SPRIK Conference
Explicit and Implicit Information in Text - Information Structure across Languages
University of Oslo, June 8-11 2006

Cohesion and explicitation in an English-German translation corpus[1]

Silvia Hansen-Schirra, Stella Neumann, Erich Steiner
Saarland University, Saarbrücken, Germany
hansen@coli.uni-sb.de, st.neumann@mx.uni-saarland.de, e.steiner@mx.uni-saarland.de

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 (cf. several contributions in Laviosa (ed.) 1998, Olohan & Baker 2000). In spite of the insight gained from this line of research, we argue that where explicitation is investigated without taking into account the source texts, the interpretation of results 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, and partly complementary to, ours with respect to corpus architecture, querying techniques and underlying linguistic modelling (for which cf. Hansen 2003, Neumann 2003, Steiner 2001, 2005a,b,c, Teich 2003).

The basic assumption for the analysis of explicitation in the present paper is that the element 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 (not in focus in this paper) and cohesion. As this stratification is still too abstract to be directly quantifiable 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 accessible level.

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 registerially parallel texts from 8 registers annotated with parts of speech, morphology, phrase structure and grammatical functions. In addition to these two sub-corpora, two further sub-copora have been compiled consisting of translations of the samples from the first two sub-corpora into the respective other language, yielding 4 sub-corpora. The overall corpus comprises 1 million words (approx. 250 000 for each of the four sub-corpora) plus 68,000 words in register-neutral (cross-register) reference corpora in both languages. A characteristic feature of our 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.

One of the methodological principles for the compilation of the resource is the distinction between lexico-grammatical/ cohesive annotation of source and target language texts (including

[1] The authors would like to thank the reviewers for helpful comments on an earlier version. The research reported on here is sponsored by the German Research Foundation (DFG) as project no. STE 840/5-1. For current information cf. http://fr46.uni-saarland.de/croco/

the alignment) 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 and/ or discontinuous units in separate files. Thus it becomes possible to view the annotation in aligned segments and to pose queries (using XSLT and XQuery) combining different layers (cf. Neumann & Hansen-Schirra 2005, Hansen-Schirra et al. 2006). 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 what follows 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 address more linguistically meaningful levels of encoding than merely string-based queries which are unable to retrieve information encoded on higher linguistic levels.
1. to 7. below are examples of hypotheses about cohesion to be tested on the data: For either a given pair of non-aligned text segments globally, or else for a given aligned source – target fragment of two texts in a translation relationship, we expect differences along the following parameters:

1. the proportion of explicit to implicit referents;
2. the proportion of phoric to fully lexical (auto-semantic) phrases;
3. the number of newly introduced discourse referents per discourse segment;
4. the amount of cohesive ellipsis and substitution;
5. the strength of lexical chains as measured by various ratios between content and function words, and as measured by type-token relationships;
6. the strength (internal connectivity) of lexical chains as measured by average number of items per lexical chains;
7. the ratio between explicit and implicit encoding of conjunctive relations.

Observe that in comparing any text fragments which are not in a unit-of-translation-relationship, as in our registerially parallel sub-corpora of originals, we are testing for the global property of (relative) explicitness. However, whenever we are comparing a specific aligned and instantiated source-target (translation) unit, we are testing "explicitation" (or its opposite, implicitation).
An indicator for the first hypothesis mentioned above could be the proportion of explicit (pronominal) referents vs. implicit ones in comparing German relative clauses with their non-finite English correspondences. The relevant evidence is reflected in the annotation and alignment at word level. Relative pronouns receive the part-of-speech tag *prels* and if they occur in both languages, they are linked to each other. Is there, however, a relative pronoun in the German translation which cannot be found in the English original text, the German relative pronoun is not aligned at all - it receives a so-called empty link. Figure 1 shows the XML

representation of the tokenised corpus, the part-of-speech tagging and the alignment on word level. In the token index file each token is assigned an index number. The part-of-speech annotation of tokens refers back to this index file via *xlinks* specifying the index number of the respective token. The word alignment then refers to the index numbers of both source and target token in turn. In cases of empty links, they receive the value *undefined*.

```
English original:      ... a  palmist,  inferring  the  future  out  of  his  own  lined  flesh
German translation:    ... ein  Handleser,  der  seine  Zukunft  aus  den  eigenen  Linien  ableitete
                           ( a    palmist    who his     future  out-of  the    own     lines   inferred  )
```

token index file	**part-of-speech annotation**	**word alignment**
<token id="t64" strg="ein"/>	<token pos="art" xlink:href="#t64"/>	<token> <align xlink:href="#t55"/>
<token id="t65" strg="Handleser"/>	<token pos="nn" xlink:href="#t65"/>	<align xlink:href="#t66"/> </token>
<token id="t66" strg=","/>	<token pos="yc" xlink:href="#t66"/>	<token> <align xlink:href="#t56"/>
<token id="t67" strg="der"/>	<token pos="prels" xlink:href="#t67"/>	<align xlink:href="#undefined"/> </token>
<token id="t68" strg="seine"/>	<token pos="pposat" xlink:href="#68"/>	<token> <align xlink:href="#undefined"/>
<token id="t69" strg="Zukunft"/>	<token pos="nn" xlink:href="#t69"/>	<align xlink:href="#t67"/> </token>

Figure 1: XML corpus annotation and alignment on word level including empty links

For the investigation of explicit pronominal referents in German relative clauses vs. implicitly encoded English referents, all German tokens with the part-of-speech tag *prels* (for relative pronoun) have to be extracted which are not aligned on the word level (since the pronominal reference is encoded in the English participle). The respective XQuery is shown in Figure 2.

```
for $k in $doc//tokens/token
  let $fileName := $doc//translations/translation[@n='1']/@trans.loc
  let $fileNameNew := replace($fileName,"tok","tag" )
    where ($k/align[1][@xlink:href != "#undefined"] and $k/align[2]
     [@xlink:href = "#undefined"] and doc($fileNameNew)//token
     [@xlink:href eq $k/align[1]/@xlink:href][@pos eq "prels"])
```

Figure 2: XQuery for relative pronouns with empty links

The output of this query are sentences like the ones displayed in Figure 1. This example (taken from the fiction sub-corpus) is interpreted as explicitation since participant role (and thus the reactivation of the referent), tense and mood are explicitly realised in the finite relative clause of the German translation, whereas they are implicit in the English original.

```
<result no="13"><ori_en>Baker Hughes Business Support Services has assumed accounting, payroll, benefits and IT
  support duties for many of the company's U.S. operations, eliminating duplicate efforts by division personnel. </ori_en>
<trans_ge>Baker Hughes Business Support Services hat die Buchführung, Gehalts- und Sozialleistungen sowie IT-
  Aufgaben für viele Niederlassungen des Unternehmens in den Vereinigten Staaten übernommen, wodurch doppelte Arbeit
  durch das Personal in den Tochterunternehmen vermieden werden konnte. </trans_ge></result>

<result no="14"><ori_en>In this environment, Baker Hughes revenue declined 22% to $4.5 billion for 1999, compared to
  $5.8 billion in 1998. </ori_en>
<trans_ge>Vor diesem Hintergrund sanken die Umsatzerlöse von Baker Hughes im Jahre 1999 um 22% auf 4,5 Mrd.
  Dollar, während sie 1998 noch 5,8 Mrd. Dollar betragen hatten. </trans_ge></result>
```

Figure 3: Results for conjunctions with empty links

A similar query could be posed for conjunctive relations. Here, all German tokens with the part-of-speech tag *kous* (for conjunction) are to be extracted which are not aligned on word level (since the conjunctive relation is encoded implicitly, for instance through a participle clause). The results for this query displayed in Figure 3 are taken from the sub-corpus of shareholder information. Here, all examples show explicitation in the German translations, since the implicit conjunctive relation encoded in the English participles (marked in bold face) are translated explicitly with German conjunctions.

Another interesting phenomenon is ellipsis in translations, since it is a direct indicator for implicitation. The alignment of our corpus on different layers enables us to find ellipsis through empty alignment links on all alignment levels. On the basis of these empty links, we can furthermore investigate in which syntactic functions ellipsis occurs, whether it prefers finite or non-finite clauses or how it is dealt with in different translation directions.

The long-term aim of the present study is to identify, count and interpret cohesion markers and lexicogrammatical markers (not addressed in this paper) and their quantitative (co-)occurrences and patterns found in translations and their source texts as indicators of explicitation/ implicitation. This will be interpreted against the background of three sources of explanation: language typology, text typology and the translator's language processing (cf. Steiner 2001, Hansen 2003, Neumann 2003, Teich 2003). 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, levelling-out (Baker 1996), culturally determined preferences (House 2002), shining through (cf. Teich 2003), density and directness (cf. Steiner 2005 a,b,c). Beyond properties of translation the resource opens up new research perspectives on basic questions of translation studies like the translation unit.

In an overall perspective, we are working towards constructing and making available our corpus as a resource, which is theory-neutral and should be usable as an empirical basis for research informed by different theories and models. At the same time, research in our own group is corpus-based, but not corpus-driven, in the sense that our research questions and hypotheses do not "emerge" out of the data, but are derived from a range of theories and models about language, texts, and translations. On the basis of the research design presented here we hope to be able to report on some initial empirical findings relating to our hypotheses by the time of the talk.

Returning, finally, to a comparison of our approach with earlier investigations of "explicitation" in translation studies (as in work by Blum-Kulka 1986, Baker 1996, Laviosa 1998, Olohan and Baker 2000, Englund-Dimitrova 2005), we are aiming at significant progress towards closing the methodological gap between high-level notions, such as "explicitness" or "explicitation" on the one hand and levels of linguistic encoding in our data on the other, while not resorting to heavily interpretative example-based examinations of individual cases.

References

Baker, M. 1996. Corpus-based Translation Studies: The challenges that lie ahead. In: Somers, H. (ed.), *Terminology, LSP and Translation Studies in Language Engineering*. Amsterdam/ Philadelphia: John Benjamins. 175-186.

Biber, D. 1993. Representativeness in Corpus Design. *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 Cognition in Translation and Second Language Acquisition Studies.* Tübingen: Narr. 17-35.

Englund-Dimitrova, B. 2005. *Expertise and Explicitation in the Translation Process*. Amsterdam: John Benjamins.

Fabricius-Hansen, C. 1999. Information packaging and translation: Aspects of translational sentence 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.

Hansen, S. 2003. *The Nature of Translated Text. An interdisciplinary methodology for the investigation of the specific properties of translations*. Saarbrücken: Saarbrücken Dissertations in Computational Linguistics and Language Technology. vol. 13.

Hansen-Schirra, S.; Neumann, S. & Vela, M. 2006. Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. In: *Proceedings of the 5th Workshop on Multi-dimensional markup in NLP*. Trento. 35-42.

Hasselgard, H.; Johansson, S.; Behrens, B. & Fabricius-Hansen, C. (eds.) 2002. *Information Structure in a Cross-Linguistic Perspective*. Amsterdam, New York: Rodopi.

House, J. 2002. Maintenance and convergence in translation – some methods for corpus-based investigations. In: Hasselgard et al. (eds.) 2002. 199-212.

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.

Laviosa, S. (ed.). 1998. *Meta Translators Journal.* vol. 43 no.4.

Neumann, S. (2003) *Die Beschreibung von Textsorten und ihre Nutzung beim Übersetzen.* Frankfurt etc.: Peter Lang Verlag.

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. *Computational Linguistics*, vol. 29, no. 1: 19-51.

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

Steiner, E. 2001. Translations English - German: investigating the relative importance of systemic contrasts and of the text-type "translation". *SPRIK-Reports* no.7. Oslo.

Steiner, E. 2005a. Some properties of texts in terms of 'information distribution across languages'. *Languages in Contrast 5:1 (2004-2005)*: 49-72.

Steiner, E. 2005b. Some properties of lexicogrammatical encoding and their implications for situations of language contact and multilinguality. In**:** Franceschini, R. (ed.), *Zeitschrift für Literaturwissenschaft und Linguistik* Jahrgang 35, Heft 139. Stuttgart: Metzler Verlag. 54-75.

Steiner, E. 2005c. Explicitation, its lexiocogrammatical realization, and its determining (independent) variables – towards an empirical and corpus-based methodology. *SPRIK-reports* no. 36. Oslo.

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