first at the comparable$_{trans}$ corpora, we see that the effect sizes for English are small for all features—despite the significant differences shown in Table 3. In contrast, the distribution of nouns (Nouns/Tok) and definite NCs (Def/NC) in German clearly differ in both subcorpora. Turning to the comparable$_{corp}$ and parallel pairs, the picture is similar: The effect sizes of *EN$_o$* and *EN$_t$* do not increase or decrease considerably. In German, however, the distribution of nouns and definite NCs in translated texts, again, deviates considerably from the distribution in original texts. In particular, we observe an overuse of nouns and definite NCs, which could be an expliciteness effect (assuming that definite NCs are more explicit than indefinite NCs).

[9] Cohen's d according to R package *MAd* (function *mean_to_d*) $d = \frac{\bar{x}_1 - \bar{x}_2}{sd_{within}}$, with $sd_{within} = \sqrt{\frac{(n_1-1)sd_1{}^2+(n_2-1)sd_2{}^2}{n_1+n_2-2}}$.

[10] This difference is of course expected in different languages like German and English.

[11] The corresponding approximations for the German and English corpora are:

- Nouns: DE: `pos="NN"`; EN: `pos="(NN|NNS)"`. The nominal tags cannot be easily compared across both languages, though. For instance, English gerunds are usually tagged as verbs, whereas nominalized infinitives in German are tagged as nouns.
- NCs: DE: `cat="(NC|PC)"`; EN: `cat="NC"`
- Definite NCs: DE: `(pos="ART" & word="d.*" | pos="(APPART|PPOSAT)")`; EN: `(pos="DT" & word="the" | pos="PP$")`, plus demonstrative NCs
- Demonstrative NCs: DE: `pos="PDAT"`; EN: `pos="DT" & word="(this|these|that|those")`

As a heuristics to determine the number of clauses, we counted verb chunks (VC) (for German and English: `cat="VC"`).

[12] Frequencies of nouns per clauses (Nouns/Clause) within turns do not differ significantly between *EN$_o$* and *EN$_t$*, according to a Welch's t-test: $t = -0.5678, df = 23439.24, p = 0.5702$.

Corpus	Nouns/Token		Nouns/Clause		Def/NC		Dem/Def	
DE_o	19.2 ± 3.5	***	78.8 ± 28.3	***	36.9 ± 9.1	***	9.9 ± 8.7	***
DE_t	20.3 ± 3.2		83.6 ± 28.3		40.4 ± 9.2		10.3 ± 8.3	
EN_o	17.3 ± 3.5	***	140.7 ± 50.5	n.s.	30.7 ± 7.8	***	17.2 ± 11.5	*
EN_t	16.9 ± 3.4		141.0 ± 45.5		29.9 ± 7.3		17.5 ± 11.1	

Table 3: Average frequencies (in %) and standard deviation in comparable$_{trans}$ corpora. Nouns/Token: nouns per tokens; Nouns/Clause: nouns per clause; Def/NC: definite (incl. demonstrative) noun chunks per noun chunks; Dem/Def: demonstrative noun chunks per definite noun chunks. Significance tests (Welch's t-test) refer to pairs of original and translated texts; significance levels: *** $p < .001$; ** $p < .01$; * $p < .05$; n.s. *not significant*.

To sum up the findings of this section, English translated texts are rather similar to English original texts, whereas German translations deviate from German originals.

4.3. Label nouns

We next investigate the use of typical label nouns such as *fact, situation*, based on our predefined sets of label nouns (see Sec. 3). In DE_o, these label nouns represent 3.48% of all nouns, in EN_o, 4.07%. For both languages, the ratios are higher in translated texts: 4.01% in DE_t, 4.25% in EN_t.

We next compare the individual frequencies of these nouns across all corpora, normalized against the total number of nouns in the respective subcorpus (and multiplied by 1 million).[13] Only nouns with (normalized) frequencies greater than 100 were considered; outliers have been removed.[14]

Fig. 2 displays the results for German (left plot) and English (right plot). The solid lines denote the number of label nouns that occur more often in translations (*overuse*), the dashed lines nouns that occur less often in translations (*underuse*). The plots show that in both languages, overuse dominates underuse: In German, 22% (100 out of 452) nouns show overuse vs. 17% (75 out of 452) nouns show underuse. In English, there are 39% (83 out of 211) overuse vs. 33% (69 out of 211) underuse nouns. The difference scores (y-axis) indicate the (normalized) ratios between original and translated frequencies. For instance, a score of 2 for an overuse noun indicates that the noun occurs twice as often in the translated than in the original texts. Especially with overuse nouns, the differences between the frequencies can be enormous, see below.

We now turn to individual frequencies of label nouns rather than general tendencies. Pairwise comparison of label nouns frequencies in original and translated texts shows that they are clearly correlated.[15]

Fig. 3 show the frequencies of label nouns in German (left plot) and English (right plot), comparing original and translated texts. Again, only nouns with more than 100 (normalized) occurrences in both texts have been included. Furthermore, outliers with extremely high frequencies in either the original or translated corpus and outliers with extreme differences between both corpora are not displayed. The heights of the boxplots clearly illustrate the overuse of label nouns in translations, already mentioned above.

Table 5 lists the label nouns with most extreme overuse and underuse. For instance, the noun *Aussprache* 'debate' occurs 7.1 times more often in DE_t than in DE_o. The figures show that the differences between original and translated noun frequencies are much more important with overuse ("Increase") than with underuse ("Decrease"). These overuses result in some sort of constricted vocabulary in the translations.

Disregarding individual noun preferences, the overall distributions of increasing and decreasing frequencies of label nouns are rather similar and they turn out not to be significantly different.[16] Taken that label nouns are approximations for abstract anaphora, we conclude that the use of abstract anaphora as such is comparable in original and translated texts. This does not hold for the lexical realizations, though. Further investigations are needed to decide whether the lexical overuse and underuse of particular nouns also effects the distribution of the semantic types of abstract anaphors, such as fact or event.

4.4. Anaphora corpus

The findings for the manually annotated Anaphora Corpus are mainly in line with the findings of the entire Europarl corpus. As an overall tendency, the differences between translated and original texts are not significant.

We found no significant difference for function (subject, object, other) in the original and translated versions (for both German and English). For position, the only statistical significant difference concerns the discourse-linked left periphery ("pre-field") in German: compared to DE_o, DE_t uses abstract anaphors less often in the pre-field. The difference could be due to shining through of English information structure, and possibly reflects the fact that English does not have a corresponding discourse-linked position that can be easily occupied by abstract anaphors. (For a more detailed discussion of the results of the Anaphora

[13] The total number of nouns (tokens) in the subcorpora are: DE_o: 890k (4.7m); DE_t: 1.2m (5.7m); EN_o: 660k (3.8m); EN_t: 860k (5.0m). As an example of normalized frequencies, consider the noun *Aussprache* 'debate': raw frequencies in DE_o: 226, in DE_t: 2106; normalized frequencies in DE_o: 255, in DE_t: 1803, cf. Table 5.

[14] Outliers are defined as data points which are more than 1.5 * the interquartile range (Q3-Q1) away from the interquartile boundaries.

[15] For German, Kendall's tau yields: $\tau = .81$ (.82); for English: $\tau = .83$ (.82) (figures in parentheses: with outliers removed).

[16] A Mann-Whitney test (= Wilcoxon rank sum test) yields for German: $W = 3981, p = 0.4871$, and for English: $W = 3335, p = 0.08133$.

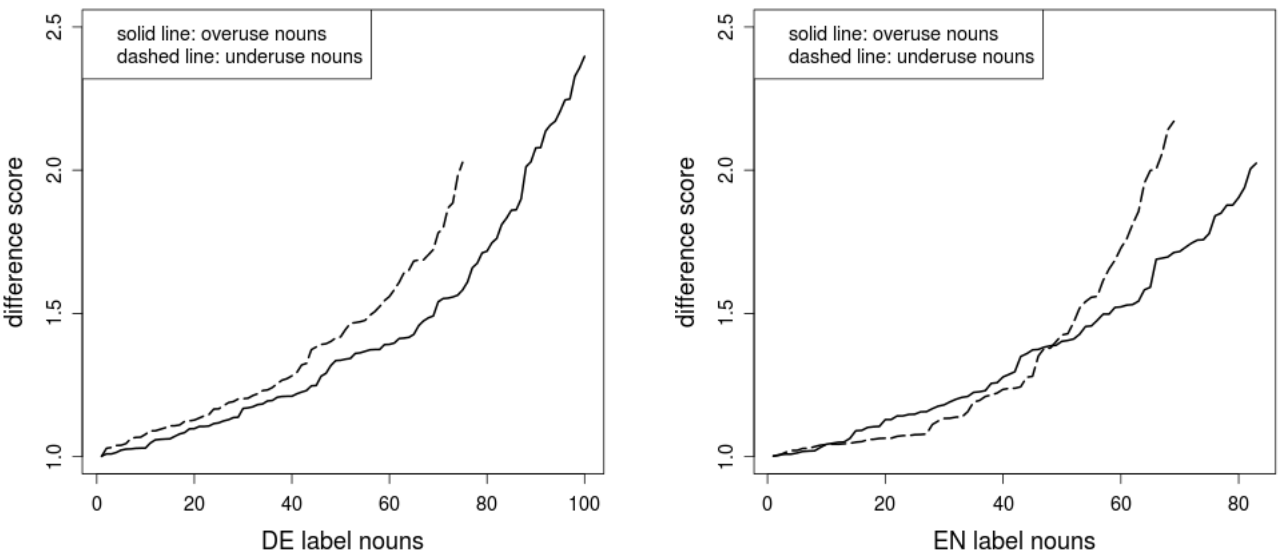

Figure 2: Number of nouns that occur more often (solid line) or less often (dashed line) in translations than in original texts (left plot: German, right plot: English). The score indicates the ratio between original and translated uses.

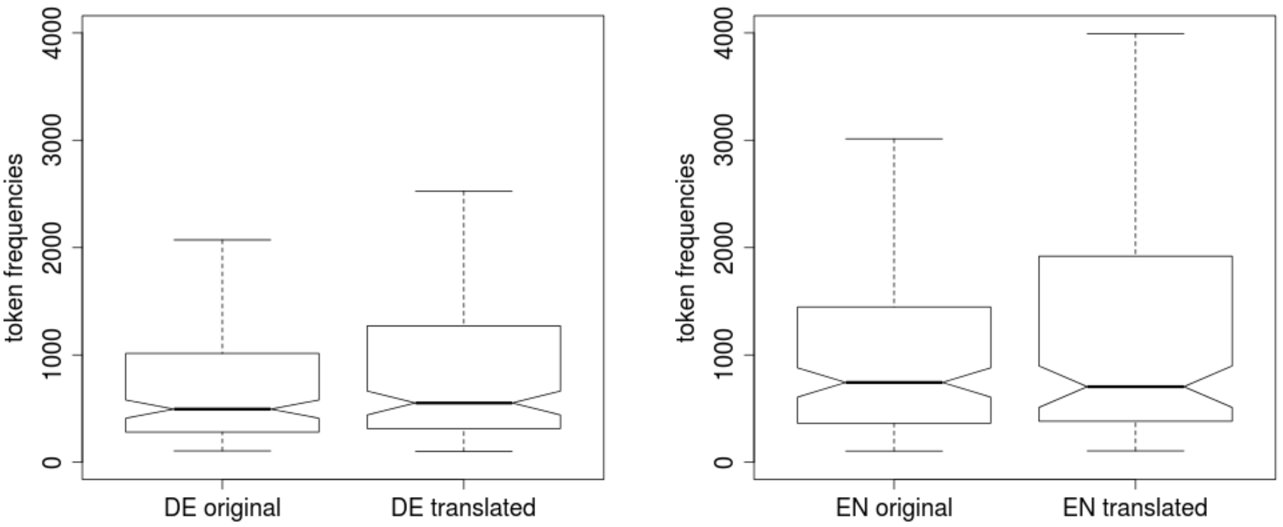

Figure 3: Boxplots of label noun frequencies in original and translated texts (left plot: German, right plot: English), with outliers not displayed.

Corpus, see Dipper et al. (2011) and Zinsmeister et al. (submitted).)

5. Conclusion

In this paper, we compared parallel and comparable subcorpora of Europarl with respect to the distribution of general features as well as features that we used for approximating abstract anaphora. Applying standard corpus linguistic methods of significance testing resulted in highly significant differences in most of the cases. Following Gries (2005), we attributed the significance effects rather to the large sample sizes than to real differences in the subcorpora and employed Cohen's d as a measure of effect size. We conclude from the evaluation of effect sizes that English texts and texts translated from German into English are sufficiently similar to be both used as (training) texts for abstract anaphora investigations even if the translated texts tend to be longer, which we attribute to explicitation. For German texts and texts translated from English into German, the conclusions are not as straightforward as with their English counterparts. The effect sizes of the differences are larger than with the English texts. The quality of the differences hints to shining through effects in addition to explicitation. Shining through of the original language would

DE label noun	Freq$_{orig}$		Freq$_{trans}$		Increase
†Aussprache ‘debate’	255	(226)	1803	(2106)	7.1
†Ansicht ‘view’	642	(569)	2782	(3250)	4.3
†Hinsicht ‘aspect’	168	(149)	666	(778)	4.0
†Angelegenheit ‘matter’	717	(635)	2096	(2449)	2.9
†Bedenken ‘worry’	281	(249)	807	(943)	2.9
					Decrease
†Konsequenz ‘consequence’	624	(553)	224	(262)	2.8
†Auseinandersetzung ‘discussion’	395	(350)	152	(178)	2.6
Voraussetzung ‘prerequisite’	951	(843)	469	(548)	2.0
Zeichen ‘indication’	432	(383)	218	(255)	2.0
Detail ‘detail’	304	(269)	161	(188)	1.9

EN label noun	Freq$_{orig}$		Freq$_{trans}$		Increase
†connection	201	(133)	921	(795)	4.6
†topic	132	(87)	424	(366)	3.2
†task	713	(471)	1871	(1615)	2.6
criticism	413	(273)	837	(722)	2.0
competition	1008	(666)	2021	(1744)	2.0
					Decrease
†reply	684	(452)	269	(232)	2.5
†scheme	979	(647)	389	(336)	2.5
†evidence	958	(633)	409	(353)	2.3
recommendation	863	(570)	397	(343)	2.2
advice	427	(282)	199	(172)	2.1

Table 5: German and English top-5 label nouns that show most extreme overuse (top tables) and underuse (bottom tables). Freq$_{orig}$ and Freq$_{trans}$ are normalized (raw) frequencies in the original and translated texts. Increase and Decrease indicate the normalized factor of overuse and underuse. Nouns marked by † are considered outliers due to extreme overuse/underuse.

corrupt the “naturalness” of the data. Since the effect sizes are only of medium size, we would not completely refrain from using the translated texts, but further investigations are necessary.

In addition to the general considerations, we also investigated the distribution of label nouns as an approximation of abstract anaphora. The subcorpora show similar overall distributions of label nouns but differ with respect to their lexical choices. It seems that translators employ a more restricted vocabulary than the original speakers. It has to be evaluated independently whether tools that are sensitive to lexical choice differ in performance when trained on either the original or translated subcorpora.

6. References

Lars Ahrenberg, Jörg Tiedemann, and Martin Volk, editors. 2010. *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora*, volume 10 of *NEALT proceedings series*, Tartu, Estonia.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Boston MA.

Mona Baker. 1993. Corpus linguistics and translation studies: implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam/Philadelphia.

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

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3).

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

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2(4):597–620. Special Issue.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the ACL-02 conference*, pages 80–87.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora, at ACL-HLT 2011*, pages 78–86.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for NLP*.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 2nd edition.

Oliver Čulo, Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2008. Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. In *Proceedings of LREC 2008*.

Stefanie Dipper and Heike Zinsmeister. 2010. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC-Workshop Language Resource and Language Technology Standards — state of the art, emerging needs, and future developments*.

Stefanie Dipper and Heike Zinsmeister. to appear. Annotating abstract anaphora. *Language Resources and Evaluation*, Online First Sep 2011.

Stefanie Dipper, Christine Rieger, Melanie Seiss, and Heike Zinsmeister. 2011. Abstract anaphors in German and English. In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011. Revised selected papers*, pages 96–107. Springer.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Gill Francis. 1994. Labelling discourse: an aspect of nominal group lexical cohesion. In Malcolm Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. London: Routledge.

Kari Fraurud. 1992. Situation reference: What does 'it' refer to? GAP Working Paper No 24, Fachbereich Informatik, Universität Hamburg.

Stephan Th. Gries. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory*, 1:277–294.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics COLING 08*, pages 937–944.

Andreas Kaster, Stefan Siersdorfer, and Gerhard Weikum. 2005. Combining text and linguistic document representations for authorship attribution. In *SIGIR Workshop: Stylistic Analysis of Text for Information Access (STYLE)*.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Iørn Korzen and Morten Gylling. 2011. What can contrastive linguistics tell us about translating discourse structure? In *Book of Abstracts of the GSCL Workshop on Contrastive Linguistics, Translation Studies, Machine Translation — what can we learn from each other?*

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-08*.

Costanza Navarretta. 2008. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 63–71.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the IEEE-ICSC*.

Marta Recasens. 2008. Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 73–82.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision tree. In *Proceedings of International Conference on New Methods in Language Processing*.

Kiril Simov, Petya Osenova, Radovan Garabik, and Jörg Tiedemann, editors. 2011. *Proceedings of The Second-Workshop on Annotation and Exploitation of Parallel Corpora*, Hissar, Bulgaria.

Elke Teich. 2003. *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. 2002. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of DAARC-2002*.

Jean-Paul Vinay and Jean Darbelnet. 1995. *Comparative stylistics of French and English: A methodology for translation*. John Benjamins, Amsterdam/Philadelphia.

Ralph Weischedel et al. 2010. OntoNotes Release 4.0, with OntoNotes DB Tool v. 0.999 beta. Technical report, Raytheon BBN Technologies et al.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. submitted. Abstract pronominal anaphors and label nouns in German and English: selected case studies and quantitative investigations.