Chapter 7

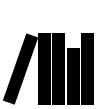
Abstract pronominal anaphors and label nouns in German and English: Selected case studies and quantitative investigations

Heike Zinsmeister
University of Stuttgart

Stefanie Dipper
Ruhr-University Bochum

Melanie Seiss
University of Konstanz

Abstract anaphors refer to abstract referents, such as facts or events. This paper presents a corpus-based comparative study of German and English abstract anaphors. Parallel bi-directional texts from the Europarl Corpus were annotated with functional and morpho-syntactic information, focusing on the pronouns 'it', 'this', and 'that', as well as demonstrative noun phrases headed by "label nouns", such as 'this event', 'that issue', etc., and their German counterparts. We induce information about the cross-linguistic realization of abstract anaphors from the parallel texts. The contrastive findings are then controlled for translation-specific characteristics by examination of the differences between the original text and the translated text in each of the languages. In selected case studies, we investigate in detail "translation mismatches", including changes in grammatical category (from pronouns to full noun phrases, and vice versa), grammatical function, or clausal position, addition or omission of modifying adjectives, changes in the lexical realization of head nouns, and transpositions of the demonstrative determiner. In some of these cases, the specificity of the abstract noun phrase is altered by the translation process.



Heike Zinsmeister, Stefanie Dipper & Melanie Seiss. Abstract pronominal anaphors and label nouns in German and English: Selected case studies and quantitative investigations. In Oliver Czulo & Silvia Hansen-Schirra (eds.), *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II*, 141–181. Berlin: Language Science Press. DOI:10.5281/zenodo.1019699

1 Introduction

Abstract *anaphora* denote an anaphoric relation between an anaphoric expression (i.e., the abstract *anaphor*) and an *antecedent* that refers to an abstract object, such as an event or a fact (Asher 1993). In the well-known example given by K. Byron (2002), the pronoun *it* (underlined in (1a)) refers to an *event*: namely, the migration of penguins to Fiji. In the alternative sequence (1b), the demonstrative pronoun *that* refers to the *fact* that penguins migrate to Fiji in the fall. In both examples, the antecedent is expressed by a clause in the preceding sentence.

(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.

Our method consists of a contrastive, corpus-based approach to investigate the properties that characterize different instantiations of abstract anaphora in English and in German. In the future, we plan to derive features from the corpus annotation that will facilitate automatic resolution of abstract anaphora.

In this paper, we focus on the realization of the anaphoric element, i.e., the *anaphor*. We restrict our investigation to a well-defined set of pronouns and lexical NPs (e.g., *this issue, this directive*, etc.).

We present the results of a comparative corpus study on the realization of abstract anaphors in a parallel bi-directional corpus of English and German. In addition to comparing the cross-linguistic realizations, we also examine these differences between original text and translated text in each of the languages. For a more detailed study on the latter differences, see Dipper et al. (2012).

In previous studies, we focused on the use of pronouns as abstract anaphors (Dipper et al. 2011; Dipper & Zinsmeister 2009). In this paper, we take into account both pronouns and a selection of full NPs. The NPs under consideration here contain a demonstrative determiner, because demonstrative NPs are likely to be used anaphorically. In addition, the NP's head must be an abstract noun such as *issue, effect*, or *process*. We contrast quantitative results from our previous studies with results from our more recent annotations of full NPs.

Furthermore, we investigate selected samples of "translation mismatches" in detail. These mismatches can include anaphors that are not translated word-for-word, but that involve *edit operations*, i.e., addition, deletion, or substitution of words. However, some such mismatches also concern *specificity*, i.e., translation mismatches that affect the amount of information available to the hearer for

the resolution of the reference of the abstract anaphor – for example, when an anaphor is not translated by the most obvious translation candidate, but instead by a target word that is more or less specific than its source word.

The annotated corpus thus far only permits tentative conclusions. We consider the research reported here to be a pilot study that highlights aspects that appear worthy of investigation on a large scale in the future.

The paper is organized as follows: §2 addresses related research; §3 introduces the corpus and the annotations upon which the study is based. In §4, we present quantitative investigations concerning selected properties of the abstract anaphors, such as grammatical category, grammatical function, and position. §5 introduces a range of case studies that address translation mismatches.

2 Related work

The majority of projects that analyze abstract anaphora deal with monolingual data. This section begins with a short, general overview of relevant projects, and then addresses in more detail projects that have examined multilingual corpora.

General studies Most annotation projects that analyze abstract anaphora are limited to pronominal markables (e.g., Byron 2003; Hedberg et al. 2007; Müller 2007). Some also annotate full NP markables, often restricted to demonstrative or possessive NPs (e.g., Vieira et al. 2002; Pradhan et al. 2007; Poesio & Artstein 2008). In projects that have analyzed pro-drop languages, zero anaphora have also been considered (e.g., Recasens 2008; Navarretta & Olsen 2008). A recent overview of projects concerned with the annotation of abstract anaphora is provided by Dipper & Zinsmeister (2010).

Multilingual studies Multilingual corpora have been annotated in Recasens (2008); Navarretta & Olsen (2008); Navarretta (2008); Pradhan et al. (2007); Weischedel et al. (2010). In contrast to the present work, these projects utilize "comparable" rather than parallel corpora (see §3).

Recasens (2008) compares the use of pronominal and NP abstract anaphors in Catalan and Spanish, determining that Spanish prefers personal over demonstrative pronouns, whereas no such preference is found in Catalan. In both languages, full NPs account for half of the abstract anaphors. The heads of these full NPs largely overlap with the "label nouns" reported by Francis (1994): Francis's list is also used in our study (see §3).

Navarretta (2008) and Navarretta & Olsen (2008) compare pronominal abstract anaphors in Danish and Italian. They find that Italian generally avoids the use of pronouns as abstract referents, preferring to use full NPs instead.

Pradhan et al. (2007) and Weischedel et al. (2010) annotate information at various linguistic levels in English, Chinese, and Arabic; a subset of the English and Chinese data consist of parallel (translated) texts. In addition to annotating nominal coreference, they also mark verbs that are coreferenced with an NP (e.g., *grew* and *the strong growth*).

Parallel studies Annotation of parallel texts has been conducted by Vieira et al. (2002), using a subcorpus from the parallel MLCC corpus.[1] The researchers investigate demonstrative NPs in French and Portuguese, finding similar attributes: In both languages, demonstrative NPs predominantly use abstract head nouns. Vieira et al. (2002) do not distinguish between texts in original and translations.

Characteristics of parallel corpora Parallel corpora, such as MLCC (see above) or Europarl (Koehn 2005), consist of original and translated texts. There has been a long-standing debate over the extent to which translated language deviates from comparable original language due to influences from both the original source language and the translation process; some arguing that such material should therefore not be used as a base for linguistic investigations (other than those focusing on translation issues such as, e.g., Čulo et al. 2008); see the related discussion in §4.

For instance, Cartoni et al. (2011) investigate the use of discourse connectives in original and translated French texts from Europarl, finding that translated texts contain significantly more discourse connectives than original texts. 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 and 96.7%.

3 The corpus

For our study, we used parts of the Europarl Corpus (release v3, 1996–2006, Koehn 2005). The Europarl Corpus consists of transcripts of European Parliament debates. Individual contributions by speakers ('turns') in the debates were

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

delivered (for the most part) in the speaker's native language. Professional translators provided official EU translations into the other EU languages.

The original contributions were spoken, but might have been based on written scripts. Speakers had the option to edit the transcripts before publication. As a result, the register of these turns is of a mixed character, varying between spoken and more standardized written language.

We created subcorpora by extracting German and English turns (contributions by German and English speakers), along with their sentence-aligned translations. This provided us with four different subcorpora; the German original turns (DE_o) and their English translations (EN_t), and the English original turns (EN_o) and their German translations (DE_t).

These four subcorpora stand in different relations to each other (see Figure 1). EN_o and DE_t (and DE_o and EN_t) are *parallel* corpora, i.e., they consist of original texts and their translations. The subcorpora DE_o and EN_o (and similarly, DE_t and EN_t) are comparable corpora, i.e., corpora in different languages that deal with the same overall topic and come from the same overall register. This notion of comparable corpora is often used in corpus-linguistic research; we therefore 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 generally refer to such corpora as comparable, thus we call this type of relation $comparable_{trans}$. We based the investigations presented in this paper on these various relations between the subcorpora.

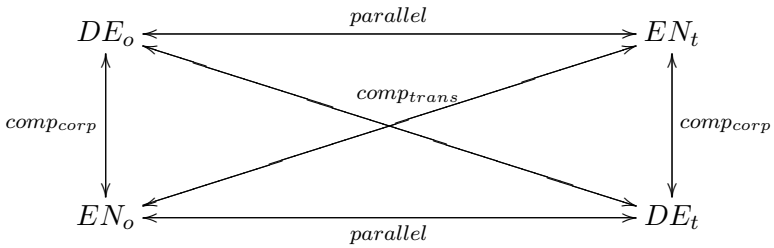


Figure 1: There are three types of relations between the four subcorpora: parallel, comparable in the corpus-linguistic sense ($comp_{corp}$), and comparable in the translation-studies sense ($comp_{trans}$)

Anaphora Corpus We created a small manually annotated corpus, which we call *Anaphora Corpus*. For this, we randomly selected about 100 turns from DE_o

and EN_o, respectively, for our manual annotation study; our goal was to investigate the properties of abstract anaphors, in particular their realization as pronouns or full lexical NPs, but also in terms of function, position, etc. To this end, a number of pre-processing steps were applied. These included verifying the native language of the speakers.[2] After this step, we were left with 94 German original turns and 95 English original turns. Further pre-processing of the data included tokenizing, POS tagging, and chunking by means of the TreeTagger (Schmid 1994). For the manual annotation of the German and English turns, we used MMAX2 (Müller & Strube 2006).

The various processing steps and manual annotations implemented are described in the following sections.

3.1 Annotating pronominal abstract anaphors

We adopted a cross-linguistic bootstrapping approach for the annotation of abstract pronouns. Starting with a well-defined set of markables in the original language, we collected all translation equivalents on the side of the "target" language (the translation of the original language).

In the first round of annotation, we chose original texts from German (DE_o), because German, unlike English, has a pronoun that is unambiguously used as an abstract anaphor: the uninflected singular demonstrative pronoun *dies* 'this'. In addition, we defined as markables the (ambiguous) demonstrative pronoun *das* 'that' and the (ambiguous) third-person neuter pronoun *es* 'it'. For all instances of these pronouns, the annotators first determined whether they were in fact being used as abstract anaphors by specifying their antecedents. In a further annotation step, the annotators had to determine how the German abstract anaphors were translated in the English data (EN_t).

For the second round of annotation, we considered the reverse translation direction: English original texts (EN_o) and their German translations (DE_t). We extended our set of markables to include the adverbs *as, so*, and *likewise*, because it was determined in the first annotation round that these adverbs often served as translations of German anaphors.[3]

In total, 871 instances of neuter pronouns were found in DE_o, and 1,224 instances of pronouns and adverbs (= the extended set) in EN_o. Of these, 203 (DE_o) and 297 (EN_o) were determined to be abstract anaphors.

[2] The language markers provided in release v3 turned out to be incomplete and partially incorrect. We therefore looked up each speaker's origin in a database of EU members of parliament.

[3] Because we used different sets of markables in the different annotation rounds, the figures from different rounds cannot be easily compared, see below.

For further details of the annotation process and the annotated features, see Dipper et al. (2011).

3.2 Annotating abstract NPs

In addition to pronominal abstract anaphors, we also annotated abstract full NPs. To accelerate the annotation process, we carefully preselected a set of NPs that seemed likely candidates for abstract anaphors by applying two constraints: First, only NPs with a demonstrative determiner were selected, because such NPs are generally used anaphorically. Second, we defined a list of admissible head nouns that refer to abstract entities.

For English, abstract nouns (such as *report, arrangement*, and *fact*) were selected. The list of nouns, which was heavily influenced by the *label nouns* defined by Francis (1994), comprised 211 abstract nouns. Table 1 provides some examples. In total, 132 instances of these nouns (in singular and plural form) occurred in EN_o of the Anaphora Corpus.[4]

We chose the most common translations for the English label nouns to create a list of German label nouns[5] and excluded non-abstract translations. This resulted in between one and ten German translations per English noun, with an average of 3.6 translations per English noun. Some example translations are provided in Table 1. The large number of German label nouns can be explained by the fact that we started out with a predefined set of English label nouns, and that these nouns are quite general in meaning; thus, depending on the context, they can be translated with a variety of German abstract nouns.

Table 1: English label nouns and their German translations

English noun	German translations
problem	*Problem* 'problem', *Fragestellung* 'question', *Problemstellung* 'problem'
activity	*Aktivität* 'activity', *Aktion* 'action', *Handlung* 'act'
subject	*Gegenstand* 'object', *Gesprächsgegenstand* 'topic'
topic	*Gegenstand* 'object', *Inhalt* 'content', *Thematik* 'subject matter', *Thema* 'matter', *Themengebiet* 'topic area'

Table 1 also shows that our method yielded multiple English translations for German label nouns as well. For example, *Gegenstand* 'object' can be translated

[4] EN_o: 132 instances of 45 different label noun types.
[5] Translations based on LEO, http://www.leo.org/.

as *subject* or *topic*. The final list consisted of 452 types of German label nouns. Of these, 134 (inflected) instances occurred in the German Anaphora Corpus DE_o.[6] Of course, not all of these were true instances of abstract anaphors (see below).

In a pre-processing step, the data was split into individual original *alignment units* as provided by the Europarl Corpus, each followed by its translation. In the units of the original text, all noun chunks with a label-noun head were pre-marked as *markables* (English label nouns in EN_o, and German label nouns in DE_o). In the translated units, noun chunks were generally pre-marked as potential translation equivalents.

In the annotation procedure, the annotators were first asked to check whether the label noun occurrences were in fact abstract. This was important because some label nouns can be ambiguous between an abstract and a non-abstract interpretation. For example, *area* can also refer to an actual geographic area, and *report* can refer to a copy of a report. This procedure resulted in 130 English and 117 German abstract NPs for further manual annotation.[7]

Annotators were next asked to align the original noun chunk with its translation. After this step, both the original label noun and the corresponding material in the translation were annotated for category, function, and position.[8] Figure 2 shows screenshots of the MMAX2 annotation windows.

In sum, for the analysis of both pronominal and NP anaphors, the same data and similar strategies were used. In both cases, we started out with a well-defined set of markables, although the set of markables for pronominals was naturally considerably smaller than the set of label nouns. In both cases, we considered how the markables had been translated and whether we could induce new markables for the next annotation round. We believe that this kind of bootstrapping approach provides a faster and more efficient method of extracting anaphors in two languages in comparison to processing contiguous text without predefined markables. Working without predefined markables would also present the risk

[6] DE_o: 134 instances of 51 different label noun types.

[7] This demonstrates that our pre-selection was highly successful in the case of abstract NPs. In contrast, occurrences of the pronominal anaphors *this, that, it*, and *das* 'that' and *es* 'it' in German most often refer to concrete referents.
Annotators did not need to determine the antecedents in the case of abstract NPs, because we could assume that most of the label nouns were abstract *per se*. In ambiguous cases, annotators did a quick check of the previous context to determine whether the noun was abstract.

[8] Admissible values were:
– Category: 'noun phrase', 'pronoun', 'pronominal adverb', 'genauso/likewise', 'sentence', 'other'
– Function: 'subject', 'object', 'object of a preposition', 'noun phrase attribute', 'other'
– Position: 'topic/prefield', 'matrix', 'embedded', 'other'.

that annotators would disagree on the set of types under consideration or, more likely still, on the markables themselves.

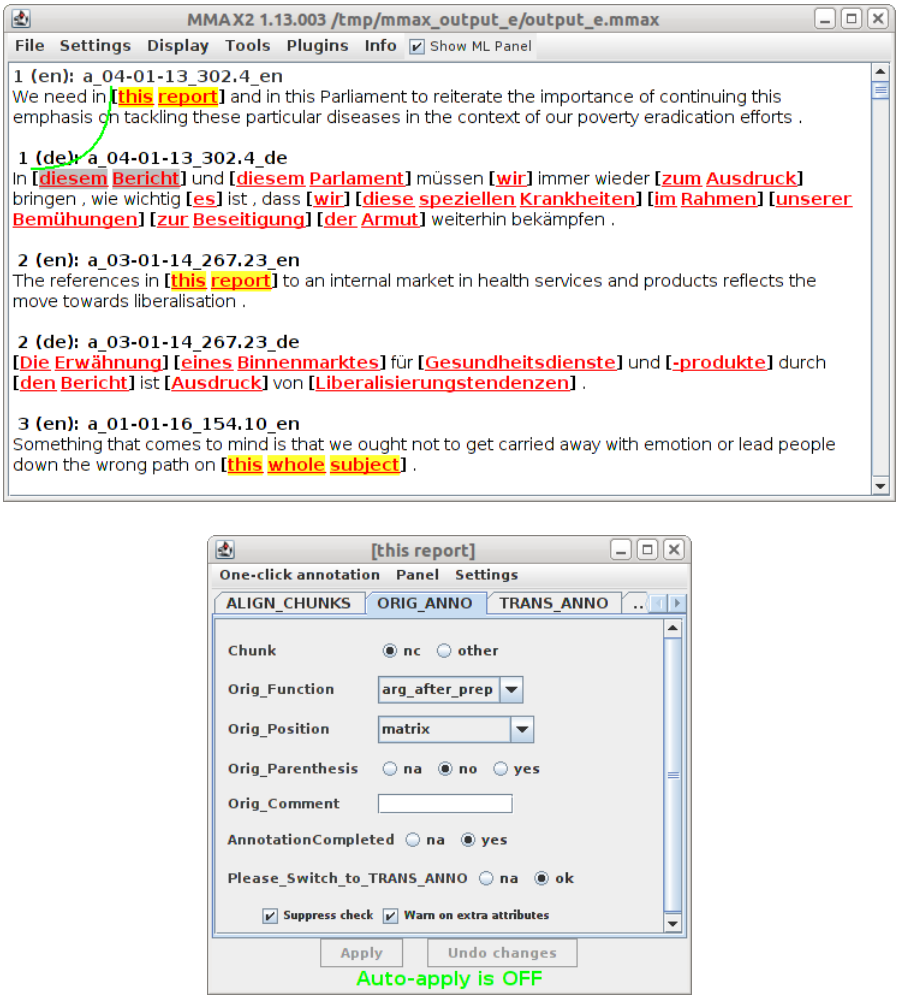


Figure 2: MMAX2 annotation windows: The upper panel shows English alignment units, along with their German translations. Noun chunks with label nouns to be processed by the annotators are highlighted in yellow. Translation candidates are marked in red. In the first alignment unit, the anaphoric abstract noun chunk 'this report' has been aligned with its German equivalent 'diesem Bericht'. The lower panel displays features that have been annotated to the English noun chunk. Similar features have also been annotated to the translated noun chunk (not displayed in the figure).

4 Quantitative investigations

This section presents our quantitative results from investigation of the Anaphora Corpus. For selected cases, findings based on our manually annotated data are complemented by evaluations of data from the entire German and English Europarl Corpus.

An obvious advantage of using parallel texts for cross-linguistic research is that the aligned units convey the same meaning and allow us a direct comparison of how this meaning is expressed linguistically in the two languages. This cross-linguistic use of parallel texts also has limitations, as many studies in translation studies have shown. The most troublesome for our research purposes are:

(i) The problem of *translation shifts* (cf. Vinay & Darbelnet 1958/1995; Dorr 1994); this refers to the fact that translated texts systematically differ from their source texts due to language-inherent differences. Further factors that can result in language-specific differences in translations are stylistic preferences (e.g. language-specific conventions that apply to parliamentary debate protocol and its translation) and cultural differences, for which the background knowledge of the hearers plays a role (Klaudy 2008).

(ii) Effects inherent to the translation process, which can affect the characteristics of translated texts in various ways. There are two subtypes that are particularly relevant for us: the *shining-through* of source-language preferences when a translation is too faithful to its source text (cf. Teich 2003), and the tendency of translated texts to be more *explicit* than their sources (Vinay & Darbelnet 1958/1995; Blum-Kulka 1986).[9] Both of these characteristics might directly affect how anaphoric links are expressed,

[9] Vinay & Darbelnet (1958/1995: 342) were the first to define the concept of *explicitation*, "a stylistic translation technique which consists of making explicit in the target language what remains implicit in the source language because it is apparent from either the context or the situation".
Blum-Kulka (1986) formulated the explicitation hypothesis: "The process of interpretation performed by the translator on the source text might lead to a TL [target language] text which is more redundant than the SL [source language] text. This redundancy can be expressed by a rise in the level of cohesive explicitness in the TL text. This argument may be stated as 'the explicitation hypothesis', which postulates an observed cohesive explicitness from SL to TL texts regardless of the increase traceable to differences between the two linguistic and textual systems involved. It follows that explicitation is viewed here as inherent in the process of translation" (Blum-Kulka (1986: 19); both citations from Klaudy 2008).
For a recent survey and critical assessment of the explicitation hypothesis, see Becher (2011: Ch. 2).

such that translated texts could end up quite different from comparable original texts.

We expect the aspects listed in (i) to result in differences between languages (*parallel* and *comparable*$_{corp}$ corpora, cf. Figure 1), and those effects in (ii) to result in differences between original and translated texts (*comparable*$_{trans}$ corpora). These differences – even if only in form and not in meaning – pose problems for approaches that target the automatic resolution of anaphora.

Having outlined the specific characteristics of translated texts, we then pursued a two-step approach. First, we compared the expression of abstract anaphors in the aligned units of the *parallel* resources. Second, we checked our results — when possible — with the *comparable*$_{trans}$ part of the corpus. This process required a number of steps, explained below in greater detail.

Step 1: We first examined parallel (translated) texts. A naïve assumption would be that in aligned units of parallel texts, abstract anaphors would be realized in the same way in both languages (e.g., with the same category and function). When we found differences between the parallel texts (e.g., a transposition,[10] as described in (a)), there were two possible explanations: either the differences were due to (i) language-specific preferences, or to (ii) effects of the translation process.

(a) *Observation of transposition*: German pronouns tend to be translated by English NPs.

To determine which explanation was applicable, we pursued various methods.

Step 2: We next checked whether the tendencies also appeared in the reverse translation direction (b).

(b) *Reverse translation direction of (a)*: English pronouns would tend to be translated by German NPs.

If (b) were true, observation (a) would likely represent an effect of the translation process. If the tendencies only showed up in one translation direction, it would indicate a language-specific effect.

Moreover, we could check whether the tendency was also observed in the reverse direction of the *transposition* (c).

[10] We use the term *transposition* to refer to changes in the grammatical category, function, etc., that occur as the result of translation.

(c) *Reverse transposition of (a)*: German NPs would tend to be translated by English pronouns.

If this were the case, the transpositions in question would seem to occur at random, and no general "rule" could be deduced from the observations.

Step 3: In addition, we checked the ratios in a *comparable*$_{trans}$ corpus (e.g., by comparing the numbers of pronouns and NPs in DE_o and DE_t, and in EN_o and EN_t). If we observed differences between original and translated texts for both German and English, this would indicate an effect of the translation process. If these differences were observed in one language only, it would indicate a language-specific effect.

We applied Steps 1 to 3 in order to shed light on the *linguistic similarity* of abstract anaphors in German and English, and in original texts and translated texts.

The following sections present quantitative results for abstract anaphors with regard to lexical choice (§4.1), grammatical category (§4.2), grammatical function (§4.3), and position in the clause (§4.4). For each of these properties, we examined pronominal anaphors (cf. §3.1) and label noun NP anaphors (cf. §3.2) annotated in the Anaphora Corpus. More detailed, qualitative discussions of translation equivalences are provided in §5.

4.1 Lexical choice

Pronominal abstract anaphors We first focused on the different lexical realizations of abstract anaphors in the original and translated texts, and compared their frequencies.

Table 2 provides a comparison of the frequency rankings in the comparable$_{trans}$ corpora (DE_o–to–DE_t, and EN_o–to–EN_t; the table is organized in accordance with the corpus scheme from Figure 1).

The table illustrates that the lexical choices lead to distributions in the translated corpora that correspond to those in their comparable$_{trans}$ counterparts: The top-ranked pronouns are equivalent in both comparable$_{trans}$ pairs. For the German corpora, *das, dies, es* are top-ranked, with *wie* 'as' intervening in DE_t; as this word was not part of the original markable set, its frequency cannot be compared. For the English corpora, *this, that, it, as* are top-ranked. The re-ranking of *it*, and *as* (in EN_t vs. EN_o) can probably be explained by the fact that *wie* (the German equivalent of *as*) was not included in the first annotation round, as just noted. A remarkable deviation is the relative overuse of *dies* 'this' in DE_t in comparison to

Table 2: Frequency rankings of original pronominal abstract anaphors and translation equivalents

Rank	DE_o pronouns	Freq	Rank	EN_t most frequent equivalents	Freq
1.	*das* 'that'	123	1.	*this*	55
2.	*dies* 'this'	45	2.	*that*	52
3.	*es* 'it'	35	3.	*it*	22
			4.	*as*	9
			5.	*which*	5
			6.	*they, these things, likewise, what, to do so, this threat* ...	< 5

Rank	EN_o pronouns	Freq	Rank	DE_t most frequent equivalents	Freq
1.	*this*	108	1.	*das* 'that'	71
2.	*that*	103	2.	*dies* 'this'	48
3.	*as*	42	3.	*wie* 'as'	31
4.	*it*	36	4.	*es* 'it'	13
5.	*so*	8	5.	*deshalb* 'therefore'	8
			6.	*damit* 'with that'	6
			7.	*was* 'what', *so* 'so', *hier* 'here', *davon* 'thereof', *dieser Prozess* 'this process', ...	< 5

DE_o if we only take into account occurrences of *das, dies*, and *es*.[11] This might be an example of *shining-through* of the frequently occurring English *this* in EN_o.

Table 3 provides a detailed view of the anaphors by aligning them with their actual translations. For each pronominal abstract anaphor, its absolute frequency in the original data and the number of different equivalence types is given. In addition, the most frequent equivalence types are listed, together with their absolute frequencies in the translated text.

[11] Chi-squared test: $\chi^2 = 7.3459, df = 1, p < 0.01$ based on R's *prop.test(c(45,48),c(203,132))*.

Table 3: Pronominal markables and their most frequent translation equivalents. The pronominal frequencies include cases in which the pronoun could not be aligned to corresponding material in the translation.

DE original			EN translations	
Pronoun	Freq	Types	Top equivalents	Freq
das 'that'	123	25	*that*	44
			this	27
			it	12
			which	5
			as	3
dies 'this'	45	9	*this*	23
			that	4
			as	3
			it	3
es 'it'	35	8	*it*	8
			this	5
			that	4
			as	3
EN original			DE translations	
Pronoun	Freq	Types	Top equivalents	Freq
this	108	42	*dies* 'this'	32
			das 'that'	21
			damit 'so that'	4
			hier 'here'	4
that	103	39	*das* 'that'	43
			dies 'this'	9
			deshalb 'therefore'	8
as	42	11	*wie* 'as'	31
it	36	16	*es* 'it'	9
			das 'that'	7
so	8	4	*dies* 'this'	4

Comparison of the anaphors with their translation equivalences in Table 3 demonstrates that in almost all cases, the literal translation is observed most frequently. *Das* 'that' is most often translated as *that*, *that* as *das*, and so forth. The only exception is the English *so*, which most often translates into *dies* 'this' — the German pronoun that unambiguously refers to abstract objects.[12]

Abstract anaphors with demonstrative label nouns An overview of the most frequent label nouns occurring in the Anaphora Corpus is provided in Table 4.

The ten most frequent types listed in Table 4 account for 59% of all instances in the original corpora, and for the considerably smaller proportion of 46% in the translated corpora.[13] This could be an effect of style in the translations, as translators might tend to show more diversity than the original authors. However, this conclusion does not hold when evaluating larger parts of the Europarl Corpus as discussed on page 171.

Examining individual translation pairs confirms the same tendency of literal translation preference as was observed with the pronominal anaphors. Most of the nouns are translated by only one or two different translation equivalences. Exceptions with greater translational variance include *agreement* (five equivalent types: *Abkommen*, *Einigung*, *Vereinbarung*, *Übereinkommen*, *Übereinstimmung*), *issue* (four types: *Angelegenheit*, *Erweiterung*, *Problem*, *Thema*), *Thema* (four types: *area*, *issue*, *subject*, *topic*), and *Frage/Fragen* (four types: *area* , *issue*, *situation*, *questions*).

Comparing the rankings in Table 4, the *parallel* rankings (horizontal neighbors, e.g., DE_o and EN_t) are more similar to each other than to the $comparable_{trans}$ rankings (diagonal neighbors, e.g., DE_o and DE_t).[14] It seems that in the case of label noun anaphors, the topic of the individual texts has a greater effect on the choice of the lexical items than language-specific conventions. This is in correspondence with findings reported in the literature.

[12] The preferences of the literal translations are significant according to a Chi-squared test for *das* ($\chi^2 = 5.0685, df = 1, p < 0.05$), *dies* ($\chi^2 = 17.1429, df = 1, p < 0.001$), *that* ($\chi^2 = 28.0137, df = 1, p < 0.001$), and *as* ($\chi^2 = 39.1301, df = 1, p < 0.001$). There is no significant difference for the translation of *this* as either *dies* or *das*. The other anaphors' frequencies are too low to be conclusive.

[13] The proportion of instances associated with the top-ten most frequent types, broken down by language, are: DE_o: 56%, EN_t: 44%, EN_o: 62%, DE_t: 48%.

[14] Some of the differences are artificial, related to the selection of label nouns that were premarked as markables. *Directive*, for example, was not in the list of English label nouns and is therefore missing from EN_o. See the discussion of the nouns *Bereich* 'area' and *directive* below.

Table 4: Frequency rankings for the most common label nouns

Rank	DE_o label nouns	Freq	Rank	EN_t label nouns	Freq
1.	*Bericht* 'report'	13	1.	*report*	13
2.	*Richtlinie* 'directive'	12	2.	*directive*	10
3.	*Thema* 'issue'	10	3.	*issue*	7
4.	*Prozess* 'process'	6	4.	*process*	5
5.	*Frage* 'question/ issue'	5	5.	*debate*	4
	Punkt 'point'	5	6.	*area*	3
7.	*Debatte* 'debate'	4		*questions*	3
	Fragen 'questions/ issues'	4		*subject*	3
	Zusammenhang 'context'	4	9.	*basis*	2
10.	*Ergebnis* 'result'	3		*connection*	2
Rank	EN_o label nouns	Freq	Rank	DE_t label nouns	Freq
1.	*report*	19	1.	*Bericht* 'report'	15
2.	*proposal*	10	2.	*Thema* 'issue'	8
3.	*area*	9	3.	*Vorschlag* 'proposal'	7
4.	*agreement*	8	4.	*Bereich* 'area'	6
5.	*issue*	7	5.	*Fall* 'case'	5
	point	7		*Punkt* 'point'	5
7.	*context*	5	7.	*Angelegenheit* 'issue'	4
	subject	5		*Berichts* 'report' (genitive)	4
9.	*debate*	4		*Gebiet* 'area'	4
	problem	4		*Problem* 'problem'	4

Usage preferences for selected nouns In addition to using the comparable corpora that form part of the Anaphora Corpus, we also took advantage of the huge amount of comparable data provided by the Europarl Corpus: 12,800 German original turns with 4.9 M tokens, and 11,500 English original turns with 3.4 M tokens. In this section, we illustrate how this data can be used to detect interesting cases that seem worthy of closer examination. Note that in this subsection, the

abbreviations DE_o, DE_t, etc., are also used to refer to the respective subcorpora of the Europarl Corpus. In most other sections in this paper, these abbreviations refer exclusively to the Anaphora Corpus.

Our starting point was the considerable divergence we found in the frequencies of certain label nouns in comparisons of original and translated turns in our Anaphora Corpus. We selected all label nouns with "considerable" differences (greater or equal to four) between the frequencies of original and translated turns, see Table 5. The columns labeled 'Anaphora Corpus' list the respective figures. A

Table 5: Label nouns with difference greater of equal four between the frequency of original and translated turns. '#' indicates absolute frequencies (as occurring in the annotated corpora): 'Diff' represents the difference between the two frequencies. 'Freq' refers to frequencies relative to the total number of nouns, multiplied by 1,000 (calculated on the basis of all Europarl turns). DE_o/DE_t etc., is the proportion of the label noun's frequency in the original turns compared to its frequency in translated turns. The entries are sorted according to the differences in frequency in the Anaphora Corpus; notable figures are printed in boldface. (For nouns marked with '*', see the remarks in the text.)

Label noun	Anaphora Corpus		Europarl Corpus			
	$\#DE_o$:$\#DE_t$	Diff	Freq DE_o	Freq DE_t	DE_o/DE_t	DE_t/DE_o
*Richtlinie** 'directive'	12 : 0	12	2.656	3.282	0.809	1.236
Vorschlag 'proposal'	1 : 7	−6	3.272	3.835	0.853	1.172
*Bereich** 'area'	0 : 6	−6	4.020	2.714	1.481	0.675
Frage 'question/issue'	5 : 0	5	6.695	5.440	1.231	0.813
Fall 'case'	0 : 5	−5	2.260	2.362	0.957	1.045
Prozess 'process'	6 : 2	4	0.482	0.776	0.621	1.611
Debatte 'debate'	4 : 0	4	2.355	1.523	1.546	0.647
Fragen 'questions/issues'	4 : 0	4	2.349	2.820	0.833	1.200
Angelegenheit 'issue'	0 : 4	−4	0.287	1.375	0.209	4.797

Label noun	Anaphora Corpus		Europarl Corpus			
	$\#EN_o$:$\#EN_t$	Diff	Freq EN_o	Freq EN_t	EN_o/EN_t	EN_t/EN_o
*directive**	0 : 10	−10	4.900	4.579	1.070	0.934
proposal	10 : 1	9	5.436	5.690	0.955	1.047
agreement	8 : 1	7	4.868	4.116	1.183	0.845
*area**	9 : 3	6	3.480	4.361	0.798	1.253
point	7 : 1	6	5.885	6.668	0.883	1.133
report	19 : 13	6	18.881	13.438	1.405	0.712
context	5 : 0	5	1.292	1.506	0.858	1.165

negative number in the 'Diff' column indicates that the label noun occurs more often in the translated turns. For example, Table 5 shows that the noun *Angelegenheit* 'issue' (ranked last in the top table) never occurs in a German original turn, but occurs four times in translations from English turns (i.e., a difference of four occurrences). In contrast, the noun *report* (see the lower table) occurs considerably more often in original English turns (19 times) than in translated turns (13 times).

Similarly, the nouns *Bereich* 'area' and *directive* (marked with '*' in the table) were only annotated in translated turns. However, this is because *Bereich* and *directive* were not included in our original set of label nouns, and thus their occurrences were not pre-marked and annotated in the MMAX2 files, although they appear quite frequently as translation equivalents in the annotated translations. In the next round of annotations, they will be included in our set of label nouns, in accordance with our general bootstrapping approach. The fact that the EN_o noun *directive* was not included in the first annotation round also had an impact on the frequency of its DE_t translation *Richtlinie* 'directive' (ranked first), which was never found in German translations for this reason. The same holds true for the frequency of the EN_t noun *area*: Its literal DE_o counterpart *Bereich* was not annotated in the original texts.

For each of the label nouns with considerable differences, we calculated its frequency in *all* original and translated turns of the Europarl Corpus (release v3).[15] We found that these frequencies differed significantly for all nouns, except for *Fall* 'case' in German and *directive* and *proposal* in English.[16]

In general, certain label nouns seem to be overused in translated texts in comparison to original texts. This can be seen in the last four columns in the tables, which list the relative frequencies of the label nouns in original and translated turns (multiplied by one thousend) and the ratio of these frequencies. For instance, the first noun is *Richtlinie* 'directive', which occurs with a relative frequency of 2.656 in original turns and of 3.282 in translated turns. This indicates that the noun occurs more often in translated turns. This is reflected by the fact that the proportion DE_o/DE_t is less than one and, consequently, the proportion DE_t/DE_o is greater than one. The last two columns show that in six instances (out of nine) in the German data, the proportion DE_t/DE_o is greater than one, and that in four times (out of seven) in the English data the proportion $EN_t/EN_o$is greater

[15]Only translations from original turns in German and English were considered.

[16]Chi-squared test with continuity correction, using the label noun vs. the class of all other nouns as features. With the noun *context*: $\chi^2 = 8.39, df = 1, p < 0.01$; all remaining nouns: $\chi^2 > 25, df = 1, p < .001$. Significant effects are easily achieved in large corpora. In Dipper et al. (2012), we discuss the results on the basis of their effect size (as suggested by Gries 2005).

than one as well. We tentatively conclude from this that the translations possibly have a more restricted vocabulary than the comparable original texts, and that individual common types thus occur with a higher relative frequency in the translated texts than in the originals.

A strikingly large frequency difference can be observed for the German noun *Angelegenheit* 'issue', which occurs 4.8 times more often in the translated turns of the Europarl Corpus; the second-ranking noun in translations is *Prozess* 'process', which occurs 1.6 times more often. Conversely, the nouns *Debatte* 'debate' and *Bereich* 'area' top the list of nouns that occur more often in the original turns — approximately 1.5 times more often. The differences in the English data are less pronounced. The top-ranked noun is *report*, which occurs 1.4 times more often in the original data.

The top-ranked nouns, i.e., those that demonstrated considerable frequency divergence both in the Anaphora Corpus and in the Europarl Corpus (indicated by figures printed in boldface in Table 5), were subject to further investigation.

***Angelegenheit* 'issue'**: The striking frequency differences that occur with *Angelegenheit* 'issue' might be attributable to the fact that the word seems to be used as a kind of "dummy" translation for English nouns that are highly unspecific, such as *issue, matter*, or *matter of concern*. (2) shows such an example.[17]

(2) EN_o: But, on this issue, I do not see any room for soft law which is why in the transition period there will be total adherence to the current financial regulation until that law is changed by due democratic process in this House and in the Council.
DE_t: Aber in dieser Angelegenheit sehe ich keinen Raum für "soft law", weshalb es im Übergangszeitraum eine strikte Befolgung der aktuellen Haushaltsordnung geben wird, bis diese Rechtsvorschrift durch das erforderliche demokratische Verfahren in diesem Hohen Hause und im Rat geändert worden ist. (ep-00-03-01/28)

***Prozess* 'process'**: Interestingly, in the Europarl Corpus, the noun *Prozess* 'process' occurs much more often in translated turns than in original ones—contrary to the ratios observed in the Anaphora Corpus. *Prozess* is always translated by its closest equivalent 'process' in the Anaphora Corpus, and vice versa: *process* is always translated by *Prozess* in this data. Our data do not permit any tentative conclusion that would explain the observed frequency differences.

[17] We mark the examples taken from the Europarl corpus with the name of the file (e.g., ep-00-03-01) and the speaker ID, as provided by release v3 of the Europarl Corpus.

***Debatte* 'debate'**: occurs more often in original German turns (no occurrence in DE_t in the Anaphora Corpus). A highly speculative explanation is that the German *translators* — in contrast to the German *speakers* — prefer the noun *Aussprache* as the translation of *debate*. *Aussprache* can mean 'discussion' but also 'interlocution, talk', whereas *Debatte*, as used in every-day language, means 'dispute, argument'. Used in the sense of 'parliamentary debates', the negative connotation is absent, the meaning being 'discussion, debate'. However, translators could be avoiding the use of the noun *Debatte* due to its negative connotations in other contexts.

***Bereich* 'area'**: As mentioned above, the noun *Bereich* 'area' was not annotated in original German turns in the first annotation round. The six examples that appeared in the translations (see Table 5) are translations of *area* (five times) and *question* (one time). In an extra step, we looked up all occurrences of *Bereich* in DE_o: this resulted in six instances that are translated in six different ways, e.g., by *area, sphere*, etc. (cf. (3)).This means that in the translation direction DE_o–to-EN_t, we observe a vast variety of English expressions that correspond to German *Bereich* 'area', whereas in the reverse direction (EN_o–to–DE_t) *Bereich* is only used as a translation of *area* (and, in one instance, of *question*).

(3) DE_o: Deswegen brauchen wir ein gemeinsames Satellitenaufklärungssystem der Europäischen Union und gemeinsame Standards für die Telekommunikation in diesem Bereich.
EN_t: That is why we in the European Union need a single satellite reconnaissance system and common standards for telecommunications in this sphere. (ep-06-05-17/20)

***Report* 'report'**: Finally, the noun *report* occurs extremely frequently in the Anaphora Corpus, both in EN_o and EN_t (and with similar frequencies in the Europarl Corpus). Some of these occurrences can be explained by the fact that in their turns, speakers often refer to reports that are up for discussion, see (4).

(4) EN_o: Madam President, I would like to thank the rapporteur for producing this report because it is a very important one.
DE_t: Frau Präsidentin, ich möchte dem Berichterstatter für seinen Bericht danken, denn es handelt sich um einen wirklich wichtigen Bericht. (98-11-17/284)

4.2 Grammatical category

Pronominal abstract anaphors In addition to lexical choice, we also investigated the grammatical properties of the anaphors. We evaluated whether pronouns were translated by pronouns — as our initial "naïve assumption" would predict (see Step 1 in §4) — or by another category (e.g., full NP, adverbial, or clause). This investigation was motivated by findings on cross-linguistic differences (e.g., between Danish and Italian and between Spanish and Catalan: cf. Recasens 2008; Navarretta 2008; Navarretta & Olsen 2008).

Assuming equivalence between the original text and the translation, we would expect to find only pronoun–to–pronoun mappings (and adverb–to–adverb, if adverbs had been included in the markable set). Our data does not confirm this equivalence. In the corpus DE_o–to–EN_t, only 65% (132) of the pronominal markables are translated as pronouns, see Table 6, first row.

Table 6: Pronouns: Categorial transposition types

	Pronoun to pronoun		Pronoun to NP		Pronoun to other		Sum	
DE_o–to–EN_t	65.0%	(132)	9.4%	(19)	25.6%	(52)	100%	(203)
EN_o–to–DE_t	70.3%	(173)	7.3%	(18)	22.4%	(55)	100%	(246)

Other target categories of translated pronouns included NPs, cf. (5), and adverbials such as *so, likewise* — which were then added to the English markable set.[18]

(5) EN_o: I do not necessarily support this.
DE_t: Diesem Standpunkt schließe ich mich nicht notwendigerweise an.
DE_{lit}: This position I do not necessarily support. (ep-00-10-03/15)

In examining the EN_o–to–DE_t corpus, we found similar results (Table 6, second row). The proportional distributions between DE_o–to–EN_t and EN_o–to–DE_t do not differ significantly.[19]

The bar plots in Figure 3 provide a more general overview by summarizing the relative frequencies of grammatical categories in the Anaphora Corpus. The

[18] DE_{lit} provides a literal translation of the German sentence.
[19] Chi-squared test: $\chi^2 = 1.5185, df = 2, p = .468$

top chart displays the data for pronominal anaphors in the source languages. For example, EN_o starts out with a larger set of markables than DE_o due to its inclusion of non-pronominal, adverbial types.

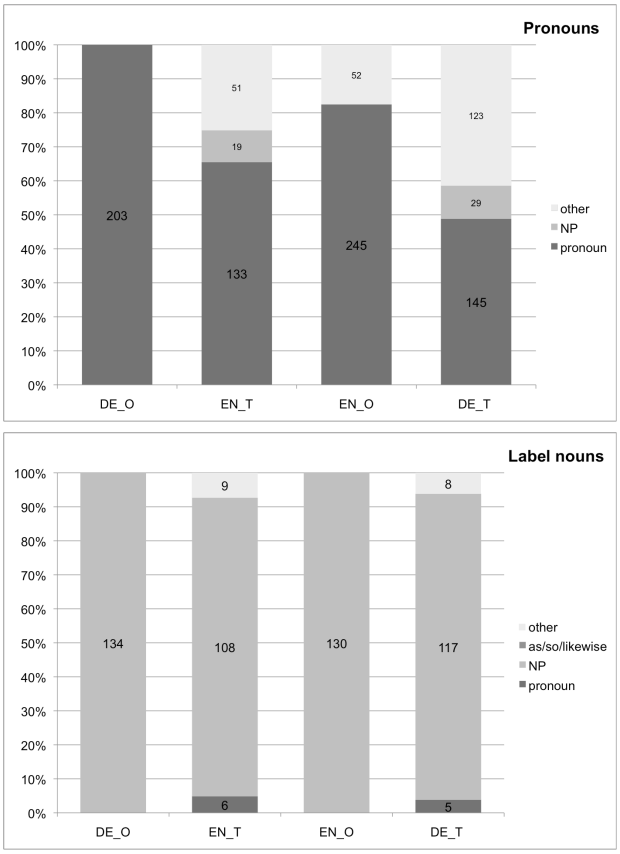


Figure 3: Relative frequencies of grammatical categories. Top chart: figures of pronominal anaphors; bottom chart: figures of the label nouns. Class 'as/so/likewise' is the markable type introduced in EN_t. Class 'other' (the white parts) consists of other cases with structural mismatches in the translations (such as translations by clauses), or cases in which anaphors could not be aligned to corresponding material in the translation.

It is clear that German and English show the same preferences with respect to the categorial realization of abstract anaphors. Similarly, translations of pronominal anaphors to more elaborate NP anaphors can be observed in both translation

directions (see the column 'Pronoun to NP' in Table 6, and the bars 'EN_t' and 'DE_t' in the top chart in Figure 3). This effect might be attributable to the translation process (and could be an example of explicitation).

However, to fully exclude language-specific tendencies, we would also need to compare relative frequencies in the comparable$_{trans}$ corpora (between DE_o and DE_t, and EN_o and EN_t, respectively), which is not possible at the current stage of the project because of the different sets of markables used in the rounds of annotation.

Table 7: Label nouns: Categorial transposition types

	NP to NP		NP to pron		NP to other		Sum	
DE_o–to–EN_t	87.2%	(102)	5.1%	(6)	7.7%	(9)	100%	(117)
EN_o–to–DE_t	90.0%	(117)	3.8%	(5)	6.2%	(8)	100%	(130)

Abstract anaphors with demonstrative label nouns Another kind of countercheck can be performed by investigating original NP anaphors and their translations. If many NPs were unexpectedly translated by pronouns, categorial transpositions from pronouns to NPs or vice versa would seem to be done at random.

In the Anaphora Corpus, the vast majority of label noun anaphors is translated by NPs, independent of the translation direction, see Table 7.[20] Only 4.5% of the label nouns are translated as pronouns (or as pronominal adverbs).

We conclude that there is a language-independent tendency that pronominal anaphors will be translated into full NPs, and that full NP anaphors till tend to remain full NPs in translation. This would conform with the explicitation hypothesis. §5.4 discusses individual translation examples in more detail.

4.3 Grammatical function

Pronominal abstract anaphors In the annotation of pronominal anaphors, only coarse-grained functions were annotated: *subject, object*, and *other*. Table 8 shows the translation equivalences for subjects and objects in both translation directions, DE_o–to–EN_t and EN_o–to–DE_t. As can be seen in the figure, German

[20] There are no significant differences between the two translation directions.

subject anaphors usually remain subjects in the English translation, whereas German object anaphors tend to become subjects in English as well. The non-literal translation in (6) results in such a transposition.

(6) *DE$_o$*: Das kann man nicht einfach so geschehen lassen.
EN$_t$: It is not such a simple matter.
DE$_{lit}$: That you cannot simply let happen. (ep-04-03-09/31)

Table 8: Pronouns: Transpositions of the functions *subject* and *object*

German original		English translation		English original		German translation	
Function	Freq	Function	Freq	Function	Freq	Function	Freq
subject	147	subject	107	subject	177	subject	114
		object	5			object	10
		other	35			other	53
object	55	object	27	object	37	object	18
		subject	12			subject	5
		other	16			other	14

As in §4.2, the bar plots in Figure 4 present a more general overview by summarizing the relative frequencies of grammatical functions in the Anaphora Corpus. The top chart in Figure 4 summarizes the distribution of grammatical functions with respect to pronominal anaphors and their translation equivalents. The cross-linguistic comparison of subjects and objects indicates significant differences: English uses more anaphoric subjects than German does.[21] In the comparable$_{trans}$ sets, we observe an overuse of anaphoric subjects in *DE$_t$*, which could be interpreted as a *shining-through* of English preferences.

Abstract anaphors with demonstrative label nouns In the annotation of the label nouns, we extended the set of functions, including a class *argument-after-preposition* ('arg-after-prep') to capture both prepositional objects and prepositional adverbials, and a class *attribute* to be used for all (prepositional and nominal) attributes of noun phrases.

In the majority of the translations, the original function is also used in the translated unit (*DE$_o$*–to–*EN$_t$*: 71.55% (83), *EN$_o$*–to–*DE$_t$*: 73.38% (91)).[22]

[21] Chi-squared test: $\chi^2 = 5.3953, df = 1, p < .05$

[22] The proportions do not differ significantly, according to a Chi-squared test: $\chi^2 = 0.0301, df = 1, p = .8622$.

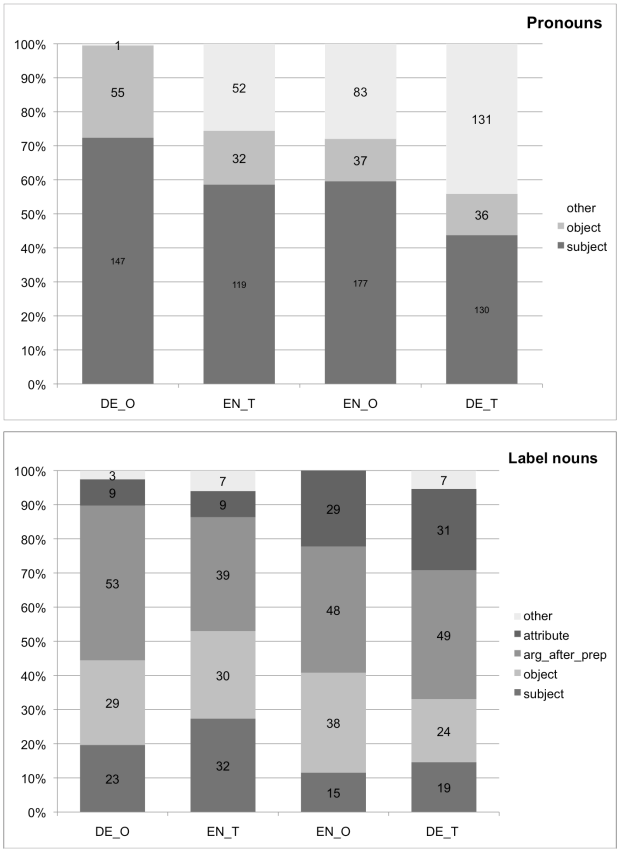


Figure 4: Relative frequencies of grammatical functions. The top chart refers to pronominal anaphors in the source languages and their translated equivalents, the bottom chart to label nouns.

However, there are some divergences: see Table 9, which lists interesting cases of transpositions of label noun functions. 17% of the 'arguments-after-prepositions' in DE_o are translated into subjects in EN_t. This is not mirrored in the opposite translation direction: only two out of 48 arg-after-preps in EN_o are translated as a subject in DE_t. We interpret this as a tendency for German prepositional phrases to be translated as subjects in English. An example is provided in (7).

(7) DE_o: Sie haben die Chance, in diesem Wettbewerb wirklich sehr vieles zusammenzuführen; regionale Kulturen können grenzüberschreitend zusammenarbeiten.

Table 9: Label nouns: Transpositions of functions. Only pairs discussed in the text are listed.

	Arg-after-prep to subject	Attribute to attribute	Object to attribute
DE_o–to–EN_t	17.0% (9/53)	66.7% (6/9)	3.5% (1/29)
EN_o–to–DE_t	4.1% (2/48)	72.4% (21/31)	18.4% (7/24)

EN_t: This competition gives them the opportunity to bring a very great deal of elements together; there can be cross-border cooperation between regional cultures.
DE_{lit}: They have the opportunity to bring a very great deal of elements together in this competition ... (ep-06-04-04/317)

English shows a characteristic tendency to realize abstract anaphors as NP attributes, in contrast to German, cf. Figure 4: 22.3% (29) of the abstract nouns in EN_o are realized as attributes, versus 7.8% (9) in DE_o.[23] If we examine the language pairs from the parallel corpora, the number of attributes do not significantly differ, because attributes are usually translated as attributes in both translational directions (cf. Table 9). The conservative mappings result in a *shining-through* effect in both directions.

As just noted, German generally avoids anaphoric attributes. Surprisingly, there are some cases in which English objects are translated by German attributes (7 cases, see the third column in Table 9), but there is only one case in the opposite direction. This is the effect of a strong tendency for nominalization in German. In (8), the English object of a subordinate clause is translated as an NP attribute in German.

(8) EN_o: Not all the decisions will be taken when we vote this report through.
DE_t: Mit unserer Zustimmung zu diesem Bericht werden nicht automatisch alle Entscheidungen getroffen.
DE_{lit}: With our agreement to this report not all points are decided automatically. (ep-00-05-16/19)

[23] The observed difference is significant, according to a Chi-squared test: $\chi^2 = 7.368, df = 1, p < .01$.

Finally, the bottom chart in Figure 4 shows the distributions of the functions observed with label nouns. The results are similar to those regarding pronominal functions.

Since the set of markables differ among the corpora, these are only preliminary conclusions. Further investigation is needed to verify the observed biases.

4.4 Clausal position

Grammatical categories (pronouns, full NPs, etc.) and grammatical functions (subject, object, etc.) are very similar in German and English, and the two languages can be directly compared to each other rather easily in these respects. In contrast, word order regularities are very different in the two languages. English has a fixed word order (S–V–O), whereas main clauses in German are *verb-second* (i.e., they allow any grammatical function to appear in the preverbal position, also called the *prefield* position).

Both languages have extra ways to mark or highlight constituents, such as cleft or topicalized constructions, which serve to place a constituent intended to be emphasized at the beginning of a sentence. Such special constructions are more often used in English than in German, probably because the prefield position in German already serves this purpose to some extent.

Sentence-initial positions play an important role in information structure: Old information tends to occur early in the sentence, new information towards the end. As abstract anaphors refer to previously mentioned referents, they represent old information. We therefore hypothesize that anaphors will tend to occur in topicalized or prefield positions.

(9) shows a relevant case: A German prefield instance is translated by a topic construction (*that is something*) in English.

(9) DE_o: Wenn es leichter ist, an die Subventionen zu gelangen, dann steigt auch die Nachfrage dafür. Dies halten wir gerade bei kleinen Programmen für notwendig.
EN_t: If subsidies are more readily obtainable, the demand for them will rise, and that is something we regard as needed, particularly by small programmes.
DE_{lit}: …This we regard as needed, particularly by small programmes.
(ep-05-10-24/68)

Our annotation distinguishes between three different positions for anaphors: in the *matrix* clause, in a *subordinate* clause, or in a sentence-initial position,

which includes topic-like constructions in English (annotated as *topic*) and the *prefield* position in German.[24]

However, as explained above, we cannot directly compare these positions to each other, due to language-inherent differences in syntax. Therefore, we must restrict our comparisons to the comparable$_{trans}$ corpora in this case.

Pronominal abstract anaphors The top charts in Figure 5 show the relative proportions of pronominal anaphors across the clausal positions.

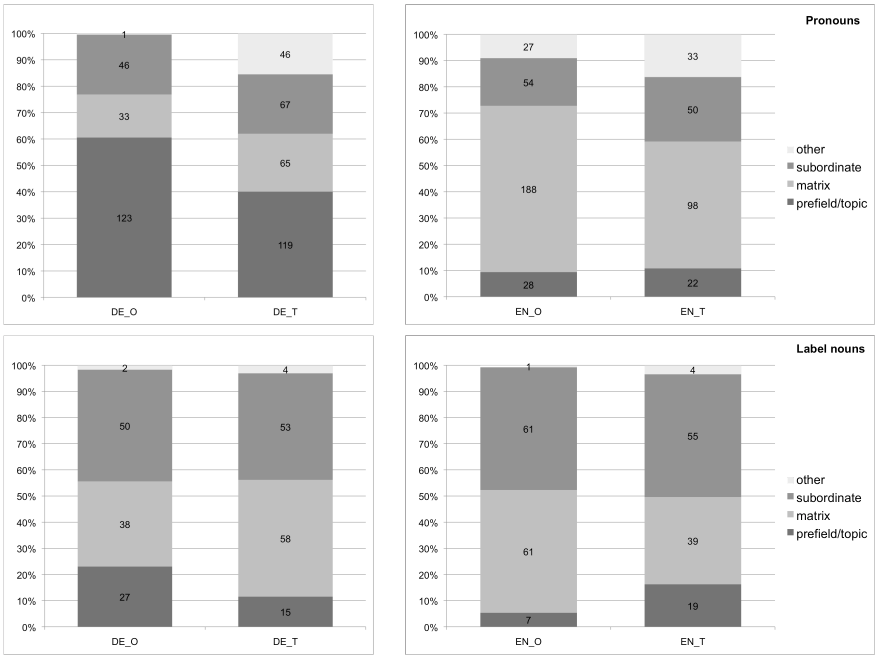


Figure 5: Relative frequencies of clausal positions. The top charts refers to pronominal anaphors, the bottom charts to label nouns. Only the pairings DE_o–DE_t and EN_o–EN_t can be compared to each other.

In comparing the two German corpora, we observe a significant underuse of prefield anaphors in DE_t: that is, pronominal anaphors in DE_o occur considerably more often in the prefield and less frequently in the (rest of the) matrix

[24]Note that our label *matrix* is assigned to constituents in the matrix clause, *except for* constituents in the topic or prefield position.

clause.[25]This indicates that translated texts do not follow our hypothesis to the same extent as original texts do.

A different effect is observed in the English corpora: EN_t shows a significant underuse of anaphors in the matrix position; this is counterbalanced by an overuse of anaphors in subordinate clauses.[26]Anaphors in topic positions are very rare, contradicting our (simplistic) hypothesis.

Abstract anaphors with demonstrative label nouns The distribution of label nouns clearly differs from the distribution of pronominal anaphors, as can be seen in Figure 5. Whereas pronouns in German are preferably realized in the prefield position (cf. top charts), there is no such preference for label noun anaphors in our data (cf. bottom charts). Instead, label nouns are preferably realized in matrix and subordinate positions.[27] For English, we observe a significant overuse of anaphors in topic constructions in EN_t.[28]

It would be interesting to relate these observations to *shining-through* effects; however, we cannot draw this conclusion on the basis of our annotations. The annotated concepts (topic, prefield) would first have to be calibrated to each other.

5 Edit operations and lexical specificity: Case studies

The previous section presented quantitative results from the comparison of our parallel and comparable$_{trans}$ corpora, focusing on various properties of pronominal and label noun anaphors, such as grammatical category and grammatical function. In this section, we investigate a range of case studies in hopes of shedding light on selected details of our data.

We focus on examples in which the translated anaphor differs from the pattern of its source, i.e., cases in which material has been added, omitted, or substituted.

[25]Proportion of matrix in DE_o: 60.9% (123/202) versus DE_t: 96.7% (119/123); Chi-squared test: $\chi^2 = 7.6415, df = 1, p < 0.01$.

[26]Proportion of matrix in EN_o: 69.6% (188/270) vs. EN_t: 57.6% (98/170); Chi-squared test: $\chi^2 = 6.0677, df = 1, p < 0.05$. Proportion of subordinate in EN_o: 20.0% (54/270) vs. EN_t: 29.4% (50/170); Chi-squared test: $\chi^2 = 4.6114, df = 1, p < 0.05$.

[27]The observed asymmetry between pronouns and label nouns is probably a reflection of the universal tendency of pronouns to occur very early in the sentence, whereas no such general tendency exists for full NPs.

[28]Proportion of topic in EN_o: 5.4% (7/129) vs. EN_t: 16.8% (19/113); Chi-squared test: $\chi^2 = 7.0016, df = 1, p < .01$.

We call these processes *edit operations*, following the common terminology in computational linguistics (Levenshtein 1965). An obvious (and highly simplistic) hypothesis would be that an increase in the length of translated anaphors could be an effect of explicitation.[29]

There are numerous ways to add, omit, or substitute material in a label noun NP, and we examine some of these in detail. We investigate the addition or omission of adjectives in label noun NPs (§5.1), the substitution of nouns by more general or more specific nouns (§5.2 and §5.3), the substitution of full NPs by pronouns and vice versa (§5.4), and the substitution of the demonstrative determiner by various types of expressions (§5.5).

Edit operations often have an effect on the *specificity* of anaphors. We refer to an expression as being *more specific* than another expression if it has fewer possible interpretations. Very often, the addition of material (such as the addition of adjectives, or the expansion of a pronoun to a full NP) results in higher specificity. As the discussions in the next sections show, translations both increase and decrease the specificity of anaphors (contrary to the assumptions made by the explicitation hypothesis).

5.1 Adjectival modifications

In this section, we consider NPs with adjectives in either the original or the translated sentences. The examples illustrate that some of these adjectives contribute to the specificity of the NP, while others do not. We observed both situations: adjectives being added in the translation, and adjectives omitted. In the Anaphora Corpus, relevant cases were found only in the translation direction EN_o–to–DE_t (but not in DE_o–to–EN_t).

In several cases, the German translated NP contains the adjective *vorliegend* 'present', but there is no correspondent in the original English sentence, cf. (10). This adjective clearly serves only a deictic function, i.e., it assumes the meaning of *this* in the English NP. Consequently, in all these cases, the demonstrative article *this* is translated by the definite article in German (which is fused with the preposition: *in dem* 'in the' becomes *im*). Thus, the German version of the abstract NP is in fact a very close translation of the original NP in English.

[29] Of course, there are clear cases of length differences that must be removed from such considerations, such as multi-word expressions and compounds, which are usually spelled in one word in German and in several words in English. Further counter-examples to this hypothesis are presented in the following subsections.

(10) EN_o: This exercise has been made possible in this case because of the work of national and international bikers' rights organisations coordinated by the Federation of European Motorcyclists, or FEM.
DE_t: Ein solcher Dialog wurde im vorliegenden Fall durch die vom Verband Europäischer Motorradfahrer, VEM, koordinierte Arbeit nationaler und internationaler Organisationen für die Rechte von Motorradfahrern ermöglicht.
DE_{lit}: This exercise has been made possible in the present case …
(ep-96-06-18/252)

In other examples, adjectives are omitted. In several cases, this concerns the adjective *whole* not being translated in the corresponding German sentences.[30] In these examples, the information provided by the original English *whole*-NP is more elaborate than the translated German NP. For instance, in the German part of (11), it is not specified that the *whole* area is involved. It would therefore be possible to continue the clause by actually limiting the area in the following way: *(much progress has been made in this area) — not in all parts/aspects, but in most of them.* This reading is not possible for the English original NP. In this sense, we can state that the original NP in English is indeed more specific than its German counterpart in these examples.

(11) EN_o: We have to note that much progress has been made in this whole area.
DE_t: Wir müssen feststellen, dass in diesem Bereich große Fortschritte erzielt wurden.
DE_{lit}: We have to note that in this area much progress has been made.
(ep-97-04-08/304)

Finally, in one example, the adjective *particular* has been omitted, see (12). The contribution by this adjective is different from the contribution of *whole* above. Here, the adjective serves as a marker of focus. In contrast to the above example, omitting the marker in German does not allow a different interpretation of the respective NP. Hence, we would not classify the German translation as less specific. (Of course, the German translation lacks the contribution of the focus marker, but this seems unrelated to specificity.)

(12) EN_o: As a British Member, I am optimistic that the British Presidency can maintain the momentum that was picked up originally by the

[30] In one case, the adjective *ganz* 'whole' was added in the translation.

Luxembourg Presidency and that will be carried on through the Austrian and German presidencies because there is much to do in this particular area.
DE_t: Als britischer Abgeordneter bin ich zuversichtlich, dass die britische Präsidentschaft den Prozess, der ursprünglich von der luxemburgischen Präsidentschaft begonnen wurde, in Gang halten wird und dass er auch unter dem österreichischen und deutschen Vorsitz weitergeführt werden wird, denn in diesem Bereich gibt es noch viel zu tun.
DE_{lit}: ...because in this area there is still much to do. (ep-98-02-19/225)

Comparing these three examples ((10)–(12)), we see that only one type of adjective actually has an impact on the specificity of the abstract NP.

5.2 Lexical semantics of nouns

In this section, we consider examples in which the lexical semantics of the nouns has an effect on the specificity of the abstract NP. Either the original or the translated noun can be more specific.

Most of the examples are found in EN_o–to–DE_t translations. In most of these cases, the German translations are more specific than the English originals. A clear example is provided in (13). The original English noun, *issue*, is highly generic: if one did not know the context, a large set of interpretations would be possible. In contrast, the German translation, *Erweiterung* 'expansion' is much more specific.

(13) EN_o: I would ask the President-in-Office to continue to champion this issue and emphasise it consistently in Göteborg, especially with a view to enabling the Irish to say "yes" to enlargement there.
DE_t: Ich bitte die Ratspräsidentin, ihr Engagement für die Erweiterung fortzusetzen und dieses Thema auch in Göteborg konsequent in den Vordergrund zu rücken, damit die Iren sich auf diesem Gipfel klar und deutlich für die Erweiterung aussprechen können.
DE_{lit}: I would ask the President-in-Office to continue to champion the expansion ... (ep-01-06-13/8)

Similar, if somewhat more ambiguous examples, can be seen in (14) and (15). In (14), the English original noun *message* is less specific than the German translation *Zusage* 'assurance'. Out of context, the English noun *message* could refer to an assurance or a denial. The *denial* reading is obviously not possible in the

German translation, which makes it more specific than the English original in this respect.

(14) EN_o: If we reverse that message now we run the risk of undermining all the reforms which have taken place at great pain in Central and Eastern Europe.
DE_t: Wenn wir jetzt von dieser Zusage abweichen, gefährden wir alle Reformen, die in Mittel- und Osteuropa mit großer Mühe unternommen wurden.
DE_{lit}: If we depart from this assurance now we run the risk of undermining all the reforms … (ep-96-04-17/58)

Similarly, in (15), the German translation *Zwecke* 'purposes' is more specific than the original English noun *way*. For example, *spending money in that way* could refer to spending money for a specific purpose, or to spending money over a certain amount of time. In contrast, the German noun *Zwecke* only permits the first interpretation.

(15) EN_o: The continued spending of money in that way is unacceptable.
DE_t: Die fortgesetzte Verwendung von Mitteln für diese Zwecke ist unvertretbar.
DE_{lit}: The continued spending of money for these purposes is unacceptable. (ep-01-04-03/46)

It should be noted that although most of the translated nouns are more specific than the original nouns, rare examples in the other direction also exist. For example, (16) involves *request* as the original English noun. The German translation is *Fall* 'case', which is clearly less specific than the English original (but connects back to a previous use of the word 'case' in the same sentence).

(16) EN_o: But the third came with the thumbprint of Government on it, unlike this request, so it is an inadequate precedent, even if it is a modest step in that direction.
DE_t: Beim dritten Fall war die Regierung involviert, anders als in diesem Fall, weshalb er als Präzedenzfall ungeeignet ist, selbst wenn er ein bescheidener Schritt in diese Richtung ist.
DE_{lit}: In the third case, the Government was involved, unlike as in this case, so it is an inadequate test case … (ep-01-05-02/31)

5.3 Impact of context

Consideration of the lexical semantics of nouns can help to locate translation examples in which specificity differs between the original and translated texts. However, it is not enough to simply consider pairs of nouns or NPs. If there is a mismatch between the NPs, the missing information can also be expressed in other parts of the sentence.

In (17), the English translation *thing* seems to be much less specific than the German original noun *Forderung* 'request'. However, the meaning corresponding to *Forderung* is instead expressed in the English verb *calling for*.

(17) DE_o: Ich sehe diejenigen, die jetzt in Briefen an uns eine Maximalharmonisierung fordern – gerade im Bereich des Verbraucherschutzes –, schon wieder sagen: Das ist zu viel Harmonisierung! Stichwort: Verbraucherkreditrichtlinie; daher sollten die Marktteilnehmer sehr vorsichtig mit dieser Forderung umgehen.
EN_t: I can imagine those who currently write to us demanding maximum harmonisation in consumer protection matters saying – yet again – that we are taking harmonisation too far with the Consumer Credit Directive; that is why they should be very careful when calling for such a thing.
DE_{lit}: …therefore the market players should be very careful with this request. (ep-05-04-27/120)

Further apparent specificity mismatches can arise when the sentence structure is changed considerably during translation. In (18), the German original NP, *diese Strategie* 'this strategy', is less specific than the translated NP *the Lisbon strategy*. However, the English translation does not actually provide any more information than the original German sentence: The German NP *diese Strategie* refers back to the antecedent *Lissabon-Strategie* (printed in bold in the example). In the English translation, the sentence structure has been changed so that the NP in question is the first mention of the abstract object, and therefore refers to *Lisbon* (the second mention being *it*).

(18) DE_o: Ich danke dem Kok-Bericht; das, was wir jetzt dringend brauchen, ist eine Ausrichtung **der Lissabon-Strategie**, denn diese Strategie ist richtig.
EN_t: I am grateful for the Kok report; what we now urgently need – as the Lisbon strategy is the right one – is an orientation for **it**.
DE_{lit}: …what we now urgently need is an orientation for **the Lisbon strategy**, as this strategy is right. (ep-04-11-17/38)

This discussion demonstrates that we must be careful in drawing conclusions from purely statistical data. Even detailed information about word-to-word correspondences (such as the noun pairs discussed in this section) can be misleading. It is therefore important to also consider the noun pairs in context. However, analysis of statistical counts and noun pairs as a first step can help to detect noteworthy examples.

5.4 Pronouns vs. NPs

As discussed in §4.2, pronouns are often translated into full NPs, both in DE_o–to–EN_t and EN_o–to–DE_t. In this section, we examine some of these cases in greater detail.

For example, in (19) (= 5), the English pronoun *this* corresponds to the full NP *diesem Standpunkt* 'this position' in the German translation. The pronominal anaphor *this* in the English original sentence can in principle refer to different kinds of objects, such as a process, a rejection, an undertaking, etc. This flexibility is eliminated in the German translation, in which it is explicitly specified that the speaker does not support the *position*.

(19) EN_o: I do not necessarily support this.
DE_t: Diesem Standpunkt schließe ich mich nicht notwendigerweise an.
DE_{lit}: This position I do not necessarily support. (ep-00-10-03/15)

In a similar way, the German original pronominal anaphor *das* 'that' is less specific than its English translation *this threat* in (20). The pronominal anaphor could also refer to a development, for example, an interpretation that is unlikely for the corresponding English expression *this threat*. In the German sentence, however, the verb *abwenden* 'avert' provides important clues and restricts the set of possible referents to those with negative connotations.

(20) DE_o: Das konnte durch die glänzende Vorsitzführung von Frau Cederschiöld, aber auch durch die sehr substanzielle Hilfe der Kommission abgewendet werden, und deswegen können wir diesem Kompromissergebnis zustimmen.
EN_t: Thanks to Mrs Cederschiöld's inspired leadership, but also due to the very substantial support from the Commission, this threat has been averted, so we can now vote in favour of this compromise result.
DE_{lit}: That could be averted by Mrs Cederschiöld's inspired leadership, but also due to the very substantial support from the Commission … (ep-04-01-28/109)

These examples were taken from a wide range of sentences in which an original pronominal anaphor was translated with a more specific full NP. In the other direction (i.e., from original abstract demonstrative NPs to translated pronouns), only rare examples can be found. (21) is such an example: German *diese Ansicht* 'this view' is translated with the pronominal *that* in English. The verb *agree*, however, is only compatible with a small range of readings for the pronoun: *that* could refer to, e.g., a judgment, assessment, opinion, or the like—quite similar concepts. Due to the use of the verb *agree*, the pronominal translation is only marginally less specific than the original full NP.

(21) DE_o: Sie schreiben, dass es nicht sinnvoll ist, Beihilfen für Investitionen an Unternehmen zu geben, die profitträchtig sind. Diese Ansicht teile ich.
EN_t: He writes that it makes no sense to give aid to businesses that are already profitable, and in that I agree with him.
DE_{lit}: He writes that it makes no sense to give aid to businesses that are already profitable. This view I share. (ep-06-02-13/115)

There are some very unusual examples in which the translated sentence is indeed less specific than its original counterpart. In (22), *dieser Effekt* 'this effect' in German corresponds to the English pronoun *this*. English *this* could refer to a development or a threat that has been exacerbated, but the German full NP does not allow these readings.

(22) DE_o: Dieser Effekt wird noch dadurch verstärkt, dass junge Mädchen nicht mehr zur Schule gehen können, weil sie ihre an Aids erkrankten Eltern pflegen müssen.
EN_t: This is exacerbated by the fact that young girls are no longer able to attend school because they have to care for their parents who are sick with AIDS.
DE_{lit}: This effect is exacerbated by the fact that … (ep-04-01-13/306)

Taking prior context into account, the discourse model that speakers and hearers have built up thus far might provide very clear constraints for the reference of *this*, so that no further specifications (such as using the noun *effect*) would be necessary. The issue of interest to us is that in most cases in which the original contribution uses a full NP, the translator also uses a full NP. In other words, if the author of the original contribution finds it necessary to spell out the referring expression in detail, this detail is probably required in order to avoid misinterpretation, and the translator will face the very same situation in the target language (especially for languages as similar as German and English).

Thus, whenever the translator deviates from the original version in this way, it could indicate an interesting example for detailed examination, both in the original and in the translated texts.

5.5 Transposition of the demonstrative determiner

In this subsection, we investigate cases that involve translations without (canonical) demonstrative articles. Remember that in the annotations with label nouns, only those noun chunks that contained a demonstrative determiner were premarked. We therefore expect close translations to contain a demonstrative determiner as well.

In total, we found 20 instances in EN_t that did not contain such a determiner, and 34 instances in DE_t. In many cases (14 in EN_t and 13 in DE_t), the abstract NP is translated either by a pronoun or by a diverging syntactic construction.

Some instances in DE_t employ a strategy that we addressed above (see Section 5.1): Adjectives, such as *vorliegend* 'present, at hand' and *last-mentioned* are used to convey the deictic meaning.

In some cases, the demonstrative pronoun is replaced by a possessive in the translated sentence. In our corpus, this occurs in English original sentences and their German translations. Some examples also involve minor changes in the overall structure of the sentence. In (23) (=4), the English speaker thanks the rapporteur for producing the report. In the German translation, *producing* is not translated but is instead replaced by the possessive pronoun.

(23) EN_o: Madam President, I would like to thank the rapporteur for producing this report because it is a very important one.
DE_t: Frau Präsidentin, ich möchte dem Berichterstatter für seinen Bericht danken, denn es handelt sich um einen wirklich wichtigen Bericht.
DE_{lit}: Madam President, I would like to thank the rapporteur for his report because it is a very important report. (ep-98-11-17/284)

In the remaining cases, we observe a variety of situations. In some sentences, the specificity of the anaphoric noun seems considerably reduced in the translation. In most of these examples, *such (a)* serves as a substitute determiner, see (24). Like canonical demonstratives, *such* has a deictic component but points to a type or set of entities that share certain properties rather than to a specific entity. In another example, the demonstrative NP is translated by an unspecific negated NP, see (25).

(24) EN_o: The Commission, however, intends [to] bring forward a Council regulation on the control of unloading and transfers: this proposal is already being prepared and the Commission believes it should provide a more appropriate framework.
DE_t: Die Kommission beabsichtigt vielmehr, eine Verordnung des Rates betreffend die Kontrolle von Aus- und Umladungen vorzuschlagen: Ein solcher Vorschlag wird bereits vorbereitet und dürfte nach Ansicht der Kommission einen angemesseneren Rahmen bilden.
DE_{lit}: ...such a proposal is already being prepared ... (ep-98-03-13/71)

(25) EN_o: It is regrettable that we cannot yet achieve that full agreement.
DE_t: Es ist bedauerlich, daß wir noch keine vollständige Einigung erzielen können.
DE_{lit}: It is regrettable that we can yet achieve no full agreement. (ep-97-04-08/304)

Finally, in (26) (= 13), the abstract label noun is translated by a lexically more specific noun. As a result, the space of possible references is narrowed and therefore use of the demonstrative determiner seems superfluous (see the discussion in §5.2).

(26) EN_o: I would ask the President-in-Office to continue to champion this issue and emphasise it consistently in Göteborg, especially with a view to enabling the Irish to say "yes" to enlargement there.
DE_t: Ich bitte die Ratspräsidentin, ihr Engagement für die Erweiterung fortzusetzen und dieses Thema auch in Göteborg konsequent in den Vordergrund zu rücken, damit die Iren sich auf diesem Gipfel klar und deutlich für die Erweiterung aussprechen können.
DE_{lit}: I would ask the President-in-Office to continue to champion the expansion ... (ep-01-06-13/8)

6 Conclusion

In this paper, we have presented a bootstrapping approach to the annotation of pronominal and label noun anaphors. Based on our annotated data, we investigated selected properties of the anaphors in greater detail. Before summarizing our findings, we would like to emphasize that all our results should be understood as valid only for the particular type of language represented in the Europarl corpus — namely, spoken and translated parliamentary debates. This holds for both

the differences between original and translated texts as well as for the language-specific properties that we have identified. It remains to be seen to what extent our findings will generalize to other domains and text types.

Lexical choice Original and translated texts showed identical preferences with regard to pronominal anaphors: *das* 'that' in German, and *this, that* in English. Translated German texts showed an interesting significant overuse of *dies* 'this', which might be an effect of *shining-through*, reflecting the high frequency of its English counterpart *this*.

Certain label nouns occurred very often in our data. This is related to the domain of our data: parliamentary debates. Nevertheless, when we compared the frequencies of selected label nouns in original and translated turns thoughout the entire Europarl Corpus, interesting (and statistically significant) discrepancies stood out.

Based on our annotated data, the German noun *Angelegenheit* 'issue' seemed to serve as a kind of "dummy" translation. With the noun *Bereich* 'area', we observed an interesting asymmetry: When translated into English, a variety of English expressions were used (e.g., *area, issue, subject, sphere*), whereas German translators employed *Bereich* quasi-exclusively as the translation for *area*.

Category, function, position Translations in general tended to preserve the anaphor's categories, functions, and positions; however, some interesting differences were observed.

With regard to category, we observed a clear asymmetry: A considerable number of pronouns were translated as full NPs, while the reverse was not true. Since the asymmetry appeared in both languages, this might have been an effect of the translation process, perhaps due to translational conventions (in the form of "do not use pronouns"). Very rarely could the opposite mapping be observed. As in the case of lexical semantics (see below), the context sometimes compensated for the loss of specificity.

At the functional level, we observed a preference for anaphoric attributes in original English texts, in contrast to German. This resulted in an overuse of these attributes in DE_t and an underuse in EN_t, i.e., a *shining-through* effect in both directions.

Finally, with respect to the positional properties of the anaphors, both languages exhibited language-typical patterns in both original and translated texts. *Shining-through* effects were found here as well: DE_t underused anaphors in the

prefield position, while EN_t underused matrix anaphors in comparison to subordinate anaphors.

Adjectival modifications Adjectives such as *whole* were sometimes omitted, even when this could result in under-specification and various possible interpretations. Such omissions mainly occurred in the translation direction EN_o–to–DE_t. That is, in these cases, the German translations were less specific than their sources.

Lexical semantics Most of the cases in which the original and translated nouns differed with respect to their specificity were found with EN_o–to–DE_t translations. The German translations were generally more specific than their English counterparts. (This might outweigh the tendencies described in the previous paragraph to some extent.)

In certain cases, the immediate context (e.g., the main verb) compensated for the loss of specificity in the nouns.

Transposition of demonstratives Two cases were of interest here: First, specific demonstrative NPs were sometimes translated by *such* (or its German equivalent). The speaker no longer referred to the specific entity in question but to all entities of the same kind. Second, the demonstrative article was sometimes translated by a definite article. In these cases, the deictic function of the demonstrative often seemed to be taken over by adjectives such as *vorliegend* 'present'.

The amount of data that we examined was rather small, so we consider the research reported here to be a pilot study that can serve as a starting point for further in-depth analyses. In order to derive more reliable conclusions, we need more data. This can be achieved in several ways.

In the next annotation round, the translated nouns that have not yet been included in our label noun list will be added and annotated.

We also plan to provide the annotators with translation candidates that have been automatically selected from all noun chunks in the aligned translated turn. To this end, we intend to use heuristics derived from our present findings, e.g., using the most common translation equivalents for nouns and marking NPs containing modifiers such as 'present' or 'at hand' as promising candidates (in addition to demonstrative NPs). Pre-selecting such candidates in the aligned translated turns will make the annotation procedure simpler and more efficient.

Thus far, we have only annotated and aligned pairs of turns that contain the pronouns *it, this*, and *that* (and their German equivalents) and demonstrative

NPs with label nouns in the original turns. No such restrictions applied to the translated turns; here the annotators were free to mark arbitrary strings as the expression that represented the translation of the anaphor in the original text. As we have seen, however, translators very often stay close to the original. Therefore, we cannot expect to discover exceptional ways of referring to abstract entities in translated texts very often. To complement our 'restricted' approach, it would be useful to annotate a sample of running text, marking all types of abstract anaphors that appear.

Finally, we would like to take advantage of the fact that Europarl provides debate protocols in many other languages, and expand our studies to include additional languages.

References

Asher, Nicholas. 1993. *Reference to abstract objects in discourse*. Boston: Kluwer.

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

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

Byron, Donna K. 2003. *Annotation of Pronouns and their Antecedents: A comparison of two domains*. Technical Report, University of Rochester.

Cartoni, Bruno, Sandrine Zufferey, Thomas Meyer & 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*, 78–86.

Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann & Mihaela Vela. 2008. Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. In *Proceedings of the LREC workshop on comparable corpora*, 47–51. Marrakesh.

Dipper, Stefanie, Christine Rieger, Melanie Seiss & Heike Zinsmeister. 2011. Abstract anaphors in German and English. In Iris Hendrickx, Sobha Lalitha Devi, António Branco & Ruslan Mitkov (eds.), *Anaphora processing and applications: 8th discourse anaphora and anaphor resolution colloquium, DAARC 2011. revised selected papers*, 96–107. Berlin: Springer.

Dipper, Stefanie, Melanie Seiss & Heike Zinsmeister. 2012. The use of parallel and comparable data for analysis of abstract anaphora in German and English. In *Proceedings of the LREC-12*. Istanbul.

Dipper, Stefanie & Heike Zinsmeister. 2009. Annotating discourse anaphora. In *Proceedings of LAW iii*, 166–169.

Dipper, Stefanie & Heike Zinsmeister. 2010. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC 2010 workshop on language resource and language technology standards*, 54–59. Valletta.

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

Francis, Gill. 1994. Labelling discourse: An aspect of nominal group lexical cohesion. In Malcolm Coulthard (ed.), *Advances in written text analysis*, 83–101. London: Routledge.

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

Halteren, Hans van. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd international conference on computational linguistics COLING 08*, 937–944.

Hedberg, Nancy, Jeanette K. Gundel & Ron Zacharski. 2007. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of daarc-2007: 6th discourse anaphora and anaphora resolution colloquium*, 31–36.

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

Klaudy, Kinga. 2008. Explicitation. In Mona Baker & Gabriela Saldanha (eds.), *Routledge encyclopedia of translation studies. 2nd edn.* 104–108. London: Routledge.

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

Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163(4). 845–848.

Müller, Christoph. 2007. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of acl-07 conference*, 816–823.

Müller, Christoph & Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*, 197–214. Frankfurt a.M.: Peter Lang.

Navarretta, Costanza. 2008. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In *Proceedings of the second workshop on anaphora resolution*, 63–71.

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

Poesio, Massimo & Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC-08*.

Pradhan, Sameer, Lance Ramshaw, Ralph Weischedel, Jessica MacBride & Linnea Micciulla. 2007. Unrestricted coreference: identifying entities and events in OntoNotes. In *Proceedings of the ieee-icsc*.

Recasens, Marta. 2008. Discourse deixis and coreference: evidence from AnCora. In *Proceedings of the second workshop on anaphora resolution*, 73–82.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision tree. In *Proceedings of international conference on new methods in language processing*.

Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin: De Gruyter.

Vieira, Renata, Susanne Salmon-Alt & Caroline Gasperin. 2002. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of daarc-2002: 4th discourse anaphora and anaphora resolution colloquium*.

Vinay, Jean-Paul & Jean Darbelnet. 1958/1995. *Comparative stylistics of french and english: A methodology for translation*. Amsterdam: Benjamins.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin & Ann Houston. 2010. *OntoNotes Release 4.0, with OntoNotes DB Tool v. 0.999 beta*. Tech. rep. http://www.bbn.com/NLP/OntoNotes.