Empirical Studies on Language Contrast Using the English-German Comparable and Parallel CroCo Corpus

Oliver Čulo*, Silvia Hansen-Schirra*, Stella Neumann†, Mihaela Vela‡
*Applied Linguistics Johannes Gutenberg University Mainz, Germersheim, Germany
†Applied Linguistics, Translation and Interpreting Saarland University, Saarbrücken, Germany
‡German Research Center for Artificial Intelligence, Saarbrücken, Germany
culo@uni-mainz.de, hansenss@uni-mainz.de, st.neumann@mx.uni-saarland.de, Mihaela.Vela@dfki.de

Abstract

This paper presents results from empirical studies on language contrasts, translation shifts and translation strategies gained by exploiting the CroCo Corpus. The aim of this paper is to show that the insights from investigating the comparable parts of the corpus can be complemented by additionally exploiting the parallel parts of the corpus using the examples of word order peculiarities and diverging part-of-speech frequencies in English and German. The exploitation of the corpus proceeds in two steps. First, contrastive differences are identified in the comparable parts of the corpus. In the second step, the solutions chosen by human translators to deal with the contrastive differences are identified. These can be used to decide between different possible translation strategies and can serve as templates for translation strategies to be adopted in the development of MT systems.

1. Multilingual Corpora in Translation

The creation of linguistic corpora in the past decades has made possible new ways of researching linguistic phenomena and refining methods of processing language with the computer. In the field of translation, corpora are making inroads as well. Corpus-based translation studies are steadily gaining interest thus potentially serving as an input to research in the field of machine translation as well.

The aspects we can study from comparable and parallel corpora differ. However, the decision is not necessarily between creating either a comparable or a parallel corpus. One outcome of the CroCo project[1] is a corpus that contains both parts.

This paper demonstrates how the CroCo corpus (Neumann & Hansen-Schirra, 2005) can be used both as comparable as well as parallel corpus and what kind of insights we can gain for each of the fields mentioned above. It also shows how techniques from both worlds can complement each other.

The paper is organized as follows. In section 2 we shortly introduce the topics of language contrasts, translation shifts, translation strategies and information structure. In section 3 we will present the design and representation of our English-German corpus of originals and translations as well as its exploitation. Section 4 discusses the findings from the corpus exploitation. Section 5 gives an overview of our conclusions and offers an outlook on computational applications of our findings.

2. Strategies for Handling Language Contrasts

Language contrasts can be studied by investigating corpora of the languages involved using multilingually comparable techniques (Granger et al., 2003). Contrasts become visible at all levels of language, in graphology (in written mode), morphology, syntax and on text level and can be investigated empirically with the help of comparable corpora. For instance, the claim that English has a more rigid word order than German with the subject mostly in sentence-initial position can easily be tested on a corpus like the annotated CroCo corpus by simply querying the number of sentences where the subject is in sentence-initial position in both languages (see section 4.1). Examples from the corpus may be helpful to understand how German word order relates to English in terms of rigidness.

When comparing source texts and their translations in another language (parallel techniques), translation shifts become apparent. Translation shifts have been discussed in translation studies since the 1950s (Vinay & Darbelnet, 1958; Catford, 1965; Newmark, 1988; van Leuven-Zwart, 1989). The accounts are similar in that they categorize lexical, grammatical, and semantic shifts. On the level of lexis, the focus is on strategies for gaps or *lacunae*, i.e. lexical items that do not exist in the target language. Grammatical shifts are often called *transpositions* and refer to changing tense, number, person, part-of-speech. They function in the target text without changing the meaning. A special case is what Catford (1965) calls "level shifts" where the shift involves both lexis and grammar, because a given grammatical construction is not available in the target language and has to be replaced by an alternative lexical item reflecting the meaning of the construction. In semantic shifts, or *modulations* (Vinay & Darbelnet, 1958), a change of perspective occurs between source and target text. This may involve concretion, explication, negation of the opposite, (de-) passivization, etc.

In computational linguistics, translation shifts of all types are a crucial issue for the development of MT systems. Identification, classification and formalization of translation shifts have received considerable attention in

[1] http://fr46.uni-saarland.de/croco, funded by the German Research Foundation as project no. STE 840/5-2 and HA 5457/1-2

the MT community (e.g. in the Eurotra project, Copeland et al., 1991). Within this context, Barnett et al. (1991) introduce a rough distinction between translation *divergences* for mere structural differences and *mismatches* for changes which also comprise shifts in meaning. Under the umbrella term *complex transfer*, Lindop & Tsujii (1991) present a comprehensive discussion of examples that appear to be problematic for MT. On this basis, Kinoshita et al. (1992) classify these divergence problems into four categories: argument-switching, head-switching, decomposition and raising. Dorr (1994) proposes a more fine-grained categorization of MT divergences. She distinguishes between thematic, promotional, demotional, structural, conflational, categorical and lexical divergences, thus using linguistic categories. Additionally, she presents a formal description of these divergences and an interlingua approach to a systematic dealing with divergences.

In more recent studies, multiply annotated parallel corpora are used to develop interlingual representations (Farwell et al., 2004) or to learn transfer rules (Čmejrek et al., 2004; Hinrichs et al., 2000). These approaches implicitly include translation shifts in MT procedures and could benefit from input from translation studies. Cyrus (2006) combines the two perspectives, but her focus on the predicate argument structure restricts the findings to semantic shifts. A further limitation of the study results from the direct annotation of translation shifts. A theory-neutral annotation and alignment on different levels like the one proposed here offers the opportunity to query the corpus for different purposes.

On sentential and textual level, the translator is faced with an information structure which, due to grammatical, lexical and other differences cannot always be directly reproduced thus entailing modulation (see section 4.1). The *translation strategies* used to map information structures from one language onto another result in shifts that may occur on all linguistic levels and are due to the translator's understanding as well as idiosyncratic preferences during the translation process, to contrastive differences between the languages involved or to different register characteristics.

The present paper presents a linguistically founded approach to detecting translation shifts and studying language contrasts and translation strategies in a multiply annotated and aligned comparable and parallel corpus.

3. Corpus Design, Representation and Exploitation

The CroCo corpus was built to investigate contrastive commonalitites and differences between the two languages involved as well as peculiarities in translations. It 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 totalling 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 including 2,000 word samples from 17 registers.

The corpus thus consists of both, comparable and parallel, parts. 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) as well as 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 including a brief register analysis that allows additional filter options following the TEI standard (Sperberg-McQueen & Burnard, 1994),
- 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, 1998) which operates on both languages,
- phrase structure again using MPRO and
- 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.

For an effective exploitation of the annotated data, the annotation and alignment is converted into a MySQL database. The information on token level, such as tokenization, part-of-speech, lemmatization and word alignment, is written into tables in the database. The tokens in one language are indexed, each index referring to a string, a lemma, a part-of-speech tag and an index for its alignment in the other language. At chunk level, the tables are filled with information about chunk type and the grammatical function it fulfills. The tables for chunks are connected to the information at token level. Analogously, the clause and sentence segmentations as well as the corresponding alignments are transformed into tables connected to the token tables in the MySQL database. This type of storage offers an easy and fast method to query the corpus. Additionally, a query interface with a menu-like, predefined set of queries can be connected to the database, also allowing non-experts to query the corpus.

4 . Findings

4.1 Information Structure in German and English-German Translations

The CroCo corpus is used to study and compare linguistic phenomena both from a cross-lingual and a monolingual perspective using original and translated texts. This has been done for grammatical functions in theme (i.e. sentence-initial) position as table 1 illustrates. The figures have been computed for the register SHARE.

	subj	obj	compl	adv	verb	other
EO_SHARE	63.43	0.15	0.15	27.14	0.80	8.35
ETRANS_SHARE	64.20	0.19	0.45	27.13	0.25	7.77
GTRANS_SHARE	55.47	2.42	0.22	36.08	0.51	5.29
GO_SHARE	50.25	8.46	1.70	31.00	1.20	7.39

Table 1: Grammatical functions in theme position (in percent).

Focussing on the grammatical functions subject (abbreviated „subj“) and adverbials („adv“), the quantitative figures confirm the widespread assumption that English has a stronger tendency than German to put the subject in theme position. The proportion of subjects in sentence-initial position in EO is more than 13 percentage points higher than in GO. The figures suggest a general tendency in German SHARE texts to vary the function located in sentence-initial position. This can be attributed to language-typological peculiarities of mapping the grammatical functions on semantic roles in the two languages involved (Hawkins, 1986). English is more restricted as to the location of the subject, but the subject can accommodate various semantic roles more easily than German. Conversely, German is more flexible as to which element goes first in the sentence, but requires different grammatical functions to reflect the various semantic roles.

Both, the human translator and the MT system, have to accommodate these differences in the translation. There are two possible solutions for cases, where a one-to-one translation is not possible. Either (1) the order of the grammatical functions remains constant and the semantic content of the original is moved to a different grammatical function or (2) the linear precedence of the semantic content is kept and the order of grammatical functions is changed.

To retrieve the strategy preferred by human translators, we query the source sentence subject chunk in combination with the word alignment. Where the semantic content is not part of the target sentence subject chunk, the word alignment points to a different grammatical function. At present, the results have a low precision and recall rate and can therefore only be seen as a first indication.

Two findings (cf. Kast, 2007) seem particularly interesting: In the translation direction German-English, the lexical content of subjects is often shifted to direct objects.

GO: Auch im Berichtsjahr setzte [die SAP] ihre bewährte Politik des offenen und intensiven Meinungs- und Informationsaustausches fort.
ETrans: [1994] saw SAP continue to pursue its proven policy of open and intensive exchange of information and values.

Here, the translator has chosen solution 2: The subject in GO (in squared brackets) is located after the verb, a position that is not easily accessible to the English subject. Consequently, the perspective is changed in the translation with the temporal information now in the subject and the former agent “SAP” now a direct object (underlined), thus leading to a modulation (see section 2.1).

The translation direction English-German highlights a shift from subject in EO to adverbial in GTrans.

EO: [Day 2] covered new thinking in Globalization, Six Sigma and Product Services.
GTrans: Am zweiten Tag widmete [man] sich dem Gedankenaustausch und neuen Ideen zu den Themen Globalisierung, Six Sigma und produktbezogene Dienstleistungen.

Again, solution 2 seems to be the preferred one: Rather than changing the precedence of the semantic content, the translator chose to map the content on another function that is more amenable to temporal information in German, namely an adverbial.

These initial findings point to a preference in human translation to preserve information sequencing while varying the mapping of grammatical functions, thus accepting a change in perspective. This result can be used in the development of MT systems when aiming at producing a more natural output.

4.2 Part-of-speech Distributions and Shifts

As on the level of chunks, parts-of-speech reflect clear differences between the two languages as can be seen from the comparable corpora displayed in table 2. Both, the reference corpora (ER and GR) and the register-controlled corpora (EO_ and GO_SHARE) show divergences that require handling during translation. The interpretation of these divergences, however, is not always straightforward.

	noun	adj	verb	adv
ER	24.60	6.24	15.72	4.63
GR	22.93	9.20	13.04	5.02
EO_SHARE	29.14	6.97	13.83	3.15
GO_SHARE	25.30	10.69	11.64	4.30

Table 2: Part-of-speech statistics in %

Interestingly, we find a higher percentage of nouns in English than in German. One reason for the former observation is a clearly technical one. German compounds are written in one word (e.g. “Gerichtsentscheidung”), whereas the parts of English compounds are mostly separated (e.g. “court decision”). The POS tagger does not decompose compounds, so where a compound containing two or more nouns is only counted once for German, each part is counted separately in English.

Furthermore, the proportion of verbs seems to be higher in English originals than in the German comparable texts. This divergence can be observed in the contrastive reference corpora as well as in the register-controlled corpora. Rather than for technical reasons, this seems to be a genuine contrastive difference between the two languages, that can be expected to have an effect on translation in the form of transpositions (see section 2). Transpositions can be retrieved from the corpus by querying for an aligned word pair with different part-of-speech tags.

Table 3 illustrates the frequency of the different transpositions for both translation directions, taken from SHARE.[2]

Type of shift	English-German	German-English
verb-noun	24.31	16.98
verb-adjective	11.69	2.80
verb-adverb	6.95	0.25
adjective-noun	17.43	9.48
adjective-verb	1.84	9.92
adjective-adverb	1.42	11.58
noun-adjective	13.89	21.63
noun-verb	5.74	16.98
noun-adverb	3.40	1.08
adverb-adjective	10.06	1.34
adverb-noun	3.05	1.59
adverb-verb	0.21	6.36

Table 3: Frequencies of transpositions in %

For this sub-corpus, we have a total of 40,090 English-German aligned lexical word pairs, among which 1,411 (3.52%) shifts are found, and 37,694 German-English aligned word pairs with 1,572 (4.17%) shifts. Comparing the types of shifts, we can generalize that we find more verb to x alignments for English-German, but fewer x to noun alignments and more noun to x alignments for German-English. This means that English translations are less nominal than their German originals. The following excerpt is taken from the English-German verb-noun list and displayed as follows: original – pos ### translation – pos.

do	-	vd0	###	Handeln	-	nn
play	-	vv0	###	Spielen	-	nn
work	-	vv0	###	Arbeiten	-	nn
programming	-	vvg	###	Programme	-	nn
communicate	-	vv0	###	Kommunikation	-	nn
believe	-	vv0	###	Auffassung	-	nn
computing	-	vvg	###	Computers	-	nn
compared	-	vvn	###	Vergleich	-	nn
learn	-	vv0	###	Lernen	-	nn
enters	-	vvz	###	Schwelle	-	nn
integrate	-	vv0	###	Integration	-	nn
develop	-	vv0	###	Entwicklung	-	nn
browsing	-	vvg	###	Browsen	-	nn
manage	-	vv0	###	Verwalten	-	nn
connect	-	vv0	###	Verbindung	-	nn
control	-	vv0	###	Kontrolle	-	nn

[2] The error rate for the part-of-speech tagger is 3.07% for the German subcorpora and 5.09% for the English subcorpora. Tested on a small sample from SHARE, the word aligner reaches 78.1% precision and 62.8% recall. Other influences on precision and recall include problems of mapping the contrastive tag sets. However, these are difficult to quantify.

The following example illustrates these English-German transpositions, which result in nominalizations in the German translation.

EO: Whether you want to <u>communicate</u>, <u>learn</u>, <u>work</u>, or <u>play</u>, the PC can enrich and improve the experience.
GTrans: Ganz gleich, ob Sie ein Hilfsmittel zur <u>Kommunikation</u> oder zum <u>Lernen</u>, <u>Arbeiten</u> oder <u>Spielen</u> benötigen, der PC kann diese Erfahrung eindringlicher und besser gestalten.

The solutions found by the human translators might be mistaken as deficient. In fact, the above example shows an appropriate solution to the difference between English and German in terms of the frequency of infinite constructions. Transpositions can therefore serve as a basis to develop transfer rules in MT systems that handle contrastive differences between the languages involved.

5. Conclusions and Outlook

The paper has presented findings from empirical studies in a German-English comparable and parallel corpus. It has shown that techniques applied for either comparable or parallel corpora can complement each other, providing explanations from the latter for observations made using the former corpus. The findings from the analyses presented here demonstrate that solutions chosen be human translators which appear to deviate from the source text must not necessarily be defective. They can rather be viewed as a valuable resource for creating a more natural output of MT systems taking into account contrastive differences in language use. These can only be identified with the help of a comparable corpus. The findings encourage further investigation into language contrasts, translations shifts and translation strategies.

For the application in MT, the use of corpora - and thus empirical resources for language contrasts, translation shifts and translation strategies - is expected to be more dynamic than rule-based approaches. The combination of linguistic corpus enrichment and the extracted translation shifts allows compiling a comprehensive set of transfer rules for MT systems, ideally evaluated on the basis of translation models from translation studies. With this approach, existing translations serve as a basis for solving translation problems thus making the MT output more similar to human translation.

6. References

M. Baker. (1993). Corpus linguistics and translation studies: implications and applications. M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair*. Amsterdam & Philadelphia: John Benjamins, pp. 233-250.

J. Barnett, I. Mani, P. Martin & E. Rich. (1991). Reversible machine translation: What to do when the languages don't line up. *Proceedings of the Workshop on Reversible Grammars in NLP*. ACL-91, Berkeley, pp. 61-70.

T. Brants. (2000). TnT - A Statistical Part-of-Speech Tagger. *Proceedings of ANLP-2000*, Seattle.

J.C. Catford. (1965). *A Linguistic Theory of Translation*. Oxford University Press, Oxford.

M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič & V. Kuboň. (2004). Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. *Proceedings of LREC 04*, Lisbon.

C. Copeland, J. Durand, S. Krauwer & B. Maegaard. (1991). The Eurotra linguistic specifications. *Studies in MT and NLP* Vol. 1, Brussels.

L. Cyrus. (2006). Building a resource for studying translation shifts. *Proceedings of LREC 06*, Genoa, pp. 1240-1245.

B. J. Dorr. (1994). Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics* 20:4, pp. 597-633.

D. Farwell, S. Helmreich, B. J. Dorr, N. Habash, F. Reeder, K. Miller, L. Levin, T. Mitamura, E. Hovy, O. Rambow & A. Siddharthan. (2004). Interlingual Annotation of Multilingual Text Corpora. *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, Boston, pp. 55—62.

S. Granger, J. Lerot & Stephanie Petch-Tyson (eds.). (2003). *Corpus-based approaches to contrastive linguistics and translation studies.* Amsterdam & New York: Rodopi.

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

M. Heyn. (1996). Integrating machine translation into translation memory systems. *European Association for MT, Workshop Proceedings*, ISSCO, Geneva, pp. 111-123.

E. W. Hinrichs, J. Bartels, Y. Kawata, V. Kordoni & H. Telljohann. (2000). The VERBMOBIL Treebanks. *Proceedings of KONVENS 2000 Sprachkommunikation*, ITG-Fachbericht 161, VDE-Verlag, pp. 107-112.

M. Kast. (2007). *Variation innerhalb der grammatischen Funktion "Subjekt" bei Übersetzungen Englisch-Deutsch und Deutsch-Englisch.* Unpublished diploma thesis. Saarbrücken: Applied Linguistics, Translation and Interpreting, Universität des Saarlandes.

S. Kinoshita, J. Phillips & J. Tsujii. (1992). Interaction between structural changes in machine translation. *Proceedings of COLING 92*, Nantes, pp. 679-685.

J. Lindop & J. Tsujii. (1991). *Complex transfer in MT: A survey of examples*. Technical report, CCL/UMIST 91/5, Manchester.

H. D. Maas. (1998). Multilinguale Textverarbeitung mit MPRO. *Europäische Kommunikationskybernetik heute und morgen 98*, Paderborn.

C. Müller & M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. S. Braun, K. Kohn, J. Mukherjee (eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Peter Lang, Frankfurt. pp. 197-214.

S. Neumann & S. Hansen-Schirra. (2005). The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations. *Proceedings of the Corpus Linguistics Conference Series* Vol. 1 no. 1.

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

F. J. Och & H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Journal of Computational Linguistics* Nr.1, vol. 29, pp. 19-51.

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

A. Schiller, S. Teufel & C. Stöckert. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. *Technical report, IMS, University of Stuttgart, Seminar für Sprachwissenschaft*, University of Tübingen.

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

K. van Leuven-Zwart. (1989). Translation and Original. Similarities and Dissimilarities, *Target* 1(2), pp.151-181.

J.-P. Vinay & J. Darbelnet. (1958). *Stylistique comparée du français et de l'anglais*. Les éditions Didier, Paris