Empty links and crossing lines: querying multi-layer annotation and alignment in parallel corpora

Oliver Čulo, Silvia Hansen-Schirra, Karin Maksymski
Johannes Gutenberg-Universität Mainz, Germersheim
culo@uni-mainz.de, hansenss@uni-mainz.de, maksymsk@uni-mainz.de

Stella Neumann
IFAAR, RWTH Aachen
neumann@anglistik.rwth-aachen.de

Translation shifts can be informative in various ways. Amongst other things, they can point to typological differences between languages or be indicators of properties of translated text like e.g. explicitation or normalisation. Detecting translation shifts in parallel corpora is thus a major task from the viewpoint of translation studies. This paper presents an analysis of translation shifts in a parallel corpus (English-German). It offers an operationalisation of queries which can exploit multi-layer annotation and alignment in order to detect various kinds of translation shifts across category boundary lines and empty alignment links. The paper furthermore discusses the shifts and links them to certain translation properties.

Keywords: parallel corpora, multi-layer annotation and alignment, corpus query, translation studies, translation shifts, translation properties

1 Introduction

In both translation studies and contrastive linguistics, multilingual corpora have recently been used to study translation phenomena, i.e. translation shifts or translation properties (as proposed by Baker 1993; 1995; Toury 1995), as well as contrastive differences between languages. One such corpus is the English-German CroCo corpus (Hansen-Schirra et al. forthcoming). The corpus contains English and German originals and their translations into German and English, respectively. It can thus be used both as a comparable and a parallel corpus, e.g. to study contrastive differences (e.g. Steiner 2008), translation phenomena (e.g. Čulo et al. 2008; Hansen-Schirra et al. 2007) or register variation (Neumann 2008). The corpus draws much of its potential from its multi-level stand-off annotation and alignment (Hansen-Schirra et al. 2006).

In this paper, we present a study based on the parallel data in the corpus, exploiting the multi-level alignment in order to detect translation phenomena. We intend to show how the annotation and alignment of linguistic structures can help detect translation phenomena and provide data for their deeper analysis and interpretation. We will demonstrate this by presenting data on and interpretations of so-called 'empty links' and 'crossing lines', two phenomena which we will characterize in section 2.

In section 3, we will briefly outline the technical background of this study, i.e. the structure of the corpus, the corpus API and how the corpus was queried. Section 4 then discusses the results and possible interpretations of the queries with respect to certain grammatical levels. Section 5 gives an overview of possible future directions for taking this study further.

2 Empty links and crossing lines

Approaching translation from a naive perspective, all translation units should match correspondent units in the source texts, both in semantics and in grammatical analysis (Padó 2007). This is, of course, unrealistic, not only because languages diverge, but also because translators make individual decisions. Very broadly speaking, originals and their translations therefore diverge in two respects. Units in the target text may not have matches in the source text and vice versa; thus no connection can be drawn and we speak of *empty links*. Units which do have a counterpart with which they are aligned may be embedded in higher units which are not aligned, resulting in *crossing lines*. This is, for instance, the case when a word is embedded in a chunk with the subject function in one language, and its counterpart in a chunk with the object function.[1] These two concepts are related, on the one hand, to concepts used in formal syntax and semantics (like null elements and discontinuous constituency types in LFG (Bresnan & Kaplan 1982) or HPSG (Pollard & Sag 1994). On the other hand, they are in the tradition of well-known concepts in translation studies such as one to zero correspondence and translation shifts (Koller 2001, Vinay & Darbelnet 1958, Catford 1965, Newmark 1988, van Leuven-Zwart 1989, Cyrus 2006 etc.).

We analyze for instance stretches of text contained in one sentence in the source text but spread over two sentences in the target text, as this probably has implications for the overall information contained in the target text. We would thus pose a query retrieving all instances where the alignment of the lower level is not parallel to the higher level alignment but points into another higher level unit. In the example below the German source sequence (1a) as well as the English target sequence (1b) both consist of three sentences, which are aligned to each other.

(1) a. *Aus dem Augenwinkel sah ich, wie eine Schwester dem Bettnachbarn das Nachthemd wechselte. Sie rieb den Rücken mit Franzbranntwein ein und massierte den etwas jüngeren Mann, dessen Adern am ganzen Körper bläulich hervortraten. Ihre Hände ließen ihn leise*

[1] The term *crossing line* does not refer to crossing edges in the alignment. The image behind the term is rather that some unit which is embedded in another unit does not follow the alignment path (if there is any) of the higher unit it is embedded in, but "crosses a line" and enters the realm of another unit.

wimmern. (GO_FICTION_002)

b. *Out of the corner of my eye I watched a nurse change his neighbor's nightshirt and rub his back with alcoholic liniment. She massaged the slightly younger man, whose veins stood out blue all over his body. He whimpered softly under her hands. (ETrans_FICTION_002)*

In German, the first sentence is subdivided into two clauses, the second one into three. The first English target sentence contains three clauses and the second sentence two. The third sentences in both versions are co-extensive with the clause contained in them. We can see in the example that the German clause 3 (*Sie rieb den Rücken mit Franzbranntwein ein*) in sentence 2 is part of the coordinated raising construction (*…and rub his back with alcoholic liniment*) in the English sentence 1. The alignment of this clause points out of the aligned first sentence, thus constituting a crossing line.

The third sentence also contains a crossing line, this time at the levels of grammatical functions and word alignment: The words *Ihre Hände* in the German subject are aligned with the words *her hands* in the English adverbial. However, this sentence is particularly interesting in view of empty links as shown in Hansen-Schirra et al. (2006). The empty links are marked by a black dot in Figure 1.

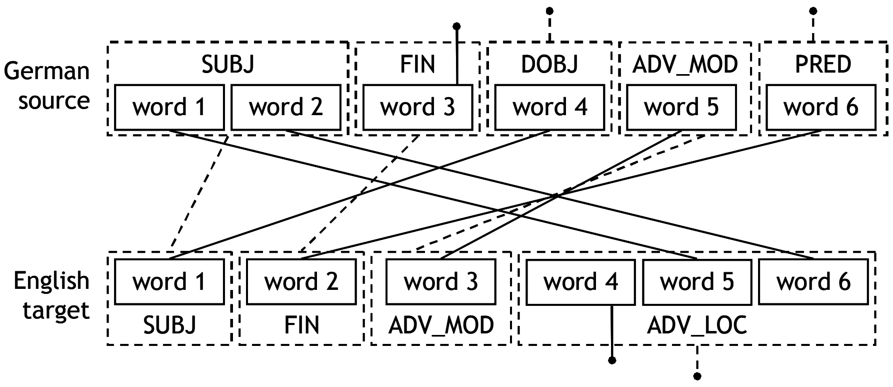

Figure 1: Alignment of grammatical functions and words in sentence 3

Our linguistic interpretation is based on a functional view of language. As explained in section 3, chunk alignment is based on the mapping of grammatical functions. Hence, the finite *ließen* (word 3) in the German sentence is interpreted as a semi-auxiliary and thus as the finite part of the verbal group. Therefore, *wimmern* (word 6) receives the label PRED (for predicator)[2], i.e. the non-finite part of the verb phrase, in the functional analysis. At word level, this German word is linked to word 2 (*whimpered*) in the target sentence, which is assigned FIN, i.e. the finite verb in the layer of grammatical functions. As FIN exists both in the source and in the target sentences, this chunk is aligned. The German functional unit PRED does not have an equivalent in the target text and receives an empty link. Consequently, word 3 in the source sentence (*ließen*) also receives an empty link. This mismatch will be

[2] We are assuming in our annotation an analysis of the verb phrase into *Finite* and *Predicator* following Halliday 1985:78ff

interpreted in view of our translation-oriented research in section 4. In the following subsection we will see how these two phenomena can be retrieved automatically.

3 Building and querying the corpus

3.1 Corpus construction

The CroCo corpus consists of English originals (EO), their German translations (GTrans) as well as German originals (GO) and their English translations (ETrans). Both translation directions are represented in eight registers, with at least 10 texts totaling 31,250 words per register. Altogether, the CroCo Corpus comprises approximately one million words. Additionally, register-neutral reference corpora are included for German and English, comprising 2,000 word samples from 17 registers.

The corpus thus consists of both a comparable and a parallel part. The registers are political essays (ESSAY), fictional texts (FICTION), instruction manuals (INSTR), popular-scientific texts (POPSCI), corporate communication (SHARE), prepared speeches (SPEECH), tourism leaflets (TOU) and websites (WEB), and were selected because of their relevance for the investigation of translation properties in the language pair English-German. All texts are annotated with

- meta information following the TEI standard (Sperberg-McQueen & Burnard 1994, Burnard & Bauman 2007) including a brief register analysis that allows additional filter options,
- part-of-speech information using the TnT tagger (Brants 2000) with the STTS tag set for German (Schiller et al. 1999) and the Susanne tag set for English (Sampson 1995),
- morphology using MPRO (Maas et al. 2009) which operates on both languages,
- grammatical functions of the highest nodes in the sentence, manually annotated with MMAX2 (Müller & Strube 2006).

Furthermore, all texts are aligned on

- word level using GIZA++ (Och & Ney 2003),
- chunk level (indirectly) by mapping the grammatical functions onto each other,
- clause level (manually) again using MMAX2,
- sentence level using the WinAlign component of the Trados Translator's Workbench (Heyn 1996) with additional manual correction.

The CroCo data are stored in an XML file format based on the XCES[3], a multi-layer stand-off markup format. The CroCoXML format is described in detail in Hansen-Schirra et al. (2006), Hansen-Schirra et al. (to appear).

[3] http://www.xces.org, last visited 3 December 2009

3.2 CroCoAPI

Processing of corpus data – annotation, querying and the like – happens on various linguistic levels and usually involves different applications suited to one particular task (e.g. PoS tagging). Thus, the necessity often arises to convert corpus data into a certain, tool-dependent input format, and then back from the output format to the corpus format. Ideally, a corpus is embedded in some sort of larger framework which manages the data streams or even already comprises a number of applications working in some sort of processing pipeline.

In the case of the CroCo corpus, we created our own *application programming interface* (API) to manage ever more complex queries, including queries operating on multiple annotation and alignment layers, and to apply Java-based annotation tools to the corpus data. The prerequisites for the API were:

- quick integration,
- support of complex queries, also on alignment,
- no complex conversion into other formats required, and
- possibly, integration of multiple formats.

The CroCoAPI presented here is a Java API which includes a light-weight, format-independent data structure that serves as communication interface to other applications. The following paragraphs describe the basic design of the API. (Java classes and API layers are typeset in capitals.)

The API is made up of three parts. On top, there is the actual interface CROCOIF, the control methods of which present the basic read/write and iteration calls for the CroCo corpus data. Under the hood, a package called CORETOOL is used to represent linguistic structures in stratified layers, and the parallel structures (e.g. aligned words, sentences, etc.) as sets of pairs. As an intermediate level, there is the CROCOXMLIO package, which handles the XCES-based CroCo data format. The CROCOIF communicates with CROCOXMLIO using the CORETOOL data structures.

Fundamental within the API is the notion of TEXT. The CORPUS is a collection of TEXTS, and each TEXT contains a thematically coherent set of linguistic structures. The list of available TEXTS can be generated for the whole corpus or per register, as singletons or as pairs of original and translation.

In the multi-layer layout of CroCo, linguistic units like sentences or chunks are defined on the basis of lists of tokens. There is no explicit information about the syntactic hierarchies, e.g. whether a certain chunk belongs to a certain sentence. However, for a number of applications it is helpful or even required to convert this representation into a stratificational structure as is provided by CORETOOL.

The CORETOOL data structure was designed to be a format-neutral representation of the linguistic structures generally found in a corpus. The data structure is used within the CroCoAPI to communicate between the interface and the input-output

(IO) level; it can, furthermore, be used as data connector to applications like in the case of the lexical chainer embedded in DKPro (Gurevych et al. 2007, see below). In general, one could enhance the CroCo corpus with various data formats and integrate these with CoReTool; this would only need additional read-/write-methods for handling the different data formats. This stratificational approach is a major difference between the CroCoAPI and other APIs like TigerAPI (Özgür 2007), where programming data structures and underlying data format are more closely linked and a conversion to TigerXML is necessary for a corpus before using it with any aspects of the TigerAPI.

CoReTool represents the linguistic data in stratified layers, following classical linguistic strata. This differs from the representation in CroCoIF, where all linguistic structures such as sentences or chunks are defined on the basis of tokens.

A Corpus is made up of an ordered collection of Texts, which again is made up of an ordered collection of Sentences, which again is made up of an ordered collection of Tokens. This structure is, so to speak, the backbone of CoReTool and the minimum of data that we expect in a corpus. In addition, a Corpus can be divided into Registers which also relate to collections of Texts (from the Corpus). Likewise, a Sentence can contain Clauses or Chunks which relate to the Tokens of the Sentence. For each of these subunits of a text (including Tokens), it is possible to have aligned counterparts. Every single alignment is represented as a pair; so if unit *U* is aligned with *U'* and *U''*, there will be two pairs <*U*,*U'*> and <*U*,*U''*>.

The CoReTool Java package uses simple data structures like ordered lists to organize the linguistics content it represents. In addition, a couple of basic methods for calculating statistics – e.g. the number of chunk types – are included. The package so far lacks a proper backend-enabled design, so that IO methods could be plugged in on demand. Also, the linguistic representation of CoReTool is currently restricted to syntactic structures.

3.3 Querying the aligned corpus

In CroCoXML, the alignment is stored in one XML file per level. Alignments between words are, for instance, represented as follows:

```
<word>
  <align xlink:href="#t3076"/>
  <align xlink:href="#t3301"/>
 </word>
 <word>
  <align xlink:href="#t3077"/>
  <align xlink:href="#undefined"/>
 </word>
```

In the pairs of words, the first entry relates to the source text word and the second to the target text word. For the word alignment, we decided to explicitly state empty links by including an element *#undefined* where no corresponding word exists for a source or target language token, which we can read off from the automatic alignment data. This is not the case for the clause or sentence alignment, which was done, or at least corrected, manually.

For the queries on empty links on word level, it would be sufficient to evaluate the XML alignment. A simple way to query for empty links would have been to query the XML annotation for pairs where one element is *#undefined*. However, the implementation results in more abstract ways to query the data. The alignment is read in from the XML files and packed into abstract data structures, representing tokens and token pairs (i.e. aligned tokens), clauses and clause pairs, etc. These abstract data structures are passed on to a query processor. This design allows both for the simple empty link queries and for the more complex crossing line queries. Also, this adheres to our aim of keeping the processing of the corpus format and the processing on linguistic structures separate.[4]

Applied to the parallel sentence from the empty link example in section 2, the empty link query returns all German original words which receive an empty link due to a missing equivalent in alignment (in this case *ließen*). The same query can also be applied to the other alignment layers: see section 4.1 for empty links at the level of grammatical functions and section 4.2 for empty links at clause level.

Querying crossing lines in the aligned source and target sentences combines the alignment on two levels, e.g. word level and the mapping of grammatical functions. Crossing lines are identified, for instance at this level, by querying for words in one grammatical function in one language which are aligned with words in a different grammatical function in the other language. An example algorithm (pseudo-code) is given in figure 2.

```
for every word_pair in word_pairs
    sl_clause : =
        get_clause(get_sl_word(word_pair))
    tl_clause : =
        get_clause(get_tl_word(word_pair))
    is_aligned?(sl_clause, tl_clause)
end
```

Figure 2: Pseudo-code for a query on crossing lines between words and clauses.

[4] Partly, the queries are realized on the format-independent CoReTool level. For the most part, however, the queries still use the proprietary CroCoXML API, because the API was still in development at the time of writing and not all levels had been sufficiently and transparently distinguished from one another.

Applying the query to example (1) returns, for instance, the German words *Ihre Hände*, which are part of the German subject and which are aligned with the English words *her hands*, which are part of the second adverbial. The query for crossing lines between words and grammatical functions is different from other queries, as there is no explicit chunk alignment. When querying for crossing lines between words and clauses, we can make use of the data from the manual clause alignment. Additionally, other alignment layers may be investigated with similar queries, e.g. crossing lines between grammatical functions and clauses.[5]

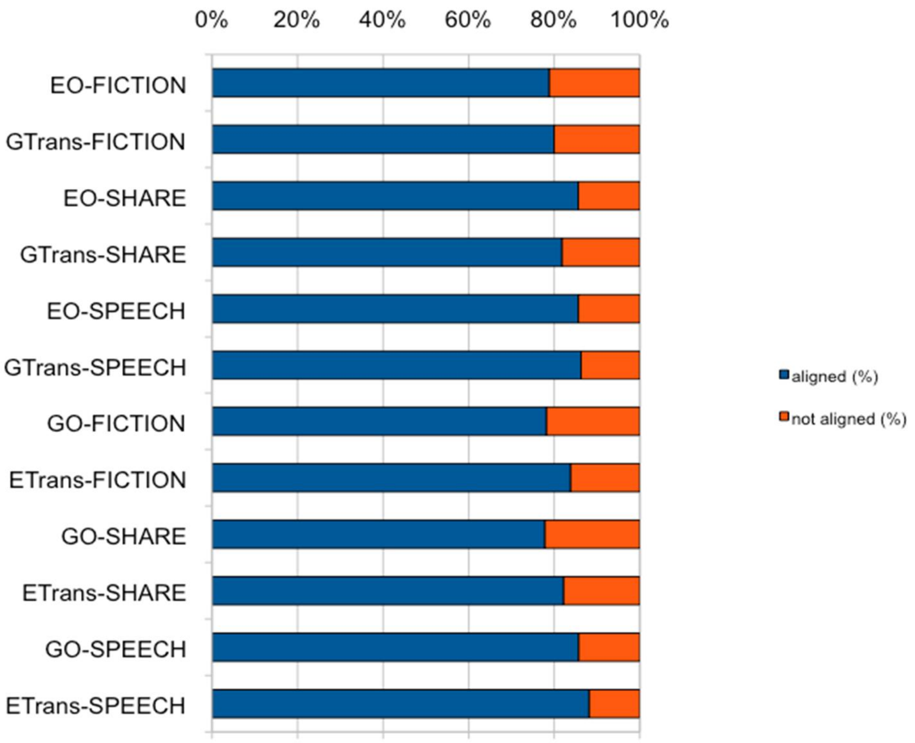

Figure 3: Statistics for alignment of grammatical function

4 Some selected phenomena

In this section, we will discuss empty links with respect to grammatical functions (subsection 4.1) and clauses (subsection 4.2) as well as crossing lines for words and grammatical functions (subsection 4.3). The three aspects were chosen because they represent a range of queries as well as translation phenomena. The discussion concentrates on the three registers FICTION, SHARE and SPEECH, which show a sufficient range of variation to detect registerial influences on translation properties.

[5] It should be noted that precision and recall of the query results can only be as precise as the word alignment provided by GIZA++ (cf. Čulo et al. 2008). This limits the validity of the query results for crossing lines and empty links on all levels involving word alignment.

4.1 Empty links at the level of grammatical functions

At the level of grammatical functions, the following tendencies in connection with empty links, i.e. non-aligned segments, can be identified. As figure 3 shows, percentages for empty links in the translation direction English-German are rather similar for originals and translations, with SHARE exhibiting a slightly higher percentage of unmapped functions for the German translations. When looking at the translations from German to English, however, there is a clear tendency for German texts to exhibit more unmapped functions than the English translations.

Tag	Explanation	EO-SHARE	GTrans-SHARE
ADV_CAUSE	causal adverbial (*therefore*)	4.00	0.83
ADV_LOC	locative (*in the house*)	3.72	2.76
ADV_MOD	modal adverbial (*with pleasure*)	4.65	12.02
ADV_TEMP	temporal adverbial (*yesterday*)	3.16	4.97
ADV_OTHER	other adverbials (*however*)	3.53	4.01
APPO	apposition (…, *which makes no sense*)	7.07	0.14
COMPL	complement (*He is a teacher*)	18.51	1.66
CONJ	sentence-initial conjunction (*but*)	7.81	12.85
DOBJ	direct object (*I hit the ball*)	16.19	11.05
FIN	finite part of the verb (*has seen*)	0.19	0.69
IOBJ	indirect object (*Tell him*)	2.51	4.97
NEG	sentence negation (*We didn't go*)	1.12	0.83
MINOR	verbless sentence (*Dear customers!*)	1.3	0.69
PART	particle (*It was just funny*)	2.79	10.91
PRED	non-finite part of verb (*has seen*)	14.6	30.11
PROBJ	prepositional object (*rely on s.o.*)	8.19	0.55
SUBJ	subject (*She is a doctor*)	0.65	0.97

Table 1: Distribution of empty links for grammatical functions (in %)

We have chosen the English-German SHARE texts for a closer look at the distribution of empty links for grammatical functions. Table 1 shows the percentage of empty links for the different grammatical functions in EO_SHARE and GTrans_SHARE. Empty links occur with different grammatical functions comparing English and German. The English originals, for example, have more empty links for appositions (APPO) and complements (COMPL), but fewer empty links for predicators (PRED) or modal adverbials (ADVmod). This means that the English original appositions and complements tend to be realized differently in the German translations. Furthermore, the German translated predicators and modal adverbials tend to have other realizations in the source language texts. These differences might be a sign of implicitation or explicitation effects (cf. Hansen-Schirra et al. 2007). They might, however, also be explained through translation shifts on the level of grammatical functions.

The following examples illustrate the observation that the frequency of empty links for appositions is higher in the English original share texts than in the German translations.

In example (2) the English apposition *a record* is an interpretation of the facts presented in this sentence. Example (3) exhibits a very similar rhetorical move in the apposition *an improvement of 2.3 turns*. In both cases, the appositions are translated by coordinated finite sentences – in the latter one even in inverse order – thus adding linguistic information by spelling out implicit information (cf. Hansen-Schirra et al. 2007 for more discussion of such phenomena). Obviously, this is one of the sources of empty links between source and target segments.

(2) a. Revenues rose 11 % to $ 112 billion, a record. (EO_SHARE_004)
b. Der weltweite Umsatz stieg um 11 % auf $ 112 Mrd. und erreichte damit eine neue Rekordhöhe. (GTrans_SHARE_004)

(3) a. Working capital turns hit an all-time high of 11.5 - an improvement of 2.3 turns. (EO_SHARE_004)
b. Die Umschlagshäufigkeit des Betriebskapitals konnte um das 2,3 fache gesteigert werden und erreichte die neue Höchstmarke von 11,5. (GTrans_SHARE_004)

The high frequency of empty links for complements may be due to registerial and typological constraints of the English SHARE texts. Example (4) shows that the English verb *name* is followed by a complement, whereas the German verb *ernannte* is followed by a prepositional object. This is, of course, an obligatory shift due to language typological differences. However, the frequent use of these constructions might be attributed to the register on the basis of a combined interpretation of verb semantics and valency. A possible explanation could then be that companies are supposed to distinguish themselves from other companies and enumerate their achievements. Example (5) again illustrates language typological differences between English and German. Whereas English uses a subject complement in the construction *We are pleased...*, the German translation is realized by the finite reflexive verb *(sich) freuen*, but no subject complement, and it is this non-mapping on the level of grammatical functions which creates the empty link here. In terms of "markedness", the original construction is typical of English, just as the translated construction is typical of German, thus explaining the number of empty links for English complements.

(4) a. Also for the second straight year, we were named "The World's Most Respected Company" by the Financial Times. (EO_SHARE_004)

b. *Ebenfalls zum zweiten Mal in Folge ernannte die Financial Times GE zum "am meisten respektierten" Unternehmen der Welt. (GTrans_SHARE_004)*

(5) a. *We are pleased to present the 2001 Annual Report of the American Institute for Contemporary German Studies (AICGS). (EO_SHARE_013)*
b. *Wir freuen uns, Ihnen den Jahresbericht 2001 des American Institute for Contemporary German Studies (AICGS) präsentieren zu können. (GTrans_SHARE_013)*

The high frequency of empty links for predicators in the German translations is due in most cases to language typological and register constraints: example (6) illustrates a shift in tense which involves using the predicator, i.e. the non-finite part of the verb phrase *geschafft*. In examples (7) and (8) the English active constructions are translated by passives in German, which include the predicators, the past participles *beschrieben* and *weiterentwickelt*. The choice of passive is motivated by the register since this German specialized register tends to favour a content-oriented style expressed by dense noun phrases as well as passivization (cf. Neumann 2008). Here, typical structures of the target language register are chosen by the translators.

(6) a. *We already have that! (EO_SHARE_004)*
b. *Das alles haben wir bereits geschafft. (GTrans_SHARE_004)*

(7) a. *In that report, we described several challenges and opportunities that we felt were going to determine the agenda of German-American relations. (EO_SHARE_013)*
b. *In diesem Bericht werden verschiedene Herausforderungen und Gelegenheiten beschrieben, die unserer Meinung nach die Beziehungen der beiden Staaten bestimmen. (GTrans_SHARE_013)*

(8) a. *It progresses with a drumbeat regularity throughout our business year - year after year.(EO_SHARE_004)*
b. *Jahr für Jahr wird das Betriebssystem mit der Regelmäßigkeit eines Paukenschlages weiterentwickelt. (GTrans_SHARE_004)*

The reasons for finding more empty links for modal adverbials in the German translations seem to be manifold: Example (9) shows an added modal adverbial in the target language text. The back-translation of the German target text reads: *Wireless networks will change the workplace fundamentally*. The English word *transform* is translated through the weaker German verb *verändern* (*change*) in combination with the modal adverb *grundlegend* (*fundamentally*). This can be interpreted as a more explicit German version of the English verbal construction.[6] Concerning the modal adverbial *persönlich* (*face-to-face*) in (10), implicit information in the source text is rendered explicit in the translation. In both cases, however, the translators probably try to emphasize relevant information, thus making the text easier or faster to understand. Example (11) illustrates a case of typologically-driven translation behavior: The English raising construction *continue to benefit* is not available in German (cf. Hawkins 1986: 75ff). Therefore, the translator chose a different lexico-grammatical realization (i.e. the addition of an adverbial), adapting the German translation to target language norms.

[6] Cf. Hansen-Schirra et al. (2007) for a discussion of explicitation vs. addition.

(9) a. *Wireless networks will transform the workplace. (EO_SHARE_005)*
b. *Drahtlose Netzwerke werden den Arbeitsplatz grundlegend verändern. (GTrans_SHARE_005)*

(10) a. *Mostly, it involves creating and distributing paper documents or telephoning and meeting with fellow employees. (EO_SHARE_005)*
b. *In den meisten Fällen erstellen und verteilen sie Papierdokumente oder telefonieren oder treffen sich persönlich mit anderen Mitarbeitern. (GTrans_SHARE_005)*

(11) a. *We continue to benefit from the strong natural gas market in North America. (EO_SHARE_002)*
b. *Wir profitieren weiterhin von einem starken Erdgasmarkt in Nordamerika. (GTrans_SHARE_002)*

In summary, empty links on the level of grammatical functions show some interesting and varied patterns. Some of the empty links may be attributed to different usage patterns, for instance in the case of English complements and German prepositional objects. Others are due to more general contrastive differences such as the (non-) availability of raising constructions in one of the languages, or different kinds of constraints on the mapping from semantic roles to grammatical functions. A more in-depth inspection of all hits for the query could provide an interesting overview of translation properties on this layer.

4.2 Empty links at clause level

For the distribution of empty links at clause level another general tendency can be observed. At clause level, it seems to be a clear characteristic of the English texts to exhibit more empty links. All English original texts as well as all English translations have more empty links than their matching German texts (see Figure 4), with English translations in SPEECH displaying the highest number: here, 35% of the clauses have no link to a clause in the German source text.

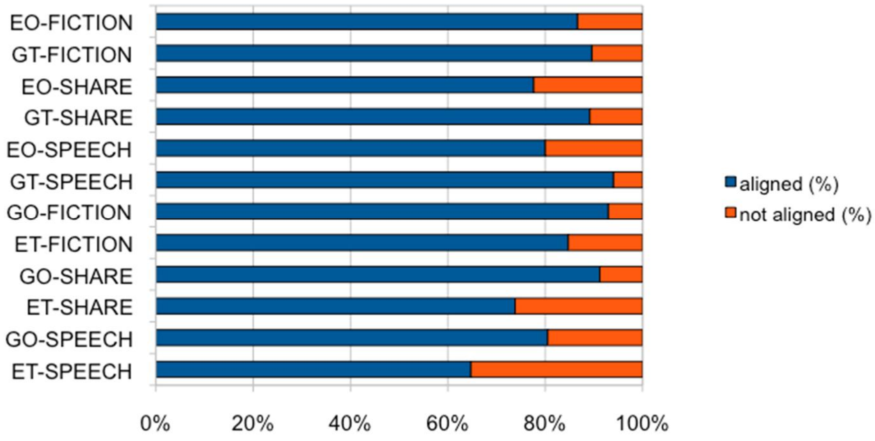

Figure 4: Clause alignment statistics

When correlating the number of empty links with the total number of clauses, we find a similar picture. In SPEECH as well as in the other registers, the English texts always display a higher number of clauses, although all corpora are of approximately the same size with respect to the number of words. Here it is important to bear the following point in mind: the clause segmentation in CroCo is verb-based, i.e. each verb (finite or non-finite) is taken as the basis of a new clause. Thus, empty links occur where a clause (containing a verb) in one text has no direct verbal equivalent in the respective text of the other language either because the content of this clause is expressed in a non-verbal construction or because it is simply left out.

	total number clauses	aligned clauses	empty links
GO_SPEECH	3,798	3,058 (80.52%)	740 (19.48%)
ETrans_SPEECH	4,856	3,144 (64.74%)	1,712 (35.26%)
EO_SPEECH	3,853	3,083 (80.02%)	770 (19.98%)
GTrans_SPEECH	3,170	2,981 (94.04%)	189 (5.96%)

Table 2: Clause alignment in SPEECH

For the register SPEECH, the numbers are as displayed in Table 2. The numbers in the second column (aligned clauses) probably represent unproblematic cases, where clauses in the source text can easily be connected to clauses in the target text, perhaps due to similar constructions or rather simple sentences.

The figures in the third column (empty links) leave room for interpretation. Concerning the translation direction German-English, we find that in many cases empty links occur in English subordinate clauses or expressions that resolve more complex structures of the German original text. These are, for example, nominalizations or nouns with premodifying participle constructions, as can be seen in (12) and (13).

(12) a. [Mittlerweile ist anerkannt,] [dass es zur Sicherung von Beschäftigung vor allem auf Flexibilität ankommt.] (GO_SPEECH_007)
b. [It has now been recognized] [that flexibility is the most important factor] [when it comes] [to safeguarding jobs.] (ETrans_SPEECH_007)

(13) a. [Die Staats- und Regierungschefs der Europaeischen Union haben in Göteborg erneut ihre Bereitschaft bekräftigt,] [die in Kyoto eingegangenen Verpflichtungen zur Verminderung der Treibhausgase zu erfüllen.] (GO_SPEECH_001)
b. [In Gothenburg the EU heads of state and government reaffirmed their willingness] [to fulfil the commitments] [they made in Kyoto] [to reduce greenhouse gases.] (ETrans_SPEECH_001)

In both examples, there are only two clauses in the German sentence; these are split into four and three clauses in the respective English translations.[7] In (12), the nominal group *zur Sicherung von Beschäftigung* is transformed into two subordinate clauses with a finite (*comes to*) and a non-finite verb (*safeguarding*). In (13), the participle of the nominal group *die in Kyoto eingegangenen Verpflichtungen* is translated with the finite verb *made*. This strategy results in one more clause in the English translation than in the German original and therefore in an empty link for this additional clause. There seems to be a tendency within the English translations to use formulations that are more explicit and less dense than those in the German texts. Fabricius-Hansen (1998) reports similar results in a comparison of German source texts and the respective translations into English and Norwegian and discusses a "tendency towards higher informational density that can be observed in German texts of the relevant type and which is correlated with a relatively high degree of syntactic complexity" (Fabricius-Hansen 1998: 197). She relates this phenomenon to different types of discourse information structure, assigning a "hierarchical type" to German texts and an "incremental" one to the English translations (Fabricius-Hansen 1998: 202-203), with the latter increasing incrementality by information splitting (Fabricius-Hansen 1998: 231). In terms of translation properties we could speak of simplification and explicitation here, i.e. a tendency in translations to simplify their texts and to spell things out rather than leaving them implicit (Baker 1996: 180-181). At the same time, the high number of clauses can be interpreted as normalization: the translation (over-)uses typical features of the target language, such as a low informational density (Baker 1996: 183).

Another example where the English translation shows a strong preference for verbal (especially non-finite) instead of nominal constructions is example (14), which consists of one single clause in German and of four clauses in English (the following segments form one discontinuous clause with several embedded clauses in between, as marked by the brackets):

(14) a. *[Mit der am 16. Juli in Bonn beginnenden Klimakonferenz der Vereinten Nationen gehen die jahrelangen Bemühungen um ein verbindliches Klimaschutz-Abkommen in die entscheidende Phase.] (GO_SPEECH_001)*

b. *[With the UN Climate Conference [beginning in Bonn on July 16] the many years of efforts [aimed at] [achieving a climate protection agreement] will enter the crucial final phase.] (ETrans_SPEECH_001)*

Here, the German nominal expression *Bemühungen um* is translated with *efforts aimed at achieving*. The decision of the translator to use this construction results in two more clauses in the English sentence: instead of translating the German expression rather literally with *efforts toward*, a longer and more explicit phrasing is used. Again, different types of information structure (hierarchical vs. incremental type, see above) could offer an explanation for the higher number of empty links in the English texts.

[7] Clauses are segmented irrespective of their dependence within the syntactic structure. Therefore, embedding cannot be retraced.

Additionally, this example illustrates a further reason: the restricted options of English concerning pre- and postmodifying. In the German sentence, the noun *Klimakonferenz* is premodified with the construction *mit der am 16. Juli in Bonn beginnenden*. Since the participle *beginnenden* is used in an adjectival way (as is almost always the case with premodifying participles) it does not form the basis of a new clause. The same information could have been conveyed using a less dense construction, e.g. a postmodifying relative clause like *Mit der Klimakonferenz, die am 16. Juli in Bonn begann*, in this way splitting the sentence into two clauses. For English, all options to translate this sequence result in a postmodifying construction containing a verb.

A considerable number of empty links in the English texts is due to properties of the language system in comparison to German. Here again a connection can be drawn to the translation property of normalization: Teich (2003: 218) relates this to contrastive differences in the range of options available in source and target language, positing that fewer options in the target language entail compensations which may then lead to normalization. English has fewer options compared to German with respect to pre- and postmodification, which leads to normalization. That in turn would explain at least in part the high number of empty links.

Still another explanation could be different registerial restrictions. In example (15), the German adverb *deshalb* is translated with the expression *that is why*, again resulting in an additional clause in the English text:

(15) a. [Deshalb machen hohe Abgaben Arbeit teuer] [und können doch nicht verhindern,] [dass unseren Sozialsystemen der Kollaps droht.] (GO_SPEECH_007)

b. [That is why] [high taxes make work expensive] [and yet cannot protect our social system from] [impending collapse.] (ETrans_SPEECH_007)

It is possible that the use of *therefore* instead of *that is why* would sound too formal for a speech or that a more explicit reference to the previous sentence has to be made. In any case, this is an example for a situation in which the individual decision of the translator influences the number of empty links. If this proves to be a typical pattern (all three occurrences of *that is why* are in fact translations of *deshalb*), it can be interpreted as a possible sign of explicitation because it shows a "rise in the level of cohesive explicitness" (Blum-Kulka 1986: 19).

For the translation direction English-German in SPEECH the picture is a different one, with only 5.96% of empty links in the target texts (GTrans_SPEECH). These are mainly cases where the translator has to opt for a different translation because of lexical differences of the verb as in (16) or where s/he uses a German non-finite construction that results in an additional clause in (17):

(16) a. [One of President Bush's primary objectives in that meeting was] [to take a further step in our efforts] [to persuade President Putin] [to join us in] [creating a new strategic framework for] [dealing with the security threats] [that we now face,] [while moving us toward a cooperative relationship with Russia and away from the adversarial legacy of the Cold War.] (EO_SPEECH_003)

b. [Eines der vorrangigen Ziele von Präsident Bush bei diesem Treffen war es,] [einen Schritt *voranzukommen bei unseren Bemühungen,] [Präsident Putin zu überzeugen,] [mit uns gemeinsam einen neuen strategischen Rahmen für die Handhabung von Sicherheitsbedrohungen zu schaffen,] [denen wir uns nun gegenübersehen,] [während wir gleichzeitig auf kooperative Beziehungen zu Russland hinarbeiten] [und die feindliche Gesinnung des Kalten Kriegs hinter uns lassen.] (GTrans_SPEECH_003)*

Here, it is semantically impossible to retain the structure *moving us toward… and away from* in the translation. Two different verbs have to be used and thus one clause in the English text is split into two clauses in the German translation.

(17) a. *[Our European friends and allies share our concern about the need] [to accord recognition to surviving Holocaust victims within their lifetimes.] (EO_SPEECH_006)*

b. *[Unsere europäischen Freunde und Bündnispartner teilen unser Anliegen,] [den überlebenden Holocaust-Opfern zu Lebzeiten Anerkennung zuteil werden] [zu lassen.] (GTrans_SPEECH_006)*

In (17), the translator uses an infinitive construction with the modifying verb *lassen*, which leads to two verbs and therefore two clauses, where the English original formulation consists of only one clause.

Apart from these few cases, the German translations adhere rather closely to the English source texts. 94.04% of the clauses are aligned, and it seems as if the translators are trying to use the same structures in the German texts that can be found in the English ones. This could be interpreted as source language shining through, which is, as it were, the 'counterpart' of normalization. Lexico-grammatical properties of the source language can be reflected in the target language as well, especially in areas where the target language is more flexible than the source language (cf. Teich 2003: 218). With regard to pre- and postmodification it is therefore possible that the German translations follow the pattern used in the English originals, because German is not confined to one specific option, but can afford to more or less copy the structures of the English text. This strategy would result in a lower number of empty links.

Nevertheless, it has to be borne in mind that there are also empty links in the English source texts. They occur, for example, where English non-finite constructions are translated with the help of nominal constructions, as can be seen in example (18).

(18) a. *[As a result: in the Middle East, countries are going back to the negotiating table,] [we have established a new relationship with Russia] [that promises] [to form the a [sic] new framework of constructive arms control agreements,] [and we are openly discussing the very real problems and the hard reality] [attached to the proliferation of weapons of mass destruction.] (EO_SPEECH_005)*

b. *[Das Ergebnis hiervon ist: - die Rückkehr der Länder im Nahen Osten an den Verhandlungstisch, - der Aufbau neuer Beziehungen zu Russland, [die das Versprechen eines neuen Rahmens für konstruktive Rüstungskontrollabkommen bergen,] und - eine offene Diskussion über die sehr realen Probleme und die harsche Wirklichkeit im Zusammenhang mit der Verbreitung von Massenvernichtungswaffen.] (GTrans_SPEECH_005)*

The results of US President Bush's policies are listed with bullet points in the English source text. For each result the author starts with a new sentence, sometimes containing several clauses. In the German translation, each result is presented as a noun phrase containing no verbs. As explained above, this rather dense discourse information structure is characteristic of German.

Empty links at clause level can be attributed in most cases to contrastive differences between English and German. In terms of translation properties, these differences often result in explicitation (mainly in the English translations) and normalization in combination with source language shining through, as a closer look at the high number of empty links in the English texts reveals. The combination of source language shining through and target language normalization leads to a hybridization in the translations.

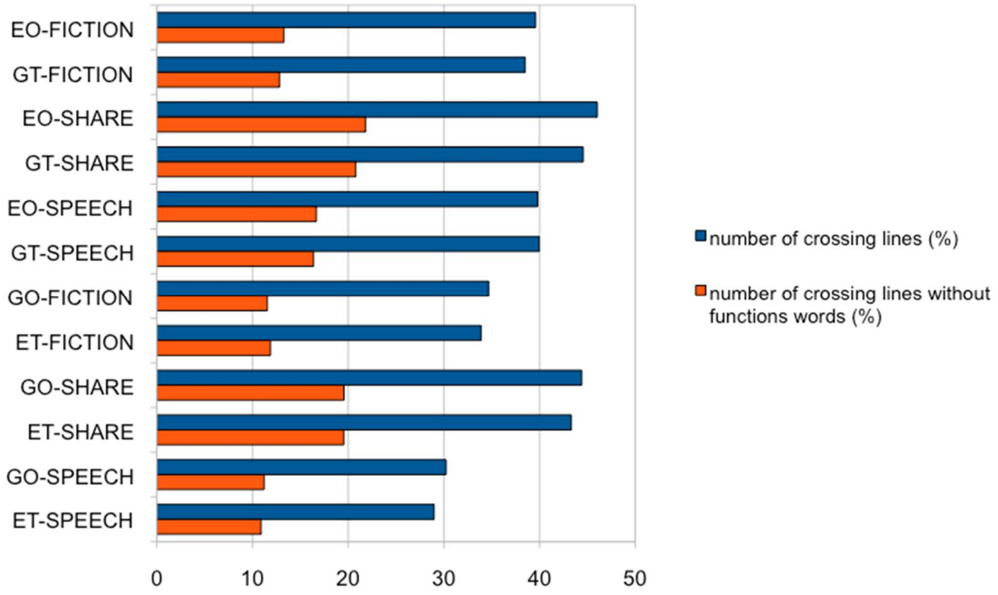

Figure 5: Percentages of crossing lines between words and grammatical functions

4.3 Crossing lines between words and grammatical functions

Crossing lines between words and grammatical functions in corresponding source- and target texts shed light on the variation in terms of grammatical "responsibility" of the words used in the parallel versions[8]. They are thus indicative of shifts in perspective as, for instance, described by Vinay and Darbelnet (1958) in terms of modulation, i.e. a semantic shift in perspective.

As mentioned previously, the validity of the query results for crossing lines on all levels involving word level is limited due to the relatively low quality of the existing word alignment (especially concerning recall; see also section 3.3). In terms of the present discussion this means that we can only draw some very preliminary

[8] The percentage of crossing lines for words and grammatical functions is calculated on the basis of the number amount of grammatical functions (per subcorpus) for which word shifts occur (the percentage of sentences containing crossing lines between words and grammatical functions in relation to the number of all sentences per register.).

conclusions from the existing figures. A cursory look at the aligned texts suggests that there are frequent candidates for crossing lines that are not retrieved by our query because recall of our word alignment tools is still lower than one would ideally hope.

Figure 5 shows that crossing lines are similarly frequent in pairs of source and target registers. The clearest pattern emerging is an organization in registers. All SHARE subcorpora display a similarly high frequency of crossing lines, just as all FICTION subcorpora display a comparably low frequency of crossing lines. The only register not showing such a clear pattern is SPEECH. Here, the pairs of original and target registers are still grouped together. This becomes particularly obvious when only taking into account lexical words and excluding function words as depicted in Figure 5.

This raises the question of why it is this level that appears to be prone to register influences. One starting point could be differing distributions of grammatical functions in the registers. If the grammatical functions are distributed differently in the four subcorpora in one register, this could be reflected in more crossing lines between originals and translations in this register. In order to assess the variation between subcorpora in the three registers, we compute the standard deviation between the values for each function in the individual registers. The sum of the individual standard deviations should be higher in a register containing more variation between the functions. As table 3 shows, SHARE in fact has more variation reflected by higher standard deviations for the individual functions. The lowest variation is found in FICTION, which has consistently lower frequencies of crossing lines.

While this appears to be a plausible explanation for the differing numbers of crossing lines, contrastive differences, i.e. an aspect not related to the register, could play a role as well. Prepositional objects and complements, for instance, display different frequencies in the two languages resulting in more similarities between originals and translations in the same language (see Table 3). Apparently, prepositional objects play a greater role in the German registers whereas complements appear to be more typical of the English registers. Consequently, it is these functions in particular that seem to be more prone to crossing lines.

Table 4 displays the most frequent crossing lines between words and grammatical functions organized by register and translation direction. Due to the abovementioned weaknesses of recall in our word alignment, we do not interpret frequencies but only the ranking of the most common shifts.

	FICTION				
	EO	ETrans	GO	GTrans	Std. dev.
adv_*	18.87	18.01	18.40	19.94	0.8335
appo	0.92	0.68	0.71	0.70	0.1141
compl	5.19	5.04	3.78	3.28	0.9389
dobj	10.77	10.26	10.82	11.76	0.6262
fin	23.43	23.20	24.39	23.87	0.5243
iobj	0.81	0.81	1.93	2.03	0.6766
other	6.76	7.61	7.75	7.09	0.4581
pred	6.04	6.75	4.83	5.26	0.8515
probj	1.74	1.75	2.49	2.27	0.3765
subj	21.08	21.27	19.86	19.37	0.9263
	SHARE				
	EO	ETrans	GO	GTrans	Std. dev.
adv_*	17.98	18.22	21.15	21.28	1.8005
appo	1.60	1.15	0.41	0.81	0.5065
compl	6.42	6.54	4.16	4.15	1.3433
dobj	12.19	10.73	10.47	11.54	0.7870
fin	22.54	21.75	20.96	21.33	0.6771
iobj	0.88	0.93	1.70	1.54	0.4196
other	11.07	12.10	12.64	11.50	0.6863
pred	7.22	9.12	8.87	8.27	0.8487
PROBJ	2.84	2.62	4.40	4.68	1.0562
SUBJ	21.32	20.82	19.78	19.17	0.9756
	SPEECH				
	EO	ETrans	GO	GTrans	Std. dev.
ADV_*	14.61	15.52	16.91	15.90	0.9534
APPO	0.81	1.41	0.83	0.42	0.4117
COMPL	6.06	8.06	5.79	5.57	1.1422
DOBJ	12.18	10.35	10.92	12.70	1.0893
FIN	22.63	21.86	21.41	22.95	0.7017
IOBJ	0.76	0.49	1.82	1.62	0.6467
OTHER	6.79	7.96	9.05	6.30	1.2312
PRED	11.08	10.21	8.27	8.92	1.2644
PROBJ	2.93	2.21	3.94	4.25	0.9357
SUBJ	22.05	21.85	21.00	21.24	0.4977

Table 3: Distribution of grammatical functions per subcorpus in per cent of all functions per subcorpus

FICTION		SHARE		SPEECH	
E2G	G2E	E2G	G2E	E2G	G2E
dobj → subj	probj → dobj	compl → probj	probj → dobj	dobj → probj	subj → dobj
compl → dobj	dobj → subj	dobj → subj	subj → compl	dobj → compl	subj → compl
subj → dobj	fin → pred	dobj → probj	subj → dobj	compl → probj	probj → compl
dobj → fin	compl → subj	compl → dobj	probj → compl	subj → dobj	dobj → compl
dobj → probj	subj → dobj	dobj → compl	dobj → compl	dobj → subj	probj → dobj
fin → dobj	dobj → compl	compl → subj	fin → pred	pred → fin	dobj → subj
adv_mod → dobj	fin → compl	probj → dobj	dobj → subj	compl → dobj	fin → compl
pred → fin	pred → fin	subj → dobj	compl → dobj	compl → subj	fin → pred
compl → subj	fin → subj	fin → pred	adv_mod → compl	subj → compl	fin → subj
adv_cause → dobj	fin → dobj	pred → fin	subj → probj	compl → fin	compl → subj

Table 4: The ten most frequent crossing lines per register and translation direction

Table 4 shows how the translators shift from prepositional object to other functions in the translation direction German-English, thus adapting to the target language preferences, e.g. prepositional objects in the German FICTION texts are frequently translated by English direct objects. When translating from English to German, translators shift words away from complements to other functions, e.g. in SHARE to prepositional objects. Table 4 indicates that this also works in the opposite direction: translators not only avoid functions that are less typical in the target language, but also shift into preferred functions. Words are moved from various German functions into English complements, as exemplified by the second to fourth rank in SPEECH translations into English in table 4. A shift from German prepositional objects to English direct objects may be a general strategy not necessarily limited to a given register, as shown by the fact that this crossing line is most common in registers as divergent as FICTION and SHARE and is still fairly common in SPEECH. (19) to (22) exemplify theses shifts for the three registers.

(19) a. *Er hat sich darauf verlassen, dass wir von drinnen sein Lächeln sehen können. (GO_FICTION_007)*

b. *He just assumed we could see his smile from inside. (ETrans_FICTION_007)*

Together with and initiated by the pronominal adverb *darauf*, the whole *dass* subordinate clause in the German original in (19) forms a prepositional object. Note that the annotation on which this discussion is based is limited to the highest node in the sentence, thus the *dass* clause is not analyzed further. This discontinuous prepositional object is shifted to a direct object in the English translation. In our query, the hit for the shift is triggered by the aligned noun pair *Lächeln* in the German prepositional object and *smile* in the English direct object. However, this analysis is somewhat problematic. Taking a closer look, we can see that *Lächeln* is actually part of a direct object in the *dass* clause, and should not account for the shift from prepositional object to the direct object. This effect is due to our top-level only annotation, an issue we will come back to in subsection 5.2.

(20) a. *1995 haben wir auf 125 Jahre Deutsche Bank zurückgeblickt. (GO_SHARE_009)*
b. *In 1995 we celebrated Deutsche Bank's 125th anniversary. (ETrans_SHARE_009)*

In (20) from the SHARE register, the name of the bank reporting to its shareholders is shifted from the postmodification within the prepositional object in German to premodification of the direct object in the English translation.

(21) a. *Nach wie vor ist der Zinsüberschuß nach Risikovorsorge mit 9,7 Mrd DM die bei weitem wichtigste Ertragskomponente. Allerdings weisen die unterschiedlichen Steigerungsraten der einzelnen Ergebniskomponenten auf die Veränderungen im Geschäft hin. (GO_SHARE_009)*
b. *Although net interest income after provision for losses on loans and advances, at DM 9.7 billion, is still by far the most important component of income, the individual figures highlight the changes in our business. (ETrans_SHARE_009)*

(22) a. *Daher setzen wir uns nachdrücklich für die Schaffung eines europäischen Systems der Finanzaufsicht ein. (GO_SPEECH_002)*
b. *Hence we expressly support the establishment of a European system of financial supervision. (ETrans_SPEECH_002)*

Example (21) still from SHARE and (22) from SPEECH underline that the specific type of crossing lines exemplified there is largely due to lexical reasons. The German verb *hinweisen* selects the preposition *auf* for its object. Possibly, this finding points to a higher frequency of verbs taking certain types of prepositional object in German than in English. Globally, however, this has to be related to phrasal verbs whose particle is annotated as part of the verb in the CroCo annotation and consequently only leaving prepositional verbs as those taking a prepositional object.

Other shifts may be more restricted to a given register, as, for instance, the shift from an English complement to a German prepositional object. This is particularly prominent in SHARE. Here, often similar reasons apply as with empty links for complements described in subsection 4.1.

Having established some potential causes for individual phenomena in the three registers, we can now return to the overall number of crossing lines on this level in the three registers. Compared to the other two registers under scrutiny here, the figures suggest that FICTION has relatively few crossing lines in both translation directions (see Figure 5). Frequently, crossing lines concern changes between finite and predicator, as is the case in example (23). The perfect tense in the English original is translated by a present tense verb in German, thus resulting in a crossing line of *happened* and *geschieht*.

(23) a. *And what has happened before a few years have passed? (EO_FICTION_006)*
b. *Und was geschieht, ehe noch ein paar Jahre vergangen sind? (GTrans_FICTION_006)*

While the shift in (23) can be attributed to a deliberate change in tense by the translator, the shift between finite and predicator in (24) is due to language contrast.

(24) a. *Aber Sie wissen nichts. (GO_FICTION_007)*
b. *But you don't know anything. (ETrans_FICTION_007)*

The English negation requires the auxiliary *do* that results in the dissociation of the predicate into the auxiliary finite and the full verb as predicator. The German text does not require this and consequently only consists of a finite.

An informationally more marked use of German as in (25) results in a frequent crossing line in this register and translation direction, a shift between direct object and subject.

(25) a. *Die Frauen hat das nicht gerade zimperlich gemacht. (GO_FICTION_007)*
b. *The women weren't exactly prudes. (ETrans_FICTION_007)*

The translator has avoided putting the direct object at the front of the sentence in the English translation, as is the case in the German original. For English, this order of grammatical functions is highly marked. Preserving the order of the content, the translator here decided to shift *women* to the subject function, adhering to the more rigid canonical order of grammatical functions in English, thus of course sacrificing some of the information structure of the original.

SPEECH contains the lowest number of crossing lines in the translation direction German to English. Even fairly complex structures as in (26) do not necessarily require numerous shifts in grammatical functions.

(26) a. *Wenn wir also in diesem Sinne unseren Interessen und Werten dienen wollen, dann muss Europa erstens wachsam gegenüber den neuen Bedrohungen sein, denen die freien und offenen Gesellschaften ausgesetzt sind. (GO_SPEECH_010)*
b. *So if we want to serve our interests and values in line with this definition, Europe must: firstly, be vigilant to the new threats to which the free and open societies are exposed. (ETrans_SPEECH_010)*

Possibly, this is due to a more canonical word order in the German SPEECH register requiring fewer adjustments in the English translation to conform to the more fixed word order of English. The percentage of subjects in sentence-initial position appears to corroborate this assumption. The percentages of grammatical subjects in relation to all grammatical functions in sentence-initial position in the German FICTION and SHARE registers are 42.16% and 45.87% respectively. By contrast, SPEECH exhibits 54.45% of subjects in this position, displaying a register-specific feature and thus making the English translators' task easier.

In the opposite translation direction, SPEECH contains more crossing lines between words and grammatical functions. A potential language contrast between English and German is a shift from coordination to subordination as in (27). This is reflected in crossing lines because the whole subordinate clause in the translation is analyzed as one grammatical function in the CroCo annotation (here an adverbial) whereas the chunks in the coordinated clause are analyzed individually (*resolution* is part of a direct object).

(27) a. *Every country has its own political issues and this makes resolution of our disputes increasingly difficult. (EO_SPEECH_009)*
b. *Jedes Land hat seine eigenen politischen Anliegen, wodurch die Streitschlichtung zunehmend erschwert wird. (GTrans_SPEECH_009)*

Example (28) displays a shift where the word *fight* is moved from the direct object in the original to the subject in the German translation. This represents a typical case of modulation, where the perspective is shifted from the persons confronted with this fight to the fight itself. Beyond the translation shift of modulation this exemplifies House's (1997) cross-cultural difference in terms of orientation towards persons in English versus orientation towards content in German.

(28) a. *And if the EU does as it has in the past, and provides financing to Airbus at below-market rates of return, we could be facing a very large and highly contentious fight in the WTO. (EO_SPEECH_009)*
b. *Und wenn die EU sich wie in der Vergangenheit verhält und dem Airbus Finanzierung zu Zinssätzen unter den auf dem Markt gültigen bietet, könnte uns ein großer und sehr kontroverser Kampf in der WTO bevorstehen. (GTrans_SPEECH_009)*

Word order contrasts combined with different mappings of semantic roles onto grammatical functions between English and German may typically result in crossing lines as represented by (29). The subject of the German passive original is positioned after the finite, which does not lead to an informationally highly marked construction in German. Rather than rearranging the linear precedence of clause elements in English, the translator has opted for rearranging the assignment of semantic roles to grammatical functions by choosing active voice. *Basis*, the aligned translation of *Grundlage*, is consequently no longer part of the subject but of the direct object. (30) displays a similar case.

(29) a. *Gleichzeitig wurde hiermit auch die Grundlage für die Einführung von Hedgefonds in Deutschland und damit für den direkten Zugang deutscher Anleger zu diesem innovativen Produkt gelegt. (GO_SPEECH_002)*
b. *At the same time it established the basis for the introduction of hedge funds in, thus affording German investors direct access to this innovative product. (ETrans_SPEECH_002)*

(30) a. *Damit werden Investitionen von rund 10 Mrd. DM angestoßen und 5 - 7 Mio. t CO2 eingespart. (GO_SPEECH_001)*
b. *It will generate investments of around 10 billion marks and reduce CO2 emissions by 5-7 million metric tons. (ETrans_SPEECH_001)*

(31) and (32) represent cases where there is no apparent reason forcing the translator to change the word order and, at the same time, the voice of the sentence. The crossing lines can be seen as symptoms of a whole range of changes that are obviously due to the translator. When seen in combination with the respective source sentence, these translations show clear indications of the translation process as a motivating variable. Nevertheless, they do not easily lend themselves to an

interpretation in terms of translation properties as described by Baker (1996) and others.

(31) a. *In Deutschland haben wir bisher noch keine Entscheidung über die Einführung von REITs getroffen. (GO_SPEECH_002)*
b. *No decision has yet been taken in Germany on the introduction of REITs. (ETrans_SPEECH_002)*

(32) a. *Dieser Markt hat sein Potenzial bei weitem noch nicht ausgeschöpft. (GO_SPEECH_002)*
b. *The full potential of this market is by no means exhausted. (ETrans_SPEECH_002)*

Concentrating on SHARE, where most of the crossing lines occur in both directions, we find examples like (33). Here, a different constituent structure (subject complement plus complementation in EO versus full verb plus prepositional object in GTrans) mapped onto very similar structures in terms of word order results in a crossing line. A certain share of instances of crossing lines may be due to cases like this. Example (34), however, is more representative of shifts occurring in translation in our data. Whereas *Der Wandel* (*the change*) constitutes the subject in the German original, it is realized as a prepositional object in the translation with the patient becoming the subject. This results in a major shift in perspective in the translation.

(33) a. *The same is true for Human Resources reviews. (EO_SHARE_004)*
b. *Das gleiche gilt für "Human Resources Reviews". (GTrans_SHARE_004)*

(34) a. *Der Wandel geht an unseren Filialen nicht vorüber. (GO_SHARE_009)*
b. *Our branches are not unaffected by these changes. (ETrans_SHARE_009)*

The crossing line in example (35) is equally interesting in that, apart from a number of shifts, the subject of the original (*die moderne Universalbank*) is hidden in the postmodification of the complement in the translation (*an impressive demonstration of a modern universal bank's capabilities*).

(35) a. *Mit ihrer Plazierungskraft im Inland hat die moderne Universalbank ihre Möglichkeiten eindrucksvoll unterstrichen. (GO_SHARE_009)*
b. *The placement of this issue in Germany was an impressive demonstration of a modern universal bank's capabilities. (ETrans_SHARE_009)*

Beyond modulation as a type of translation shift these crossing lines do not easily lend themselves to interpretations in terms of translation properties. Instances like (35) point to implicitation rather than explicitation in terms of constituency structure, because the referent (and the words) contained in the subject in the original is not only shifted into the complement in the translation, but is additionally reduced to postmodification instead of representing the head of the phrase in the original.

The discussion of crossing lines between words and grammatical functions has shown that these crossing lines are symptomatic of a whole range of factors relevant to translation. Of course they are subject to a wide range of influences that prohibit mono-causal explanations. They are, however, indicative of differences between registers as well as contrastive differences in the frequency of certain grammatical functions and in word order. Furthermore, they show translation shifts, typically in the area of modulation, which must often be attributed to translator behavior. Finally, we have also shown dimensions of cross-cultural differences in House's sense at work.

A direct and simplistic association between crossing lines between words and grammatical functions and translation properties should be avoided: while crossing lines definitely have implications for properties such as explicitation, normalization, simplification, shining through and others, the relationship is complex and needs further evidence.

5 Future work

We have shown in this paper the query power which can be provided by an annotation which involves multi-level annotation and alignment and which to a considerable extent can be done (semi-)automatically, at least when it comes to tagging and chunking. The value of the CroCo-specific annotation lies on the one hand in the alignment which was partly done by human annotators (for the clause and sentence level). On the other hand, the manual annotation of levels like phrase structure and grammatical functions delivers a high-quality set of data. Moreover, we have demonstrated the methodological value of querying empty links and crossing lines for the detection of translation shifts and investigation of translation properties. Within the context of the CroCo project there are a number of spin-off projects, e.g. further investigating cohesion in originals and translations, or how "parallel" valency is between English and German. In some of these projects, the limitations of the CroCo annotation – esp. the decision to keep the functional annotation on the top level, with the exception of clauses which are annotated for their functions as well – become obvious.

The following subsection 5.1 outlines some thoughts on how the findings in this paper will help realize a project on valency queries. In order to study valency and other phenomena in a more detailed fashion and on all linguistic levels, i.e. with respect not just to main and subordinate clauses, but also to embedded structures, we will add deeper annotation levels to CroCo. Subsection 5.2 briefly sketches these plans.

5.1 Valency queries

One of the big hopes in parallel corpora is that they may enable us to build multilingual valency dictionaries (semi-)automatically. This would facilitate the work of the lexicographer enormously. Corpora allow for the extraction of large amounts of data in a short time and may contain examples a lexicographer would not easily think of. Examples for monolingual valency dictionaries based on corpora are the Czech PDT-VALLEX[9] and the English Erlangen Valency Pattern Bank[10].

[9] http://ufal.mff.cuni.cz/vallex/2.5/doc/home.html

[10] http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi

In order for valency queries to work, we must rely on the fact that the structures between original and translation are maximally equivalent. As we have seen in our results, this is more valid for some linguistic levels than for others. For the sentence level, for instance, we found that in all registers and all translation directions at least 99% of the sentences have an equivalent. If we see the sentence as a valency carrier plus the complements and adjuncts accompanying it, this means for the purpose of valency extraction that in 99% of all cases we will have a pair of structures which can be used for further investigation.

The results on empty links and crossing lines for grammatical functions, which we presented in this paper, will be most valuable for our valency studies as well. The considerable number of occurrences for these two phenomena already suggest that we are likely to find quite a number of valency-related phenomena which occur in translation. In example (12), for instance, we have a case in which the nominal group *zur Sicherung von Beschäftigung* was translated with a verbal expression *to safeguarding jobs*, resulting in an empty link on the clause level. From a valency point of view, the shift from noun to verb also shifts the syntactic valency frame of *Sicherung* which adds the object as a *von*-PP, compared to the direct object *jobs* that the verbal equivalent *safeguarding* requires. Another kind of valency shift involves cases of shifts in grammatical functions, which have been described in subsection 4.3. Furthermore, a pilot study has revealed that there is a considerable percentage of cases in which the main verbs do not perfectly match. This was the case for about 20-40% in our sample of 300 sentence pairs (50 from each register and translation direction). For the instances of divergences found, there was either a shift in meaning (e.g. *jmdm. gut tun* 'do so. good' vs. *benefit from sth.*) or the full verb on the one side has a syntactically more complex equivalent on the other side, e.g. a copula construction, an idiomatic expression or a support verb construction, often changing the overall structure of the sentence. As for copula constructions, it has already been outlined in subsection 4.1 that they are more frequent in English and thus account for quite a number of empty links for (or?) shifts departing from (predicative) complements. There seems to be only a small minority of cases in which a sentence has been completely re-phrased, thus rendering the sentence pair useless for the study of valency-related phenomena.

In order to study these phenomena, we will need a deeper annotation of structures, which will be provided by converting (parts of) the CroCo corpus to a parallel dependency treebank, the plans for which are briefly outlined in the following subsection.

5.2 Towards a parallel treebank

Let us go back to our *Lächeln*-example, number (19) from subsection 4.3. We can see in this example, as has already been discussed in 4.3, that the top-level-only annotation in CroCo sometimes negatively affects our queries. The *dass*-clause is combined into a prepositional object together with the *darauf*-adverb. When querying for the word pair *Lächeln* and *smile*, we get a shift from prepositional object to direct object, which is triggered by our method of analyzing the structure rather than a real shift. This kind of annotation is also disadvantageous when looking into valency phenomena. Elements might be deeper embedded when shifting from full verb to copula plus adjective-constructions, for instance. We would like to be able to detect these kinds of shifts automatically as well.

We have thus decided to transform at least parts of the CroCo-annotation into a parallel dependency treebank, in a spin-off project. When tentatively translating our functional analysis of the German original sentence from the *Lächeln*-example into a dependency tree, we could get an analysis as exemplified in Figure 6. From a dependency tree like that depicted in the figure, we can deduce the correct grammatical function for *Lächeln*, but still preserve the information that the whole subordinate clause with *sehen* as root functions as a prepositional object.

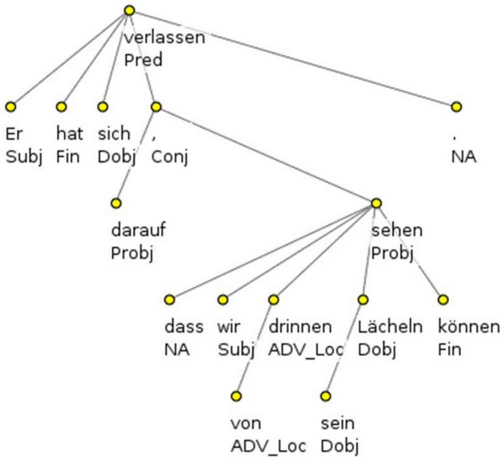

Figure 6: A possible dependency analysis for example (19a)

We will be using the tools created within the Prague Dependency Treebank project, namely TrEd[11] plus some extensions for working with parallel data which it delivers (Böhmová et al. 2000). We will annotate dependencies at the functional level, using grammatical categories as subject, object etc. Annotation of deep syntactic or semantic roles is not planned at present, but may be added at a later stage. The trees will be aligned on the level of the grammatical functions. This alignment will allow us to more reliably query shifts on this level.

6 References

Baker, M. 1993. "Corpus linguistics and translation studies. Implications and applications". In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology. In Honour of John Sinclair*. Amsterdam, Philadelphia: Benjamins, 233–250.

Baker, M. 1995. "Corpora in translation studies: an overview and some suggestions for future research". *Target*, 7 (2), 223–243.

[11] http://ufal.mff.cuni.cz/~pajas/tred/

Baker, M. 1996. “Corpus-based translation studies: the challenges that lie ahead”. In H. Somers (Ed.), *Terminology, LSP and Translation. Studies in Language Engineering*. Amsterdam/ Philadelphia: Benjamins, 175-186.

Blum-Kulka, S. 1986. “Shifts of cohesion and coherence in translation”. In J. House & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies.* Tübingen: Gunter Narr, 17-35.

Böhmova, A., Hajič, J., Hajičová, E. & Hladká, B. 2000. “The Prague Dependency Treebank: A Three-Level Annotation Scenario”. In A. Abeillé (Ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluver Academic Publishers.

Brants, T. 2000. “TnT: a statistical part-of-speech tagger”. *Proceedings of the Sixth Conference on Applied Natural Language Processing*, 224-231.

Bresnan, J. & Kaplan, R. 1982. “Lexical-Functional Grammar: a formal system for grammatical representation”. In J. Bresnan, (Ed.), *The Mental Representation of Grammatical Relations*. MIT Press, 173-281.

Burnard, L. & Bauman, S. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative.

Catford, J. C. 1965. *A Linguistic Theory of Translation. An Essay in Applied Linguistics.* Oxford: Oxford University Press.

Čulo, O., Hansen-Schirra, S., Neumann, S. & Vela, M. 2008. “Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus”. *Proceedings of the LREC 2008 Workshop "Building and Using Comparable Corpora"*, Marrakesh, Morrocco.

Cyrus, L. 2006. “Building a resource for studying translation shifts”. *Proceedings of LREC* 2006, 1240–1245.

Fabricius-Hansen, C. 1998. “Informational density and translation, with special reference to German – Norwegian – English”. In S. Johansson & S. Oksefjell, *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies*. Amsterdam/Atlanta: Rodopi, 197–234.

Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M. & Zesch, T. 2007. “Darmstadt Knowledge Processing Repository Based on UIMA”. *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany.

Hansen-Schirra, S., Neumann, S. & Vela, M. 2006. “Multi-dimensional annotation and alignment in an English-German translation corpus”. *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, 35-42.

Hansen-Schirra, S., Neumann, S. & Steiner, E. 2007. “Cohesive explicitness and explicitation in an English-German translation corpus”. *Languages in Contrast*, 7 (2), 241–265.

Hansen-Schirra, S., Neumann, S. & Steiner, E. forthcoming. *Cross-linguistic Corpora for the Study of Translations - Insights from the language pair English-German.* Berlin: de Gruyter.

Hawkins, J. A. 1986. *A Comparative Typology of English and German. Unifying the Contrasts.* London: Croom Helm.

Heyn, M. 1996. "Integrating machine translation into translation memory systems". *European Association for Machine Translation - Workshop Proceedings*, 111—123.

House, J. 1997. *Translation Quality Assessment. A Model Revisited.* Tübingen: Gunter Narr.

Koller, W. 2001. *Einführung in die Übersetzungswissenschaft*. Tübingen: Gunter Narr.

van Leuven-Zwart, K. 1989. "Translation and original. Similarities and dissimilarities". *Target*, 1 (2), 151–181.

Maas, H.-D., Rösener, C. & Theofilidis, A. 2009. "Morphosyntactic and semantic analysis of text: the MPRO tagging procedure". In C. Mahlow & M. Piotrowski (Eds.), *State of the Art in Computational Morphology. Workshop on Systems and Frameworks for Computational Morphology 2009*. New York: Springer, 76-87.

Müller, C. & Strube, M. 2006. "Multi-level annotation of linguistic data with MMAX2". In S. Braun, K. Kohn & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy. New Ressources, New Tools, New Methods*. Frankfurt a.M.: Peter Lang, 197-214.

Neumann, S. & Hansen-Schirra, S. 2005. "The CroCo project: cross-linguistic corpora for the investigation of explicitation in translations". *Proceedings from the Corpus Linguistics Conference Serie*s, 1-11.

Neumann, S. 2008. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German.* Saarbrücken: Universität des Saarlandes. Unpublished Habilitationsschrift.

Newmark, P. 1988. *A Textbook of Translation.* New York: Prentice Hall.

Och, F.-J. & Ney, H. 2003. "A systematic comparison of various statistical alignment models". *Computational Linguistics*, 29 (1), 19–51.

Özgür, D. 2007. *TIGER API 1.8 – A Java interface to the TIGER corpus*. Available at: http://www.tigerapi.org (accessed August 2010).

Padó, S. 2007. "Translational equivalence and cross-lingual parallelism: the case of framenet frames". *Proceedings of the nodalida workshop on building frame semantics resources for scandinavian and baltic languages*, Tartu, Estonia.

Pollard, C. & Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Sampson, G. 1995. *English for the Computer. The Susanne Corpus and Analytic Scheme.* Oxford: Clarendon Press.

Schiller, A., Teufel, S., Stöckert, C. & Thielen, C. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart, Universität Tübingen.

Sperberg-McQueen, C. M. & Burnard, L. (Eds.) 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* Chicago and Oxford: Text Encoding Initiative.

Steiner, E. 2008. Empirical studies of translations as a mode of language contact - "explicitness" of lexicogrammatical encoding as a relevant dimension. In P. Siemund & N. Kintana (Eds.), *Language Contact and Contact Languages*. Amsterdam: Benjamins, 317-346.

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

Toury, G. 1995. *Descriptive Translation Studies and Beyond.* Amsterdam, Philadelphia: Benjamins.

Vinay, J.-P. & Darbelnet, J. 1958. *Stylistique Comparée du Francais et de l'Anglais. Méthode de Traduction.* Paris: Didier.