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

Heike Zinsmeister*
University of Stuttgart
heike.zinsmeister@ims.uni-stuttgart.de

Stefanie Dipper
Ruhr-University Bochum
dipper@linguistics.rub.de

Melanie Seiss
University of Konstanz
melanie.seiss@uni-konstanz.de

Abstract anaphors refer to abstract referents such as facts or events. This paper presents a corpus-based comparative study on German and English abstract anaphors. Parallel, bi-directional texts from the Europarl Corpus have been annotated with functional and morpho-syntactic information, focussing on the pronouns 'it', 'this', 'that' and 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 also looking into the differences between original text and translated text in each of the languages. In selected case studies, we investigate "translation mismatches" in detail, which involve changes of the grammatical categories (from pronouns to full noun phrases and vice versa), grammatical functions or clausal positions, addition or omission of modifying adjectives, changes in the lexical realization of the head nouns, and transpositions of the demonstrative determiner. In some of these cases, specificity of the abstract noun phrase was altered by the translation process.

1 Introduction

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

*We would like to thank our two anonymous reviewers for their helpful comments and Claire Bacher for proof reading our text. Heike Zinsmeister's research was partly financed by Europäischer Sozialfonds. All URLs were accessed on May 16, 2012. We used R (http://www.r-project.org/) to compute statistical significance.

We pursue a contrastive, corpus-based approach to investigate the properties that characterize different instantiations of abstract anaphora in English and German. In the long run, we envisage to derive features from the corpus annotation that will serve us to tackle the automatic resolution of abstract anaphora.

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

We present results of a comparative corpus study on the realization of abstract anaphors in a parallel bi-directional corpus of English and German. Besides comparing the cross-linguistic realizations, we also look into the 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 and Zinsmeister, 2009). In this paper, we take into account both pronouns and a selection of full NPs. The NPs considered here contain a demonstrative determiner because demonstrative NPs are most likely used anaphorically. In addition, the NP's head must be an abstract noun, such as *issue, effect, process*. We contrast quantitative results from our previous studies with results from annotations of full NPs that we accomplished recently.

In addition, we investigate selected samples of "translation mismatches" in detail. On the one hand, these include anaphors that are not translated word by word but involve *edit operations*, i.e. addition, deletion, or substitution of words. On the other hand, such mismatches also concern *specificity*, i.e., translation mismatches that have an impact on the amount of information that is available to the hearer to resolve reference of the abstract anaphor, when anaphors are not translated by the most obvious translation candidate but by some target word that is more or less specific than its source word.

The corpus that has been annotated so far only allows for tentative conclusions. We consider the research reported here as a pilot study that highlights aspects and areas that seem worth being investigated on a large scale in the future.

The paper is organized as follows. Sec. 2 addresses related work, Sec. 3 introduces the corpus and its annotations that this study is based on. In Sec. 4, we present quantitative investigations concerning selected properties of the abstract anaphors, such as grammatical category, grammatical function, and position. Sec. 5 introduces a range of case studies that deal with translation mismatches. Sec. 6 provides the conclusion.

2 Related work

The majority of projects that analyze abstract anaphora deal with monolingual data. The presentation here starts with a short overview of relevant projects in general, and then addresses projects in more detail that consider multilingual corpora.

General studies Most annotation projects that analyze abstract anaphora restrict themselves 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 and Artstein (2008)). In projects that analyze pro-drop languages also zero anaphora have been considered (cf., e.g., Recasens (2008), Navarretta and Olsen (2008)). A recent overview of projects

annotating abstract anaphora is provided by Dipper and Zinsmeister (2010).

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

Recasens (2008) compares the use of pronominal and NP abstract anaphors in Catalan and Spanish. She shows 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. It turns out that the heads of these full NPs largely overlap with the "label nouns" reported by Francis (1994), which we also use in our study (see Sec. 3).

Navarretta (2008) and Navarretta and Olsen (2008) compare pronominal abstract anaphors in Danish and Italian. They find that Italian in general disprefers pronouns as abstract referents, and seems to use full NPs instead.

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

Parallel studies Annotation of parallel texts has been performed in Vieira et al. (2002), who use a subcorpus from the parallel MLCC corpus.[1] They investigate demonstrative NPs in French and Portuguese. Results for both languages are similar: demonstrative NPs predominantly have an abstract head noun. Vieira et al. (2002) do not distinguish between texts in original vs. translated language.

Characteristics of parallel corpora Parallel corpora such as MLCC (see above) or Europarl (Koehn, 2005) consist of original and translated texts. There is a long debate to what extent translated language deviates from comparable original language due to influences from both the original source language and the translation process and, hence, should not be used as the base of linguistic investigations (other than ones that focus on translation issues such as, e.g., Čulo et al. (2008)); see the discussion in Sec. 4.

For instance, van Halteren (2008) shows that based on word *n-grams* it is possible to identify the source language in Europarl translations with accuracies between 87.2–96.7%. Cartoni et al. (2011) investigate the use of discourse connectives in original and translated French texts from Europarl. They find that translated texts contain significantly more discourse connectives than original texts.

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 delivered (mostly) in

[1] The MLCC corpus contains 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. [All urls in the article have been accessed on May 15, 2012]

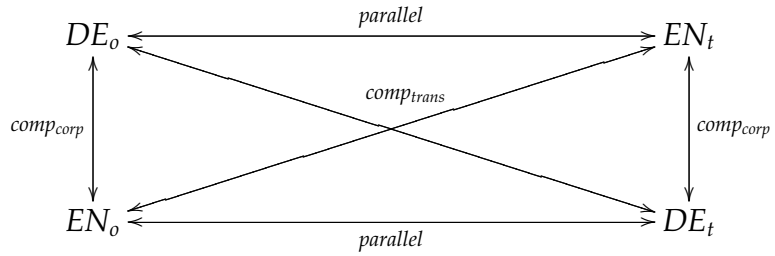

Figure 1: Three types of relations hold between the four subcorpora: parallel, comparable in the corpus-linguistic sense ($comp_{corp}$) and comparable in the translation-studies sense ($comp_{trans}$)

the speaker's mother tongue. 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. Hence, the register of the turns is of a mixed character, between spoken and a more standardized written language.

We created different 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) as well as the English original turns (EN_o) and their German translations (DE_t).

These four subcorpora stand in different relations to each other, cf. Figure 1. EN_o and DE_t as well as DE_o and EN_t are *parallel* corpora, i.e., they are original texts and their translations. On the other hand, the subcorpora DE_o and EN_o (and similarly, DE_t and EN_t) are comparable corpora, i.e. corpora in different languages which deal with the same overall topic and are from the same overall register. This notion of comparable corpora is usually used in corpus-linguistic research. Hence, we call this type of relation $comparable_{corp}$. Finally, the subcorpora DE_o and DE_t (and EN_o and EN_t) are also comparable corpora, in that they represent varieties of the same language. Translation studies usually refer to such corpora as comparable, hence we call this type of relation $comparable_{trans}$. We base our investigations, which are presented in this paper, on the different types of relations between the subcorpora.

Anaphora Corpus We randomly selected about 100 turns from DE_o and EN_o respectively, for our manual annotation study to investigate properties of abstract anaphors, in particular the realization as pronouns or full lexical NPs as well as function, position etc. For this, different preprocessing steps were applied. They included verifying the mother tongue of the speakers.[2] This left us with 94 German original turns and 95 English original turns. Further preprocessing 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 and Strube, 2006).

[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 all EU parliament members.

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

We call our collection of annotated subcorpora the *Anaphora Corpus*.

3.1 Annotating pronominal abstract anaphors

We took a cross-linguistic bootstrapping approach for the annotation of abstract pronouns: we started out with a well-defined set of markables in the original language and 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 in German there is—in contrast to English—a pronoun that is unambiguously used as an abstract anaphor: the uninflected singular demonstrative pronoun *dies* 'this'. In addition to this, 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 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 and included the adverbs *as, so* and *likewise*, because these adverbs were found in the first annotation round to often serve 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. Among them, 203 (DE_o) and 297 (EN_o) turned out to be abstract anaphors.

For further details of the annotation process and the annotated features, the reader is referred to Dipper et al. (2011).

3.2 Annotating abstract NPs

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

For English, abstract nouns such as *report, arrangement, fact*, etc. have been selected. The list of nouns is highly inspired by the *label nouns* defined by Francis (1994) and comprises 211 abstract nouns.[4] Table 1 provides some examples. In total, 132 instances of these nouns (in singular and plural form) occur in $EN_o$of the Anaphora Corpus.[5]

We chose the most common translations of the English label nouns to create a list of German label nouns[6] and excluded non-abstract translations. This resulted in 1–10 German translations per English noun, with an average of 3.6 translations per English noun. Some translation examples are provided in Table 1. The high number of German

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

[4] XXX

[5] EN_o: 132 instances of 45 different label noun types.

[6] Based on LEO, http://www.leo.org/.

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: English label nouns and their German translations

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, so that depending on the context, they can be translated with various German abstract nouns.

Table 1 also shows that our method yields multiple English translations for German label nouns, too. For example, *Gegenstand* 'object' can be translated as *subject* or *topic*. The final list consists of 452 types of German label nouns. From these, 134 (inflected) instances occurred in the German Anaphora Corpus DE_o.[7] Of course, not all of them were true instances of abstract anaphors (see below).

In a preprocessing 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 in general were 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 is important because some label nouns can be ambiguous between an abstract and a non-abstract reading. For example, *area* can also refer to an actual geographic area, and *report* can refer to a copy of a report. This way, we ended up with 130 English and 117 German abstract NPs for further manual annotation.[8]

Annotators were next asked to align the original noun chunk with its translation. After this, both the original label noun and the corresponding material in the translation were annotated for category, function and position.[9] 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 have been used. In both cases, we started out with a well-defined set of markables, although of course the set of markables for pronominals is considerably smaller than the set of label nouns. In both cases, we consider how the markables have been translated and whether we can induce new markables for a next annotation

---

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

[8] This shows that our preselection is highly successful in the case of abstract NPs. In contrast, occurrences of pronominal anaphors *this, that, it*, and *das* 'that' and *es* 'it' in German very 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 are abstract *per se*. In case of doubt, annotators did a quick check of the previous context to determine abstractness of the noun.

[9] Admissible values are:
– 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'.

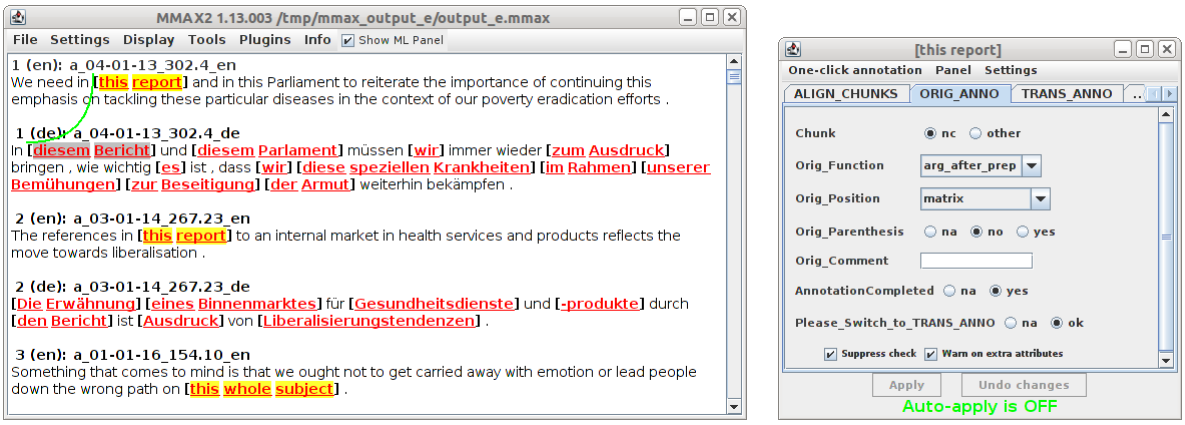

Figure 2: MMAX2 annotation windows: The left panel shows English "alignment units" along with their German translations. Noun chunks with label nouns, which have to be annotated by the annotators, are highlighted in yellow. Translation candidates are marked in red. In the first "alignment unit", the anaphoric abstract noun chunk 'this report' has been aligned to its German equivalent 'diesem Bericht'. The right 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).

round. We believe that this kind of bootstrapping approach allows for a faster and more efficient way of extracting anaphors in both languages in comparison to going through contiguous text without predefined markables. Working without predefined markables would also bear the risk that annotators disagree on the set of types they consider and even more so on the markables they annotate.

4 Quantitative investigations

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

An obvious advantage of using parallel texts for cross-linguistic investigations is that the aligned units convey the same meaning and allow for direct comparison of how this meaning is expressed linguistically in the two languages. This kind of cross-linguistic use of parallel texts also has its limitations as numerous works in translations studies have shown. For our purposes, we roughly summarize these as:

(i) The problem of *translation shifts*, cf. Vinay and Darbelnet (1958/1995); Dorr (1994), which 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 protocols of parliament debates and their translations) and cultural differences (which background knowledge of the hearers can be assumed) (Klaudy, 2008).

(ii) Impacts of the translation process which can affect the characteristics of the translated texts in different ways. There are two subtypes that are particularly relevant for us, the *shining through* of source language preferences if the translation is too faithful to the source text, cf. Teich (2003), and the tendency of translated texts to

be more *explicit* than their sources (Vinay and Darbelnet, 1958/1995; Blum-Kulka, 1986).[10] Both characteristics might directly affect the way anaphoric links are expressed such that translated texts turn out looking 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), those in (ii) in differences between original and translated texts ($comparable_{trans}$ corpora). Such differences – even if only in form and not in meaning – pose problems for approaches that target automatic resolution of anaphora.

Having outlined the specific characteristics of translated texts, we pursue a two-step approach. First, we compare the expression of abstract anaphors in the aligned units of the *parallel* resources. Second, we control our results — if possible — on the $comparable_{trans}$ part of the corpus. This means, in more detail:

Step 1: We first look at parallel (translated) texts. A naïve assumption would be that in aligned units of parallel texts, abstract anaphors are realized in the same way in both languages (e.g. with the same category and function).

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

If we find differences between the parallel texts, e.g., a transposition[11] as described in (a), there are two possible explanations: the differences are due to (i) language-specific preferences or (ii) effects of the translation process.

To test which of the explanations apply, there are different ways to pursue:

Step 2: We next check whether the tendencies also turn up in the reverse translation direction (b).

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

If this is the case, we probably observe an effect of the translation process. If the tendencies only show up in one translation direction, it is a language-specific effect.

Moreover, we can check whether the tendencies also turn up in the reverse direction of the *transposition* (c).

---

[10] Vinay and Darbelnet (1958/1995, p. 342) are the first to define the concept of *explicitation*, as "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".
The *explicitation hypothesis* by Blum-Kulka (1986): "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, p. 19) (both citations after Klaudy (2008)).
For a recent survey and critical assessment of the explicitation hypothesis, see Becher (2011, Ch. 2)).

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

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

If this is the case, the transpositions at hand seem to occur at random, and no general "rule" can be deduced from the observations.

Step 3: In addition, we can check the ratios that hold in a $comparable_{trans}$ corpus (e.g., compare the numbers of pronouns and NPs in DE_o and DE_t on the one hand, an in EN_o and EN_t on the other hand). If we observe differences between original and translated texts for both German and English, this hints at an effect of the translation process. If differences show up in one language only, it is a language-specific effect.

Steps 1–3 are applied in the following sections, with the aim to shed light on the *linguistic similarity* of abstract anaphors in German and English on the one hand, and in original and translated language on the other hand.

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

4.1 Lexical choice

Pronominal abstract anaphors We first focus on the different lexical realizations of abstract anaphors in the original and the translated texts, and compare 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 to the corpus scheme in Figure 1). The equivalents are the most frequent ones *in general*, and do not align directly to the original pronouns on the left side.

The table shows that the lexical choices lead to distributions in the translated corpora that correspond to the ones in their $comparable_{trans}$ counterparts: the top rankings are equivalent in both $comparable_{trans}$ pairs. For the German corpora, *das, dies, es* are top-ranked, with *wie* 'as' intervening in DE_t, which was not part of the original markable set and can, thus, not 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*) had not been considered in the first annotation round, as just mentioned. A remarkable deviation is the relative overuse of *dies* 'this' in DE_t in comparison to DE_o if we only take occurrences of *das, dies, es* into account.[12] This might be an example of *shining through* of the highly frequent English *this* in EN_o.

Table 3 provides a detailed view on 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. Furthermore, the most frequent equivalence types are listed together with their absolute frequencies in the translated text.

Comparing the anaphors with their translation equivalences in Table 3 shows that in almost all cases it is the literal translation which is observed most frequently. *Das* 'that'

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

Rank	DE_o pronouns	Freq
1.	*das* 'that'	123
2.	*dies* 'this'	45
3.	*es* 'it'	35

Rank	EN_t most frequent equivalents	Freq
1.	*this*	55
2.	*that*	52
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
1.	*this*	108
2.	*that*	103
3.	*as*	42
4.	*it*	36
5.	*so*	8

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

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

is most often translated as *that*, *that* as *das* and so forth. The only exception is English *so*, which translates into *dies* 'this' most often — the German pronoun that unambiguously refers to abstract objects.[13]

Abstract anaphors with demonstrative label nouns An overview of the most frequent label nouns occurring in the Anaphora Corpus corpus is provided in Table 4. As before, the nouns are the most frequent ones in these subcorpora and do not necessarily correspond to the nouns on the other side.

Looking at individual translation pairs the same tendency of literal translation preference is confirmed as 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*), *Frage/Fragen* (four types: *area* , *issue*, *situation*, *questions*).

The ten top-frequent types listed in Table 4 account for 59% of all instances in the original corpora, and for a considerably smaller proportion of 46% in the translated corpora.[14] This could be an effect of style in the translations, i.e., translators would tend to show more diversity than the original authors.

Comparing the rankings in Table 4, the *parallel* ones (displayed left to right) are

[13] 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.

[14] The precise figures for the coverages are: DE_o: 56%, EN_t: 44%, EN_o: 62%, DE_t: 48%.

DE original		EN translations (pronominals etc.)		
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 (pronominals etc.)		
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

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

Rank	DE_o label nouns	Freq
1.	*Bericht* 'report'	13
2.	*Richtlinie* 'directive'	12
3.	*Thema* 'issue'	10
4.	*Prozess* 'process'	6
5.	*Frage* 'question/issue'	5
	Punkt 'point'	5
6.	*Debatte* 'debate'	4
	Fragen 'questions/issues'	4
	Zusammenhang 'context'	4
7.	*Ergebnis* 'result'	3

Rank	EN_t label nouns	Freq
1.	*report*	13
2.	*directive*	10
3.	*issue*	7
4.	*process*	5
5.	*debate*	4
6.	*area*	3
	questions	3
	subject	3
7.	*basis*	2
	connection	2

Rank	DE_o label nouns	Freq
1.	*report*	19
2.	*proposal*	10
3.	*area*	9
4.	*agreement*	8
5.	*issue*	7
	point	7
6.	*context*	5
	subject	5
7.	*debate*	4
	problem	4

Rank	EN_t label nouns	Freq
1.	*Bericht* 'report'	15
2.	*Thema* 'issue'	8
3.	*Vorschlag* 'proposal'	7
4.	*Bereich* 'area'	6
5.	*Fall* 'case'	5
	Punkt 'point'	5
6.	*Angelegenheit* 'issue'	4
	Berichts 'report' (genitive)	4
	Gebiet 'area'	4
	Problem 'problem'	4

Table 4: Frequency rankings of the most common label nouns

more similar to each other than the $comparable_{trans}$ ones (displayed in the diagonals).[15] 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.

Usage preferences of selected nouns In addition to the comparable corpora that are part of the Anaphora Corpus, the entire Europarl corpus provides us with a huge amount of comparable data. In this section, we illustrate how this data can be used to detect interesting cases that seem worth to be looked at in detail. Note that in this subsection the abbreviations 'DE_o, DE_t' etc. are additionally used to refer to the respective subcorpora of the Europarl Corpus. In all other sections of this paper, these abbreviations exclusively refer to the Anaphora Corpus.

Our starting point is that we found considerable divergencies in the frequencies of certain label nouns, comparing original with translated turns in our Anaphora Corpus. We selected all label nouns with "considerable" differences (≥ 4) between the frequencies of original vs. translated turns, see Table 5. The columns labeled 'Anaphora Corpus' list the respective figures. A negative number in column 'Diff' indicates that the label noun occurs more often in the translated turns. Table 5 shows that, e.g., the noun *Angelegenheit* 'issue' (ranked last in the top table) never occurs in German original turns but 4 times in translations from English turns (i.e. a difference of 4 occurrences). In contrast, the noun *report* (see the bottom table) occurs considerably more often in original English turns (19 times) than in translated turns (13 times).

Similarly, the nouns *Bereich* 'area' or *directive* (marked by '*' in the table) were only annotated in translated turns. The reasons here are different, though: *Bereich* and *directive* had not been included in our original set of label nouns, and, hence, their occurrences were not pre-marked and annotated in the MMAX2 files. However, they turned up quite often as translation equivalents in the annotated translations. In the next round of annotation, 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* had not been included in the first annotation round also had an impact on the frequencies of its DE_t translation *Richtlinie* 'directive' (ranked first), which never turned up in German translations, for this reason. The same holds for the frequencies of the EN_t noun *area*, whose direct DE_o counterpart *Bereich* had not been annotated in original texts.

For each of these label nouns, we calculated its frequencies in *all* original and translated turns of the Europarl Corpus (release v3).[16] It turned out that these frequencies differ significantly with all nouns except *Fall* 'case' in German, and *directive* and *proposal* in English.[17]

In general, preferences for specific label nouns show up more often in the translations than in the original versions. This can be seen from the four last columns in the tables: they list the relative frequencies of the label nouns in original vs. translated turns (multiplied by 1,000), and the ratio of these frequencies. For instance, the very

[15] Some of the differences are artificial ones, related to the selection of label nouns that were pre-marked 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.

[16] Only translations from original turns in German and English have been taken into account.

[17] 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: $\chi^2 > 25, df = 1, p < .001$.

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

Table 5: Label nouns with differences ≥ 4 between the frequencies of original vs. translated turns. '#' indicates absolute frequencies (as occurring in the annotated corpora), 'Diff' is the difference between the two frequencies. 'Freq' are 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; noticeable figures are printed in boldface. (For nouns marked by '*', see the remarks in the text.)*

first noun is *Richtlinie* 'directive', which occurred with relative frequencies of 2.656 in original turns, and of 3.282 in translated turns. That is, the noun occurs more often in translated turns. This is reflected by the fact that the proportion DE_o/DE_t is below 1 and, consequently, $DE_t/DE_o > 1$. The last two columns show that in 6 times (out of 9) in the German data, $DE_t/DE_o > 0$, and in 4 times (out of 7) in the English data, $EN_t/EN_o > 0$.

A strikingly large frequency difference can be observed with the German noun *Angelegenheit* 'issue', which occurs 4.8 times more often in the translated turns of the Europarl Corpus; the second rank is taken by *Prozess* 'process', which occurs 1.6 times more often. Conversely, the nouns *Debatte* 'debate' and *Bereich* 'area' are the "top nouns" that occur more often in the original turns, namely roughly 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, which showed considerable frequency divergences both in the Anaphora Corpus and in the Europarl Corpus (indicated by figures printed in boldface in Table 5), were subject to further investigations.

***Angelegenheit* 'issue'**: The striking frequency differences that occur with *Angelegenheit* 'issue' might be attributed to the fact that it seems to be used as a kind of "dummy" translation for English nouns that are highly unspecific, such as *issue, matter, matter of concern*. Ex. (2) shows such an example.[18]

(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 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). Hence, our data do not allow for any tentative conclusion that would explain the observed frequency differences.

***Debatte* 'debate'** occurs more often in original German turns (no occurrence in DE_t in the Anaphora Corpus). A highly-tentative explanation could be 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 'Parliament debates', the negative connotation is absent, the meaning being 'discussion, debate'. Still, translators could avoid 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 turned up in the translations (see Table 5) are translations of *area* (5x) and *question* (1x). In an extra step, we looked up all occurrences of *Bereich* in DE_o (which had not been annotated beforehand): there are six instances, which are translated in six different ways, e.g. by

[18] 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.

area, sphere, cf. Ex. (3).[19] This means, 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, once, 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: Finally, the noun *report* occurs extremely often in the Anaphora Corpus, both in the original and translated versions (and, similarly, 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 under discussion, see Ex. (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 xxx ich hab probiert, die Charts voranzustellen und dann erst die Table zu nehmen, aber dann muss man die Section komplett neu schreiben... (mein Versuch findet sich in der Datei quant_cat.tex.versuch, habe aber irgendwann die Umarbeitung abgebrochen)

In addition to the lexical choices, we investigate grammatical properties of the anaphors. We evaluate whether pronouns are translated by pronouns — as expected according to our initial "naïve assumption" (see Step 1 in Sec. 4) — or by some other category (e.g. full NP, adverbial or clause). This investigation is motivated by findings about cross-linguistic differences, e.g., between Danish and Italian, and Spanish and Catalan, respectively (cf. Recasens (2008); Navarretta (2008); Navarretta and Olsen (2008)).

Assuming equivalence between original text and 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 such an 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.

	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)

Table 6: Pronouns: categorial transposition types

Other target categories of translated pronouns include NPs, cf. Ex. 5, and adverbials like *so, likewise* — which were then added to the English markable set.[20]

[19] The translator's choice fits nicely in the current context.
[20] DE_{lit} provides a literate translation of the German sentence.

(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 follow.* *(ep-00-10-03/15)*

Looking at the corpus EN_o–to–DE_t, the results are similar (Table 6, second row). The proportional distributions between DE_o–to–EN_t and EN_o–to–DE_t do not differ significantly.[21]

The bar plots in Figure 3 summarize the relative frequencies of grammatical categories in the Anaphora Corpus. The top chart displays the figures for pronominal anaphors in the source language.

We see 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 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 attributed to the translation process (and could be an instance of the explicitation hypothesis.

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, due to the different sets of markables used in the different annotation rounds.

	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)

Table 7: Label nouns: categorial transposition types

Abstract anaphors with demonstrative label nouns Another kind of counter-check can be performed by investigating original NP anaphors and their translations. If, unexpectedly, many of them were translated by pronouns, categorial transpositions from pronouns to NPs or vice versa would seem due to pure chance.

In the Anaphora Corpus, the vast majority of label noun anaphors is translated by NPs, independently of the translation direction, see Table 7.[22] Only 4.5% of the label nouns are translated as pronouns (or pronominal adverbs).

We conclude that there is a language-independent tendency that pronominal anaphors are translated as full NPs and full NP anaphors tend to be kept as full NPs in translation. This would conform to the explicitation hypothesis. Sec. 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*, *other*. The top chart in Figure 4

[21] Chi-squared test: $\chi^2 = 1.5185, df = 2, p = .468$
[22] There are no significant difference between both translation directions.

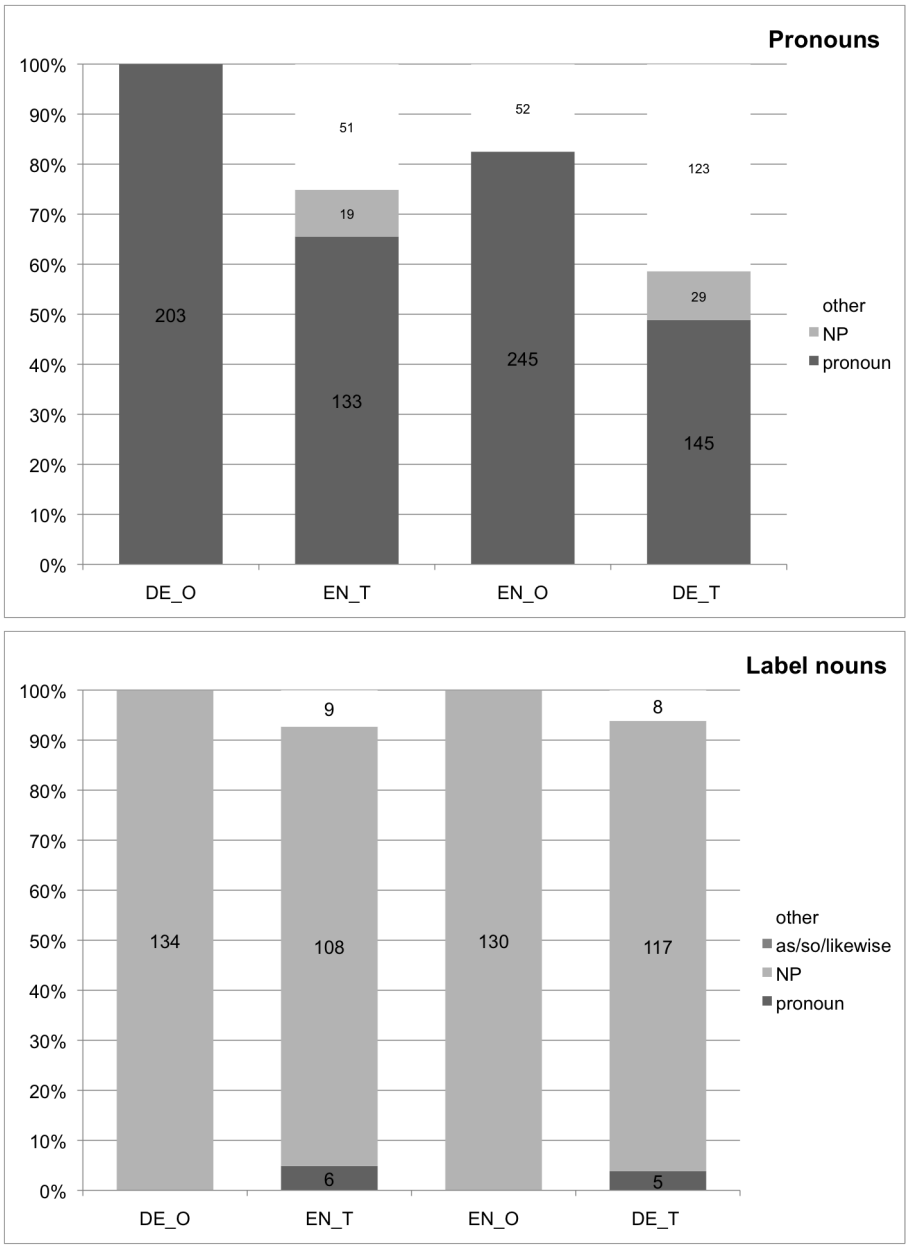

Figure 3: Relative frequencies of grammatical categories. Top chart: figures of pronominal anaphors; bottom chart: figures of the label nouns. Class 'as/so/likewise' is the markable type that has been introduced in EN_t. Class 'other' (the white parts) contains other cases with structural mismatches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

German original		English translation	
Function	**Freq**	**Function**	**Freq**
subject	147	subject	107
		object	5
		other	35
object	55	object	27
		subject	12
		other	16

English original		German translation	
Function	**Freq**	**Function**	**Freq**
subject	177	subject	114
		object	10
		other	53
object	37	object	18
		subject	5
		other	14

Table 8: Pronouns: transpositions of the functions subject and object

provides an overview of the distribution of grammatical functions in the four subcorpora. Crosslinguistic comparison of subjects and objects shows significant differences: English uses more anaphoric subjects than German.[23] 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.

This is also confirmed by looking at the translation equivalences in more detail. 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 subject anaphors usually remain subjects in the English translation, whereas German object anaphors tend to become subjects in English, too. The non-literal translation in Ex. (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)*

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

The bottom chart in Figure 4 shows the distributions of the functions observed with label nouns. The picture is similar to the ones of pronominal functions. 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)).[24] **xxx Problem: das kann man den den charts so nicht ablesen, aber die entsprechende Tabelle haben wir auch nicht im Paper. Also Formulierung evtl. einfach so lassen? xxx**

However, there are some divergencies, 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 2 out of 48 arg-after-preps in EN_o are translated as a subject in DE_t. We interpret this as a tendency of German prepositional phrases being translated as subjects in English. An example is provided in Ex. (7).

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

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

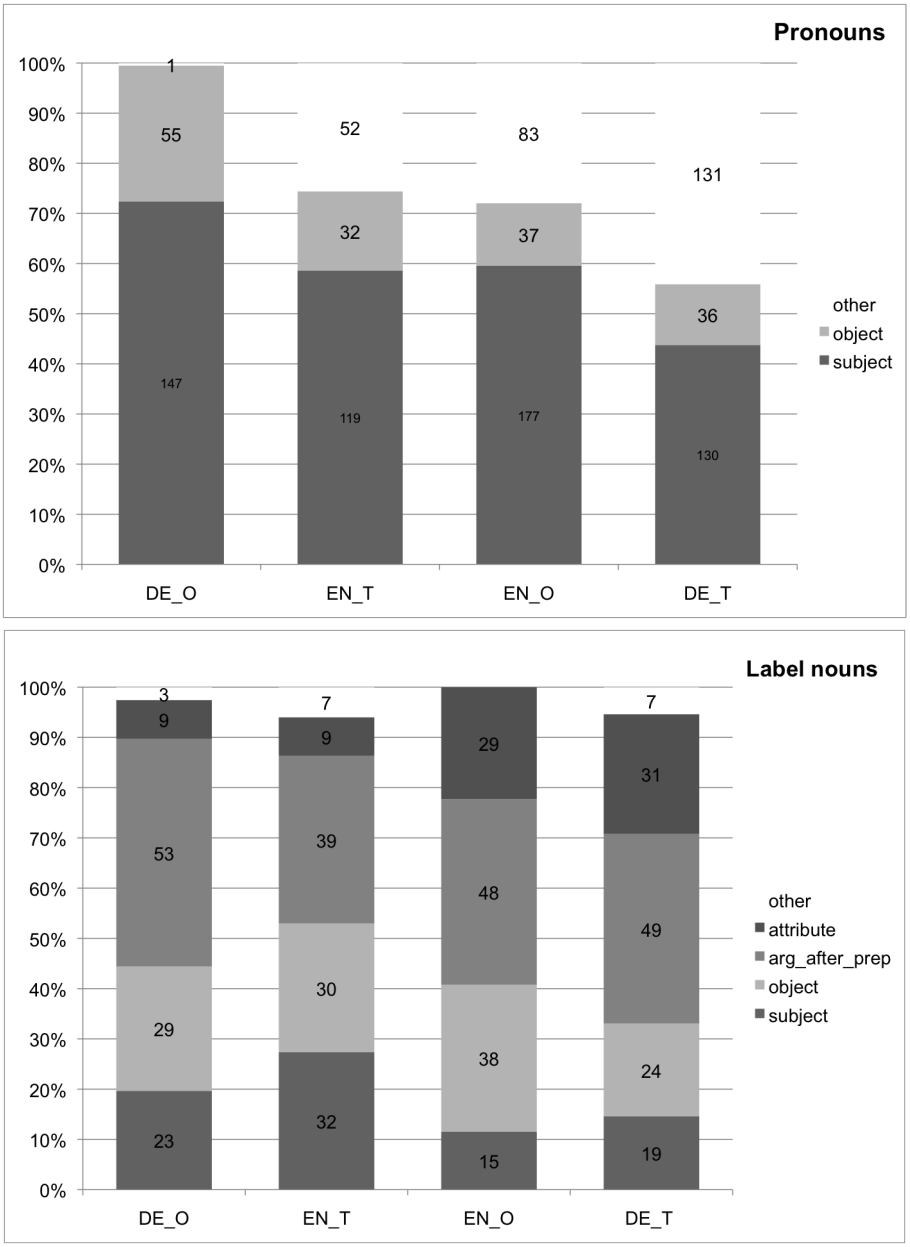

Figure 4: Relative frequencies of grammatical functions. The top chart refers to pronominal anaphors, the bottom chart to label nouns.

	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)

Table 9: Label nouns: transpositions of functions. Only pairs with mayor divergencies discussed in the text are provided.

(7) DE_o: *Sie haben die Chance, in diesem Wettbewerb wirklich sehr vieles zusammenzuführen; regionale Kulturen können grenzüberschreitend zusammenarbeiten.*
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}: *You 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.[25] If we look at the language pairs from the parallel corpora, the number of attributes do not differ significantly, 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 mentioned, German disprefers anaphoric attributes in general. Suprisingly, there are some cases where English objects are translated by German attributes (7 cases, see the third column in Table 9), but there is only one case in the other direction. This is the effect of a strong tendency for nominalisations in German. In Ex. (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)*

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

4.4 Clausal position

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

Both languages have extra means to mark or highlight constituents, such as cleft or topicalized constructions, which serve to put the constituent that is to be highlighted at the beginning of the sentence. Such special constructions are more often used in English than in German, though, probably because the prefield position in German already serves (parts of) this purpose.

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

Ex. (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.*

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

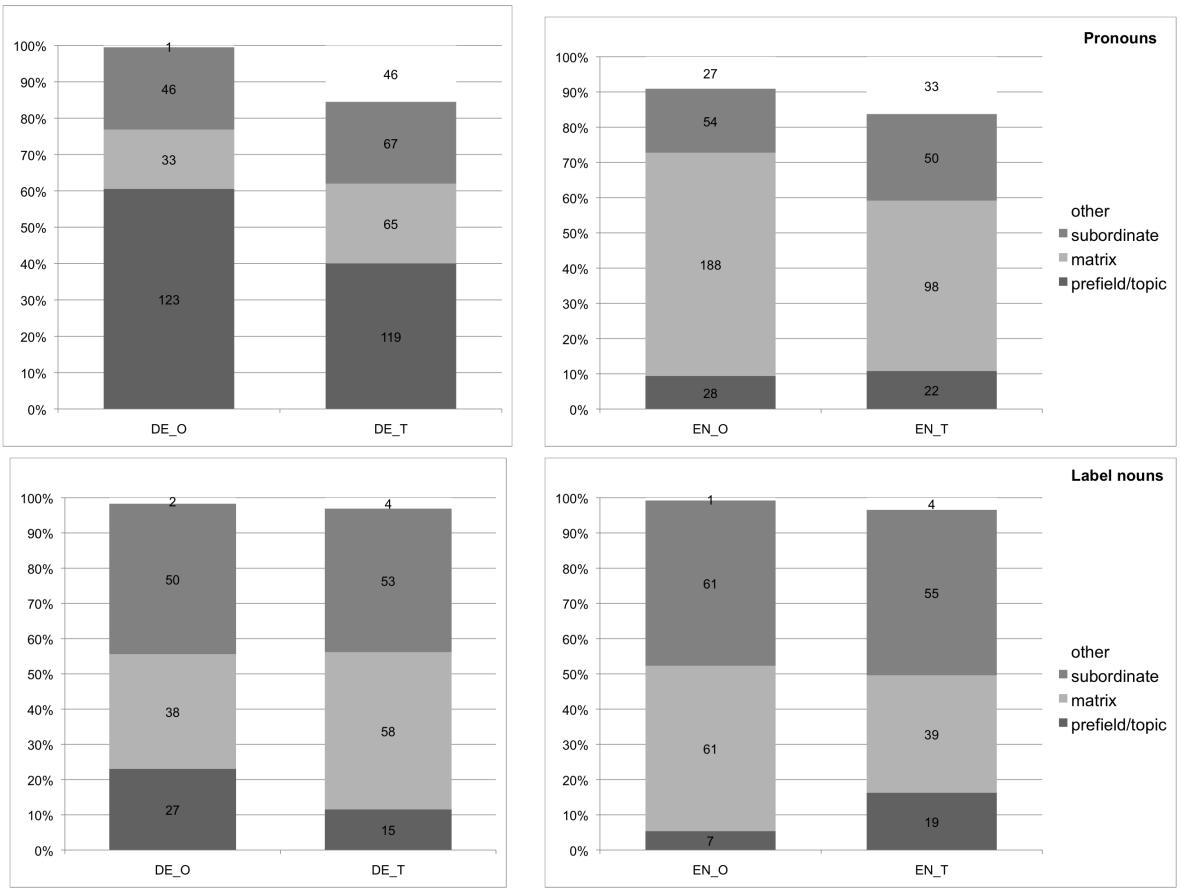

Figure 5: Relative frequencies of clausal positions. The top charts refers to pronominal anaphors, the bottom charts to label nouns. Only the pairings DE_o–DE_t and EN_o–EN_t can be compared with each other.

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: anaphors in the *matrix* clause, in a *subordinate* clause, or in some sentence-initial position, which covers topic-like constructions in English (annotated as *topic*) and the *prefield* position in German.[26]

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

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

Comparing the two German corpora with each other, we observe a significant underuse of prefield anaphors in DE_t, i.e., pronominal anaphors in DE_o occur considerably

[26] xxx Our label *matrix* might be misleading here. It is assigned to constituents in the matrix clause *except for* constituents in the topic or prefield position.xxx

more often in the prefield and less frequently in the matrix clause.[27] That is, translated texts do not follow our hypothesis to the same extent as the original texts.

A different effect is observed in the English corpora: EN_t shows a significant underuse of anaphors in a matrix position, which is outweighed by an overuse of anaphors in subordinates.[28] Anaphors in topic positions are very rare, contradicting our (simplifying) 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.[29] For English, we observe a significant overuse of anaphors in topic constructions in EN_t.[30]

To be able to relate these observations to *shining through* effects, the annotated concepts (topic, prefield) would first have to be adjusted 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 different properties of pronominal and label noun anaphors, such as grammatical category and grammatical function. In this section, we are concerned with a range of case studies to shed 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. We call these processes *edit operations*, following common terminology in computational linguistics (Levenshtein, 1965). An obvious (and highly simple) hypothesis would be that an increase in length of translated anaphors could be an effect of explicitation.[31]

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

---

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

[28] 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: $\chi^2 = 4.6114, df = 1, p < 0.05$.

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

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

[31] Of course, there are clear cases of length differences that have to be taken out from such considerations: multi-word expressions and compounds, which are usually spelt in one word in German and in several words in English. Further counter-examples to this hypothesis are presented in the following subsections.

Edit operations often have an effect on the *specificity* of the anaphors. We refer to an expression as *more specific* than another expression if it has fewer possible interpretations. Very often, addition of material (such as adding adjectives, or expanding a pronoun to a full NP) results in higher specificity. As the discussions in the next sections show, translations both increase and decrease 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 observe both situations: adjectives being added or omitted in the translation. In the Anaphora Corpus, relevant cases only occurred 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', while there is no correspondent in the original English sentence, cf. Ex. (10). This adjective clearly serves a deictic function only, i.e., it takes up 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*). Hence, the German version of the abstract NP is in fact a very close translation of the original NP in English.

(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*, which has not been translated in the German corresponding sentences.[32] 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 Ex. (11), it is not specified that the *whole* area is involved. Hence, it would be possible to continue the clause by actually restraining 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 available 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 Ex. (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 above, omitting the marker in German does not allow for a different interpretation of the respective NP. Hence, we would not classify the German translation as less specific. (Of course, the German

[32] In one case, the adjective *ganz* 'whole' has been added in the translation.

translation is missing the contribution by 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 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 all three examples (10)–(12), we see that only one type of adjective actually has 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 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 Ex. (13). The original English noun, *issue*, is highly generic, i.e., if one does not know the context, *issue* could receive a large set of possible interpretations. On the other hand, 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 this expansion . . .*
(ep-01-06-13/8)

Similar, though probably less clear examples, can be seen in Ex. (14) and (15). In Ex. (14), the English original noun *message* is less specific than the German translation *Zusage* 'assurance'. Without context, the English noun *message* could refer, e.g., to an assurance or a denial. The *denial* reading is obviously not available for 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)

In a similar way, the German translation *Zwecke* 'purposes' is more specific than the original English noun *way*, as shown in Ex. (15). *Spending money in that way* could refer, e.g., to spending money for specific purposes, or to spending money during a certain amount of time, etc. In contrast, the German noun *Zwecke* only allows the first reading.

(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 needs to be pointed out that, although most of the translated nouns are more specific than the original nouns, rare examples of the other direction also exist. For example, Ex. (16) involves *request* as the original English noun. The German translation is *Fall* 'case', which is clearly less specific than the English original (but takes up a prior mention of the word 'case').

(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

Considering the lexical semantics of nouns can serve to find translation examples in which specificity differs between original and translated texts. However, it is of course not enough to just 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 Ex. (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 arise when the sentence structure is changed considerably from the original to the translated sentence. In Ex. (18), the German original NP, *diese Strategie* 'this strategy', is less specific than the translated NP *the Lisbon strategy*. However, the English translation does actually not provide more information than the original German sentence: The German NP *diese Strategie* refers back to the antecedent *Lissabon-Strategie* (put in italics in the example). In the English translation, the sentence structure has been changed so that the NP at hand 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}: *I am grateful for the Kok report; what we now urgently need is an orientation for* ***the Lisbon strategy****, as this strategy is right.* *(ep-04-11-17/38)*

The discussion shows that we have to 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, analyzing statistical counts and noun pairs as a first step can serve to detect note-worthy examples.

5.4 Pronouns vs. NPs

As was discussed in Sec. 4.2, pronouns are quite 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 more detail.

For example, in Ex. (19), 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 given up in the German translation. Here it is specified explicitly that the speaker does not support a *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 follow.* *(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*, see Ex. (20). The pronominal anaphor could refer, e.g., to a development, a reading which is not easily available for the English corresponding expression *this threat*. In the German sentence, however, the verb *abwenden* 'avert' provides important clues and constrains the set of possible referents to negatively connotated entities.

(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}: *This 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 are taken from a wide range of sentences in which an original pronominal anaphor is translated with a more specific full NP. When considering the other direction, i.e. from original abstract demonstrative NPs to translated pronouns, only some rare examples can be found. Ex. (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 rather small range of readings for the pronoun: *that* could refer, e.g., to a judgement, assessment, opinion, or the like—highly similar concepts. Due to 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)*

Some very rare examples exist in which the translated sentence is indeed less specific than its original counterpart. In Ex. (22), *dieser Effekt* 'this effect' in German corresponds to the English pronoun *this*. English *this* could refer to a development or a threat which is exacerbated, but the German full NP does not allow for 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)*

Of course, taking prior context into account, the discourse model that speakers and hearers have built up so far could provide very clear constraints for the reference of *this*, so that no further specifications (such as using the noun *effect*) might be necessary. — The issue that is of interest to us is that in most cases where the original contribution makes use of 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 more detail, it is probably (often) done to avoid misinterpretation, and the translator faces the very same situation in the target language (for languages as similar as German and English).

Hence, whenever the translator deviates from the original version in such ways, this could indicate interesting phenomena to look at in detail, 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 were pre-marked that contained a demonstrative determiner. Hence, we 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 Sec. 5.1): they use adjectives, such as *vorliegend* 'present, at hand' and *last-mentioned*, 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 for English original sentences and their German translations. The examples can also involve some minor changes in the overall structure of the sentence. In Ex. (23) (= Ex. 4), the English speaker thanks the rapporteur for producing the report. In the German translation, *producing* is not translated but 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 can observe different situations. In a range of sentences, specificity of the anaphoric noun seems considerably reduced in the translation. In most of these examples, *such (a)* serves as a substitute determiner, see Ex. (24). Similarly to canonical demonstratives, *such* has a deictic component but serves to point 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, Ex. (25).

(24) EN_o: *The Commission, however, intends 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(not) yet achieve no full agreement.* *(ep-97-04-08/304)*

Finally, in Ex. (13) (repeated here as Ex. (26)), the abstract label noun is translated by a lexically more specific noun. Due to this, the space of possible references is narrowed and, hence, use of the demonstrative determiner seems superfluous (see the discussion in Sec. 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 this expansion . . .* *(ep-01-06-13/8)*

6 Conclusion

In this paper, we have presented a bootstrapping approach to annotating pronominal and label noun anaphors. Based on the annotated data, we investigated selected properties of the anaphors in more detail. Before summarizing some interesting findings in our data, we want to emphasize that all our findings have to be understood as holding only for the particular type of language represented in the Europarl corpus — spoken and translated parliament 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 shown to what extent they generalize to other domains and text types.

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

Judging from the frequencies of the top-frequent label nouns, translators of both languages exhibit a more diverse vocabulary than the authors of the original source.

Certain label nouns occur very often in our data. This, of course, is connected with the domain of our data: parliament debates. Nevertheless, when comparing the frequencies of selected label nouns in original vs. translated turns in the entire Europarl Corpus, interesting discrepancies stand out (which are statistically significant).

Judging from our annotated data, the German noun *Angelegenheit* 'issue' seems to serve the translators as a kind of "dummy" translation. With the noun *Bereich* 'area', we observe an interesting asymmetry: when translated into English, a vast variety of English expressions is found (e.g. *area, issue, subject, sphere*), whereas German translators employ *Bereich* quasi exclusively as the translation of *area*.

Category, function, position Translations in general tend to preserve the anaphor's categories, functions and positions. Interesting differences are:

With regard to category, we observe a clear asymmetry: a considerable amount of pronouns are translated as full NPs, while the reverse is not true. Since the asymmetry shows up with both languages, we might observe an effect of the translation process. Maybe this is due to translational conventions (in the form of "do not use pronouns"). Very rarely can the opposite mapping be observed. As in the case of lexical semantics (see below), the context sometimes compensates for the loss of specificity.

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

Finally, looking at positional properties of the anaphors, both languages exhibit language-typical patterns, in the original and translated texts. Still, *shining through* effects show up here as well: DE_t underuses anaphors in the prefield position, EN_t underuses matrix anaphors as compared to subordinate anaphors.

Adjectival modifications Adjectives such as *whole* are sometimes omitted even if this can result in underspecification and different interpretations possibilities. Such omissions mainly occur in the translation direction EN_o–to–DE_t. That is, the German translation is less specific than its source.

Lexical semantics Most of the cases where the original and translated nouns differ with respect to their specificity are found with EN_o–to–DE_t translations. And usually, the German translations are 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) compensates for loss of specificity of nouns.

Transposition of demonstratives Two cases seem of interest here: First, specific demonstrative NPs are sometimes translated by *such* (or its German equivalent). Hence, the speaker does no longer refer to the specific entity under discussion but to all entities that are of the same kind. Second, the demonstrative article is sometimes translated by a definite article. In these cases, the deictic function of the demonstrative often seems to be taken over by adjectives such as *vorliegend* 'present'.

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

In the next annotation round, the translated nouns that have not yet been part of our label noun list will be included and annotated as well.

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

So far, we have only annotated and aligned pairs of turns that, in the original language, contain the pronouns *it, this, that* (and their German equivalents) or demonstrative NPs with label nouns. No such restrictions apply to the translated turn; here the annotators are free to mark arbitrary strings as the expression that represents the translation of the anaphor in the original text. As we have seen, however, translators very often stay close to the original. Hence, we cannot expect to hit on rather marked and exceptional ways to refer to abstract entities. To complement our approach, it would be useful to annotate a sample of running text, marking all kinds of abstract anaphors that turn up.

Finally, we want to profit from the fact that Europarl provides the debate protocols in many more languages, and include further languages in our studies.

7 References

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

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

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

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

Donna K. Byron. Annotation of pronouns and their antecedents: A comparison of two domains, 2003. Technical Report, University of Rochester.

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

Stefanie Dipper and Heike Zinsmeister. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC 2010 workhop on Language Resource and Language Technology Standards*, pages 54–59, Valletta, Malta, 2010.

Stefanie Dipper and Heike Zinsmeister. Annotating discourse anaphora. In *Proceedings*

of LAW III, pages 166–169, 2009.

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

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

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

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

Nancy Hedberg, Jeanette K. Gundel, and Ron Zacharski. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of DAARC-2007: 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 31–36, 2007.

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

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

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

Christoph Müller. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of ACL-07 conference*, pages 816–823, 2007. URL http://www.aclweb.org/anthology/P07-1103.

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

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

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

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

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

Ralph Weischedel et al. OntoNotes Release 4.0, with OntoNotes DB Tool v. 0.999 beta. Technical report, Raytheon BBN Technologies et al., 2010. http://www.bbn.com/NLP/OntoNotes.

Marta Recasens. Discourse deixis and coreference: Evidence from AnCora. In *Proceed-*

ings of the Second Workshop on Anaphora Resolution, pages 73–82, 2008.

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

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

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

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

Renata Vieira, Susanne Salmon-Alt, and Caroline Gasperin. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of DAARC-2002: 4th Discourse Anaphora and Anaphora Resolution Colloquium*, 2002.

Jean-Paul Vinay and Jean Darbelnet. *Comparative stylistics of French and English: A methodology for translation*. John Benjamins, Amsterdam/Philadelphia, 1958/1995.