Quantitative and qualitative analyses of explicitation in translations

Stella Neumann & Silvia Hansen-Schirra
Universität des Saarlandes
st.neumann@mx.uni-saarland.de, hansen@coli.uni-sb.de

Introduction

This paper presents research on an assumed property of translations according to which information is expressed more explicitly in translated texts than in originals. We combine the interpretation of quantitative linguistic information obtained from a large corpus of English and German originals and translations with qualitative analyses of smaller subcorpora.

Explicitation in translation studies

Blum-Kulka (1986) formulates the hypothesis that explicitation is a characteristic phenomenon of translated versus original texts on the basis of linguistic evidence from individual sample texts showing that translators explicitate optional cohesive markers in the target text not realised in the source text. This is an important exploratory step to describe properties of translated texts which make translations distinct from originals in the same language. However, a characteristic of a class of texts, i.e. a property, can only be identified on the basis of statistically relevant linguistic evidence. Corpus linguistic methods can be applied to investigate the assumed properties of translations. Building on Blum-Kulka's work, it is assumed that the strikingly frequent use of cohesion markers should result in increased text length of the translations as compared to original texts in the same language (Baker 1996). However, text length is only a very weak indicator.

Olohan and Baker (2000) therefore concentrate on the frequency of the optional 'that' versus zero-connector in combination with the two verbs 'say' and 'tell'. They analyse concordances of the respective lexical strings in a corpus of translations into English as compared to a corpus of English originals. Although Olohan and Baker show that translators use the explicit 'that' significantly more often than authors of English originals do, this finding is limited to the strings they analyse. In the case of 'say' and 'tell', categories of verbs expressing verbal meaning would yield more com-

prehensive findings than selected lexical strings. While being extensive enough for statistical interpretation, corpus-driven research like Olohan and Baker's is limited in its validity to the selected features. More generally speaking, there is a gap between the abstract research object and the low level features used as indicators. This gap can be reduced by operationalising the notion of explicitation into syntactic and semantic categories, which can be annotated in a corpus. Intelligent queries will then produce linguistic evidence with more explanatory power than low level data obtained from raw corpora. The annotation may comprise both automatic and manual analyses: Automatic annotation is suited for processing large corpora and thus for quantitative analyses. Manual annotation is typically used for analyses involving a high degree of human interpretation. As this is very costly, it is better suited for small corpora and thus for qualitative analyses.

Using annotated corpora for the analysis

We combine both types of analyses by applying automatic annotation to a one million word corpus as well as manual annotation to smaller subcorpora sampled from the large corpus. This approach requires a flexible corpus design permitting easy drawing of samples as well as XML stand-off mark-up supporting annotations on different layers with overlapping annotation units. The corpus compiled for this purpose consists of English and German originals and matching translations taken from eight registers as well as a register-neutral reference corpus in both languages. This corpus design allows intralingual and interlingual comparisons within and across registers with the possibility to factor out language typological characteristics as found in the reference corpora. All of the data, i.e. the raw texts, the metadata and the various annotations, are kept in separate files with the annotated elements linked to the indexed corpus through unique IDs. This kind of XML encoding takes advantage of the rich set of tools readily available to edit, validate, transform and query the linguistic annotation across layers. Part-of-speech tagging, morphology and phrase chunking represent the basic annotation of the present research. In order to view the different units in original and translation together, we align words, phrases, clauses and sentences. The corpus enriched with this information covers a wide range of indicators for explicitation and forms the basis for the remaining analyses requiring human interpretation.

The following example combines quantitative and qualitative analyses for the investigation of explicitation. Cohesive ties are very susceptible to explicitation as already shown for connectors by Blum-Kulka (1986) and Olohan and Baker (2000). Anaphoric relations are also useful indicators. Direct anaphoric relations are categorised in recurrences (total or partial), pronominal relations (personal, possessive or demonstrative) and IS-A relations (synonyms, hypernyms or hyponyms). Based on our automatic part-of-speech tagging, i.e. our quantitative analysis, we can search all pronominal relations using the part-of-speech tags for personal, possessive or demon-

strative pronouns. Furthermore, we can also display all corresponding pronominal relations in the source and target texts with the help of the alignment. The mismatches can be found and analysed as well: e.g. pronominal relations in the target text which are, for instance, aligned to total or partial recurrences in the source text (here, string-matching algorithms are applied). In order to analyse IS-A relations, however, deeper linguistic knowledge is required. Neither synonyms nor hypernyms nor hyponyms can be retrieved on the basis of part-of-speech tags or phrase chunks. For this purpose, a coreference annotation is necessary, as it is proposed for example in Kunz and Hansen-Schirra (2003). This kind of annotation depends of the linguistic interpretation of a human annotator and can thus only be applied to small sub-corpora. Using this kind of qualitative annotation, antecedents and the subsequent anaphors can be compared throughout the source and target text. While this linguistic evidence does not allow statistical processing necessary for general statements, it supplements the quantitative data and deepens our understanding of explicitation.

Conclusion

The research described here shows the use of comprehensive quantitative and qualitative annotation for the analysis of explicitation. The fact that the annotation contains as little theoretical bias as possible makes it also suitable for the investigation of other assumed translation properties like simplification or normalisation. It may even be used beyond translation studies in cross-linguistic natural language processing.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers, ed., Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager. Benjamins, Amsterdam, pp. 175-186.

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

Kunz, K. and S. Hansen-Schirra (2003). Coreference annotation of the TIGER treebank. In Proceedings of the Workshop Treebanks and Linguistic Theories 2003. Vaxjo, pp. 221-224.

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