Cohesion and explicitation in an English-German translation corpus

Silvia Hansen-Schirra, Stella Neumann & Erich Steiner

In translation studies, cohesive features as indicators for explicitation have been analysed either in an example-based way (Blum-Kulka 1986) or as concordances in monolingually comparable corpora of raw text (Olohan & Baker 2000). Where this is done without taking into account the source texts, the interpretation of explicitation remains restricted and problematic. Work on translations against a more linguistic background has addressed some of these restrictions and problems (cf. relevant work as in Johansson & Oksefjell eds. 1998; Fabricius-Hansen 1999); the focus of these research interests and methodologies is however different from ours.

The basic assumption for the analysis of explicitation in the present paper is that the element that is explicitated in the target text has to be present implicitly in a linguistically traceable way in the source text and vice versa for implicitated elements. Explicitation is thus defined as a relationship and a process between instantiated and aligned pieces of translated texts. Furthermore, we stratify the notion of explicitation according to the linguistic levels of lexicogrammar and cohesion. As this stratification is still too abstract to be directly quantified on linguistic data in an electronic corpus, a series of further micro-level operationalisations is undertaken which are meant to bring the relevant phenomena down to an empirically measurable level.

For this purpose, our investigation of explicitation and implicitation of cohesion markers in translations is based on a cross-linguistic corpus containing statistically meaningful and representative samples (cf. Biber 1993) of German and English parallel texts from 8 registers annotated with parts of speech, morphology, phrase structure and grammatical functions. Altogether, the corpus comprises 1 million words plus 68,000 words in register-neutral reference corpora in both languages. A characteristic feature of this corpus is the alignment of source and target texts on different linguistically motivated layers: we not only align sentences (which is state of the art in Translation Memories; e.g. Johansson et al. 1996) and words (which is state of the art in Machine Translation; cf. Och & Ney 2003) but also clauses and syntactic functions.

A methodological principle for the compilation of the resource is the distinction between strictly lexico-grammatical annotation of source and target language texts including the alignment of these annotations on the one hand, and the interpretation of the data in view of more abstract concepts like “explicitation” on the other. This distinction allows us to pose queries on (combinations of) lower level linguistic features assumed to be indicators of the more abstract concept. One technical precondition for the comprehensive analysis of the corpus is the use of XML stand-off mark-up as representation format for annotation and alignment. This is necessary because we annotate the corpus on different layers, thus keeping the annotation and alignment of overlapping units in separate files. Thus it becomes possible to view the annotation in aligned segments and to pose queries (using XSLT) combining different layers (cf. Neumann & Hansen-Schirra 2005), which is essential in order to get meaningful information on the workings of explicitation in texts. The resource thus permits the analysis of a wealth of linguistic information on each level helping us to understand the interplay of the different levels and the relationship of lower level features to more abstract concepts such as explicitation.

In the present paper we will exemplify the queries possible on the basis of the annotation and alignment for the cohesion markers described by Halliday & Hasan (1976) and their equivalents for German. It is, for instance, a straightforward step to retrieve (co-)reference markers separately from the source and target language corpora. The part-of-speech information contained in the two corpora permits precise queries. Specific queries into these reference markers in the target texts which have no equivalent in the source texts are more complex. However, they have more explanatory power than queries of the type Olohan and Baker describe. We will also address even more complex questions like the following: Where do reference markers tend to occur within a sentence? Are they typically realised in preferred syntactic functions? Do these occurrences differ when comparing source and target language texts?

Another interesting phenomenon is ellipsis in translations, since it is a direct indicator for implicitation. The alignment of our corpus on different layers means that ellipsis can be found through empty alignment links on the word level as well as on the level of syntactic functions. On the basis of these empty links, we can furthermore investigate which syntactic functions ellipsis occurs in, whether it prefers finite or non-finite clauses or how it is dealt with in different translations directions.

In the paper we will show how to query the annotated and aligned corpus in order to identify the above mentioned and other cohesion markers. We will discuss query results in a crosslinguistic perspective and draw conclusions on explicitation and implicitation in translated text. The long-term aim of the present study is then to interpret quantitative (co-)occurrences and patterns found in translations and their source texts as indicators of explicitation/ implicitation against the background of three sources of explanation: language typology, text typology and the translator's language processing. Our methodologically motivated separation of linguistic annotation/alignment from their interpretation in pursuing a research question (in our case explicitation) makes the corpus resource flexible enough to allow research into other phenomena of interest in connection with translation, such as simplification, normalisation, shining through (cf. Teich 2003) etc.

References

Biber, D. 1993. Representativeness in Corpus Design. In *Literary and Linguistic Computing* 8/4, 243-257.

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

Fabricius-Hansen, C. 1999. Information packaging and translation: Aspects of translationalsentence splitting (German – English/ Norwegian). In Doherty, M. (ed.) *Sprachspezifische Aspekte der Informationsverteilung*. Berlin: Akademie-Verlag, 175-214.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.

Johansson, S., Ebeling, J., & Hofland, K. 1996. Coding and Aligning the English-Norwegian Parallel Corpus. In Aijmer, K., Altenberg, B. & Johansson, M. (eds.) *Papers from Symposium on Text-based Cross-linguistic Studies*. Lund, 87-112.

Johansson, S. & Oksefjell, S. (eds) 1998. *Corpora and Cross-linguistic Research*. Amsterdam: Rodopi.

Neumann, S. & Hansen-Schirra, S. 2005. The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations. In *Proceedings from the Corpus Linguistics Conference Series*, vol. 1, no. 1, ISSN 1747-9398.

Och, F. J. & Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, vol. 29, no. 1, 19-51.

Olohan, M. & Baker, M. 2000. Reporting *that* in Translated English. Evidence for Subconscious Processes of Explicitation? In *Across Languages and Cultures* 1(2), 141-158.

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