*Josef Schmied**

Translation and Cognitive Structures[1]

Abstract

This project is based on a corpus of English and German source and target texts, ranging from contemporary literature to scientific textbooks. We try to create a machine-readable and aligned corpus which will allow us to discover and categorize translation equivalents for a number of linguistic items, such as prepositions, subordination, deictic elements, metaphors or culture-specific structures. On this basis we look for regularities in the configuration of factors that influence equivalent choices for each of the phenomena in question. Apart from theoretical insights into contrastive language structures as well as cognitive aspects of the translation process, the purpose of the project is to discover and categorize prototye and non-prototype equivalents in two closely related languages. Research results could, for instance, be applied to bilingual lexicography or other language learning and translation aids.

1. A new approach to contrastive and cognitive issues: an example

German and English are genetically and typologically closely related languages. Thus language learners and bilingual language users (e.g.

[1] The project group consists of a team at Chemnitz (led by Josef Schmied) and a team at Cambridge (led by Kirsten Malmkjær) and it works parallel to other translation workgroups (cf. Johansson/Hofland 1994). The basic research design was agreed on in February/March 1993, when I worked as an honorary research fellow at the *Research Centre for English and Applied Linguistics* in Cambridge. I have to thank all supporting agencies that have enabled me and my colleagues to pursue the project so far (the German Research Association/DFG, the British Council and the Technical University of Chemnitz-Zwickau), all the colleagues who have assisted and encouraged us (Gillian Brown, Cambridge; Michael Claridge, Chemnitz; Elke Esders, Brussels and Paul Procter, Cambridge) and all the publishing houses and agencies that have contributed texts (sometimes even in machine-readable form).

Although the basic research framework is the same for both groups, the Chemnitz team is specialized in corpus collection, whereas the Cambridge team adds their translation experience. Analyses will be carried out parallel (in each case source- and target-language specialists have to work together), but each team will be responsible for separate sections of the comparison.

* *Josef Schmied*
Englische Sprachwissenschaft
TU Cemnitz-Zwickau
Postfach 964
09009 Chemnitz /D)

translators) can in many cases start from the assumption that structures are parallel. Thus it appears fairly safe to expect a German *in* as an equivalent for an English *in*. This crude juxtaposition however does not consider external and internal boundaries of the underlying cognitive[2] principles, i.e. the basic status of *in* in the respective prepositional system and the polysemous subcategories (including the respective quantitative weight). Therefore, most dictionaries would also give at least *into* as an English equivalent for directional German *in* (the choice is normally easy for the dictionary user as the form is usually disambiguated by case morphology in German; local *in* entails dative, directional *in* accusative inflection). However, this only applies for prepositional English *in*, adverbial *in* has various other equivalents in German (*herin*, *hinein*, *herein*, *ein*, etc.). Thus it would be interesting to analyse all cases where German *in* is not rendered by English *in* on the most general level of analysis. As we have seen, this example is subcategory- and word-class sensitive, thus we can establish a lower, more detailed level of analysis where *in*, even local *in*, is not rendered by *in* in English, but by *at* for instance - and exclude adverbs from the analysis. For other prepositions this juxtaposition of structures can be easier - for less polysemous prepositions (e.g. English *besides* is less complex than German *neben*), or more difficult for more grammaticalized ones (e.g. English *of* is more complex than German *von*). In addition to these increasingly specific semantic levels (general local prepositional usages versus two-dimensional local ones), lexeme-specific idiomatic structures tend to complicate the matter, as they are by definition larger structures that are unmotivated by the individual meaning components and often very language- and culture-specific.

The term culture- and language-specific contrasts does not only apply to the famous untranslatables (like *gentleman*) and false friends (like German *Stuhl* [English *chair*] and English *stool* [German *Hocker*]), but also to more pervasive conventions like politeness and tentativeness, which are pragmatically and semantically defined but formally very difficult to grasp. These examples illustrate how a) speci-

[2] I am using the term cognitive in two meanings in this contribution: first with reference to the translation-specific cognitive processes and second with reference to the language-independent cognitive space on which language structures are mapped. This is not unusual in an age where everything seems to drift towards cognitive linguistic approaches, though very different types. The interesting question is in what way the different cognitive perspectives are related.

fic forms in each language (e.g. *in*) cover partially overlapping areas of the cognitive spaces (e.g. location, direction, etc.) and how b) the results can be used to set up finer context-specific equivalents for improving (human) translation teaching and for improving (semi-)automatic machine translation programs. Thus comparative studies based on real language in translations can be used for theoretical and practical purposes.

2. Compilation and categorization of corpus material

2.1. General principles

The text collection will consist of an English-into-German and a German-into-English translation part, although we have to bear in mind that it may not be possible to find parallel texts for both directions for all categories, criteria for textual equivalence are not very strict, however (cf. 3.2 below).

The choice of text-types for a translation corpus (cf. Appendix 1) has to be governed by theoretical considerations, i.e. in which areas of language usage, genres/text-types translations are particularly important or functional. As we wanted to exclude learner language and interference phenomena as far as possible, we only took translations that were not made for teaching and testing purposes (i.e. from schools), texts that could be assumed to have been translated “for a real purpose” carefully and by professionals, so that the quality of the work could remain unquestioned up to a certain point. This lead us to three types of public translation domains:

a) international publishers (and newspapers), where internationally interesting books and articles from one language and culture are transferred into another language or culture,

b) international agencies, where information has to be distributed to various national agencies, and

c) bicultural institutions, where texts from one target culture have to be translated specifically for a readership in another target-culture.

When we thought of various modern access routes to translated (preferably machine-readable) texts we arrived at similar results. We thought of materials from book publishers and newspapers (e.g. translations of German newspaper articles from *Die Zeit* in the *Guardian*

Weekly) or from multilingual international (standard examples are agencies of the European Union in Brussells) or bilingual national agencies (e.g. texts distributed abroad to propagate national interests and cultures in a foreign country).

Because most comparative studies, contrastive and translation, have (in the past?) been carried out using literary texts, our core corpus will consist of non-literary texts, although a subcorpus of literary texts will be compiled for comparative purposes. The latter may be used to highlight the special nature of literary language; in some cases it can reflect "real" language, depending on the writer's "linguistic" perception, in others it includes all the genre- and culture-specific literary conventions that may force the author to deviate from naturalness, in which case it may be more creative and metaphorical, the author testing the flexibility of a language, stretching it a little further than "normal".

The most general criterion for categorization and corpus compilation was the attempt to cover a broad variety of texttypes from technical to literary language, as the composition of a corpus depends on its intended uses (Johansson 1978:i), texttype is one of the basic variables in the linguistic analysis of the corpus, so that questions about language- or culture-specific technical or special languages could be addressed.

2.2. Text-type specific problems

Texts translated for a mother-tongue readership are usually available through publishers in the source- or target-language country. So far we have started collecting excerpts[3] from scientific monographs from different (academic) subjects. The subjects range from pure to applied, from natural to social sciences, so that linguistic features could be correlated and genre-specific clines or subject-specific structures isolated. In these cases we aim at collecting 20,000 word excerpts from at least 3 books for each direction per subject category.

Some translations in our corpus come from the European parliament in Brussels and have their own specific problems. As many of the diplomats and the translators there are multilingual they do not necessarily write in their mother tongue. The special problem of English texts is

[3] As we have some entire books in machine-readable form we will be able to do some quick quantitative studies on the representativeness of our extract and thus contribute to the general corpus-linguistic discussion.

that many non-native speakers write in English in order to make their ideas more directly or quickly accessible to their colleagues and many non-native speakers translate into English (just as many native speakers of the so-called smaller European languages use the English version rather than the original as a source text, e.g. when a Greek translator does not know Danish he uses the English translation). A further complication is that political texts are often difficult to categorize for subject, although we would consider it desirable to have parallel categories (e.g. economics) for scientific and political texts. Unfortunately, many political speeches are a heterogeneous mixture of politics, law, economics and culture - to say the least.

A different type of translations are texts on the source-language country produced for foreigners in the target-language country in their language. The British Embassy in Bonn, for instance, distributes a publication entitled “Britische Dokumentation”, a translation of a selection of culture-specific texts that are assumed to be of interest to a foreign, usually professionally Britain-related or anglophile readership. Here through the choice of special topics intercultural and cultures-specific peculiarities can be expected to play a more prominent role than in other texts. A German equivalent to this English-into-German translation is more difficult to find, the German Foreign Office does not invest in a comparable periodical, its periodicals in English are less target-culture specific and cater for a more international readership - in other words it does not use English as culture-specific national language for communication with Britain specifically, but as an international language for communication with “the world at large”.

A different problem arises with tourist brochures: regional tourist boards in Germany and Britain often seem to prefer unprofessional do-it-yourself strategies, so that the target-language versions are unsatisfactorily rendered from a native-speaker perspective (cf. 5.2 below).[4]

Many of these compilation problems may be due to culture-specific and text-type specific traditions. The fact that corpus texts are culture-specific and that it may be difficult to find equivalents in a different speech community has been mentioned in particular in connection with parallel variety corpora, such as the Brown-LOB-Kolhapur family

[4] A more technical problem is that the latter categories (Embassy publications and tourist brochures) are generally shorter than 20,000 words, so that more categories from a different region (and possibly by different writers) have to be combined.

(Shastri 1988) or the *International Corpus of English* (Schmied forthcoming). These difficulties arise less in a German-English corpus than in many others involving countries where English occupies almost second language status and languages that are characterized by a unilateral translation process; thus whereas it is no problem finding modern literature translated from Norwegian into English the number of computer handbooks is very limited, because the developers would write directly in English, assuming that any Norwegian user would know English anyway and few non-Norwegian users would know Norwegian. On the other hand, English originals would not be translated into Norwegian either.

3. Analysis of corpus material

3.1. The theoretical framework

The theoretical framework for this project can be drawn from three different linguistic perspectives. The study on philosophy of language have been raising the question of translatability for a very long time, a corpus-based analysis can shed new light on an old issue. Modern translation studies are interested in the product- as well as in the process-oriented approach, modern cognitive linguistic studies seem to converge with this "constructive" view of translation in so far as it concentrates on the processes that construct mental concepts between language and the real world. The latest expansion of contrastive studies led to the typological comparison of language structures (particularly Hawkins 1986 for English and German) emphasizing possible and actual developments of language structures. Now the four threads are combined: Non-prototype form equivalents indicate where the translator felt a need to deviate from the staightforward routine because the formal equivalent would not render the appropriate effect among the target readership. This indicator is our starting-point for the analyses of the fuzzy edges of our cognitive structures in language[5]. This product-oriented analysis leads to reflections upon the process of translation and the identification of the interestingly different structures in the source

[5] Because the Chemnitz-Cambridge collaboration intends to exploit only the results for the empirical description of English and not of German, the German material may be underused. It will however be made available to other researchers for scholarly use as early as possible.

and and target texts. Thus we are not only concerned with texttype-specific translation strategies but also with the regularities in the configuration of factors which influence equivalent choices for each of the language phenomena investigated contrastively, taking into account cotextual as well as contextual factors (such as the scope of translation).

3.2. Linguistic categories

The linguistic categories that will be investigated range from the more form-oriented to the function-oriented side (cf. Appendix 2). Although the quantitative side of the analysis requires more surface-oriented approaches, partial tagging can help to combine items that are semantically related, be it in semantic fields (e.g. prepositions covering the cognitive space of directionality), be it in cohesive reference chains (e.g. anaphoric and cataphoric relationships). Two examples of more sophisticated features may serve as examples for addressing cognitive questions by surface-oriented tools.

A well-known example of linguistic and cultural differences between German and English is modality. The prototype view is that both languages have a similar and related system of modal auxiliaries, but historical analyses (e.g. Lightfoot 1979) have shown that the strategies of expressing epistemic modality have been developed relatively late over the last few centures and more so in English than in German. Nowadays the fact that modal auxiliaries have been grammaticalized to a greater extent more in English than in German can be seen from their greater deficiency in verbal features (e.g. expressing past time reference and attracting an object). Correspondingly, a much larger part of modal functions of auxiliaries has to be taken over by adverbials in German. This is a typical corpus-linguistic problem. Only a quantitative context-sensitive analysis can verify this statement, the structural options being the same but the preferences different.

3.3. Problems of data analysis: validity, alignment and retrieval

For translated text collections the general corpus-linguistic problem of data validity applies: Can all data be taken as acceptable or do we have to use (near)native-speaker intuition to judge at least single or author-specific occurrences? This may not be an important issue for quantitative analyses, but when fine-grained analyses of gradient structures are undertaken it becomes an important issue. In addition, the problem of interference from the original presents a special challenge. The language-contact situation involved in translations allows us to see the same contrastive structures in a different light from either perspective: the researcher analysing structures from his mother tongue might be able to suspect interference, the researcher analysing structures into his mother tongue might be able to detect unnaturalness. What we are interested in is the unconscious twisting of target texts through source-text structures. The big question is what interference contributes to translators' jargon, translationese, whether translations are somewhere between the more "natural" texts from parallel corpora, or whether stereotyped notions of target-language structures occur more frequently than in natural texts. English-specific forms like cleft constructions or continuous forms are inserted by translators in order to increase the acceptability for the target-language readership[6].

A major developmental problem is presented by the necessary software tools, which have to consist of three components: an alignment program, a retrieval program and a statistics program.

Alignment programs are often non-language specific, based on punctuation marks and average (and comparatively adjusted) word lengths (German words are usually longer than their English equivalents, because of morphosyntactic suffixes). Language-specific alignment programs usually work with lists of corresponding anchorwords. In fairly idiomatic translations however the vocabulary as well as the word length and punctuation may differ quite a lot. Thus a combination of purely mathematical and language-specific parameters may be desirable. Therefore a list of anchorwords has to be drafted and tested. The problem with anchorwords is that they should be specific enough to occur regularly in the corresponding text and frequent enough to work in as many cases as possible. Unfortunately, grammatical morphemes such as articles occur often enough, almost too often, but their occurrence or omission follows some very language-specific rules. Lexical morphemes do not occur often enough, although they would be safe cases as they can hardly be omitted, even if they can be

[6] In terms of language acquisition, the question is whether translators maintain strategies of overrepresentation, which are characteristic of advanced learners, compared to the common underrepresentation of intermediate learners.

rendered in the target language by various synonyms (or even paraphrases?). Clear cases are usually numerals, where we have to consider however that they may occur in digits or in words (7 or seven) and that enumerations may be rendered by other symbols (e.g. letters). In many cases a limited set of words (about 200) should be sufficient, if problem clauses are corrected by hand.

For the retrieval it would be desirable to have a Word-Cruncher type program with interrelated search and statistical facilities in two aligned windows. This would enable us to search in one text version for a certain string and in the corresponding aligned text immediately underneath for an equivalent. The two structures should be visible directly below each other (like an interlinear text) in a key-word-in-context window or in a full-text window, so that one could compare prototype and non-prototype equivalents immediately. These search procedures should be linked with the Boolean parameters AND, OR, NOT.

Furthermore, it should be possible to measure prototype versus non-prototype equivalents on various levels: using the above example, we would need a crosstabulation that shows us the absolute or relative frequency of German *in* correlating with English *in* and the most important other forms. This would, for instance, allow us to decide on the relative overlap between German and English prepositions. Jumping back into the texts we would be able to check whether the co-text or the context betray text-type specific or collocational clusters. From the quantitative perspective there are three levels of analysis, the individual texts, the text-types and the corpus as a whole.

4. Practical applications of research results

4.1. Improving bilingual lexicography

Bilingual dictionaries are the basic traditional tool of bilingual speakers, including translators. But whereas modern English monolingual learner dictionaries have made enormous progress in recent years, bilingual dictionaries have hardly participated in these developments. A major basis for this rapid development was the use of modern computer equipment, including large databases and efficient retrieval tools, for the choice of lexemes, the extraction of examples, the indication of collocations and the syntactic patterns, etc. (the COBUILD publications are the most famous example for marketing “real English” nowadays). Taking over modern lexicographic research techniques from monolingual to bilingual dictionaries is therefore a logical and necessary step in this development. A translation corpus can provide the empirical bases for translation equivalents in the broadest sense. The special emphasis on cultural contrasts in our studies coincides with the recent development of culture-specific lexicography, which breaks down the traditional borders between a dictionary and an encyclopedia (cf. the recent *Longman Dictionary of Language and Culture*). The cultural perspective can however also be applied to pragmatic and syntactic levels of language analysis. Just as the grammatical information is attached to individual lexemes in the bilingual dictionary, text-type specific contrastive information could be added. Most of this information is however of a gradient nature, especially in closely related languages like English and German. Thus even advanced dictionary users are torn between the fear of interference and translationese, either over- or underusing source-language structures in target-language texts. More paradigmatic and syntagmatic information would provide useful guidelines here. Paradigmatic information could include the relative frequency of near-synonyms and of hyper- or hyponyms in specific text-types, syntagmatic information would be provided in list of collocates and clause pattern wherever the slightest contrast can be detected (analogue to the monolingual BBI dictionary). Besides cotextual, wider contextual factors of regional, social, register, domain- or text-specific marking should be considered consistently. In addition, the subcategorization of meanings could be improved, such as prepositions, and notes on specific usage could be inserted at “neuralgic points” such as false friends.

4.2. Providing an empirical basis for translation textbooks and translation programs

The results of these analyses could be used in textbooks on translation where, so far, prototype standard examples abound. Introspective examples are invented to illustrate rules and categorizations given in standard textbooks for translation. These have to be checked against corpus material, where ambiguity and indeterminacy create more fuzzy edges than clear cases. But even clear cases are not always recognized, because prototype views tend to change very slowly. Famous examples for the divergence of old versus new prototype collocates are *run*, which is used transitively (with objects like *shop*) more often than intransitively, or *mouse,* which collocates with *click* more often than with *catch* or occurs more often in the domain computer than household.

Infamous examples of prototype views can also be found in computer dictionaries attached to translation programs, where often any cotext variables are missing.

4.3. Limitations

Finally, it has to be admitted that there are some specific limitations to the application of (individual language structures from) translation corpora.

First, in many cases the target texts in the corpus represent different levels of adequacy. This leads to difficulties in finding regularities in the configuration of factors which influcence the choice of a specific translation equivalents, especially if some of the texts contain inadequate (not necessarily wrong) translations. The solution of using translations of the same text by different translators is not possible in our case as they do not exist in translation reality (except in translations of very famous literary works and in the testing situation, which we excluded on principle).

The second major issue is how to account for extralinguistic, contextual factors (which mainly apply to individual texts) when creating a matrix of factors which influence the choice of a translation equivalent. Many of these factors are inaccessible for the compiler of a translation corpus who has access only to the finished product and cannot use the translator as an informant.

Finally, it has to be borne in mind that a translation corpus is difficult to use as a basis for EFL-applications. Not only because of the inadequate translations mentioned, but also because real language data are not adapted or graded according to the learner-specific language-learning process.

These problems lead to two conclusions: Whenever qualitative results of corpus-based translation studies are used, the analysis has to be combined with a translation critique (as one branch of translation studies). But even here inadequate translations can possibly be identified not only on the basis of introspective knowledge but also on the basis of quantitative, comparative corpus analysis. In our case, however, the results are used for a more adequate typological description of English, where individual deviant cases are of no significance.

References

Hawkins, John A. (1986): *A Comparative Typology of Englsh and German*. London: Croom Helm.

Johansson, Stig (1978): “Manual of information to accompany the Lancaster - Oslo/Bergeb Corpus of British English, for use with digital computers”. Oslo.

Johansson, Stig/Knut Hofland (1994): “Towards an English-Norwegian parallel corpus”. In: Udo Fries et al., eds. *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*, Zürich 1993. Amsterdam/Atlanta, GA: Rodopi, 25-37.

Lightfoot, David (1979). *Principles of Diachronic Syntax*. Cambridge: Cambridge U.P.

Schmied, Josef (forthcoming). “Sociolinguistic Problems of Second Language Corpora”. In: Sidney Greenbaum, ed. *The International Corpus of English*. Oxford University Press.

Shastri, S.V. (1988): “The Kolhapur Corpus of Indian English and work done on its basis so far”. In: *ICAME Journal* 12, 15-26.

APPENDIX 1: Text-types for a German - English Translation (GET) corpus (with provisional size)

	size
1. Scientific textbooks (3 books, 2 directions, 20.000 words each)	
a) Physics	120,000
b) Engineering and Computer Science	120,000
c) Economics	120,000
d) Religion and Philosophy	120,000
e) History	120,000
	[600,000]
2. Newspapers	100,000
3. EU-texts	
a) economy	100,000
b) aspects of society and culture	100,000
	[200,000]
4. Publications of the British Embassy in Bonn	100,000
5. Tourist brochures	100,000
TOTAL	1 million
subcorpus:	
6. Literature (contemporary British/German literature)	200,000

APPENDIX 2: Linguistic categories investigated

- anaphora, cataphora, exophora
- deixis
- thematisation
- transitivity
- modality
- prepositions
- de-lexicalised function verbs
- non-finite constructions
- ing-constructions/nominalisations
- culture-specific structures (epistemic modality, metaphors, lexicon)