

## **Types of contrasts, linguistic levels, their relationship to *explicitation and explicitness*, and possible explanations**

Short version, electronically available through *CroCo*-website  
The longer version containing an additional discussion of the individual studies,  
is available upon request from the project.

### **Croco Work Package 2.3.**

**18 January 2008**

**Erich Steiner**

The following report on Workpackage 2.3. summarizes, contrasts and interprets 12 studies coming out of the first phase of Croco and closely related work (cf. <http://fr46.uni-saarland.de/croco/>). A structured listing of most of these, with initial interpretations, can also be found in Kerstin Kunz September 2007-deliverable on the issue.

The report gives an introduction and overview, covering the corpus, the types of contrast represented in its architecture and a first overview of studies. Each of the 12 studies (**S1-S12 below**) is discussed

- in terms of the types of contrast investigated in the corpus,
- in terms of the linguistic phenomena covered on their respective linguistic levels
- in terms of their results, and in their role as operationalizations of/ relationship to *explicitness* and *explicitation*,
- and finally in terms of what kinds of explanation can be evoked against the architecture of the Croco corpus and in view of independent sources of explanation (language type, register, translation as a process of text production).

Our main aim is to get a better understanding of what types of results are possible with our research strategy, and what changes and improvements are necessary in order to make further progress.<sup>1</sup>

---

<sup>1</sup> Studies number 1,4,5 reported on here are accepted for publication. Study 2 has been published. Study 3 is available as a project report and has not been published yet. Studies 6-12 are Diploma Thesis of our Department, all produced as part of, or at least in close collaboration with, *Croco* work.

# Contents

## 1. Introduction and overview

### 1.1. The corpus

### 1.2. Types of contrast

### 1.3. Overview of studies

#### 1.3.1. Types of contrast

#### 1.3.2. Linguistic levels and operationalizations

#### 1.3.3. Results and their relationship to explicitation/ explicitness

#### 1.3.4. Explanations: corpus architecture and sources of explanations

## 1. Introduction and overview

### 1.1. The corpus

We shall start by graphically representing the structure of the *CroCo*-Corpus (Figure 1), to be followed by a list of types of contrast investigated within this corpus:

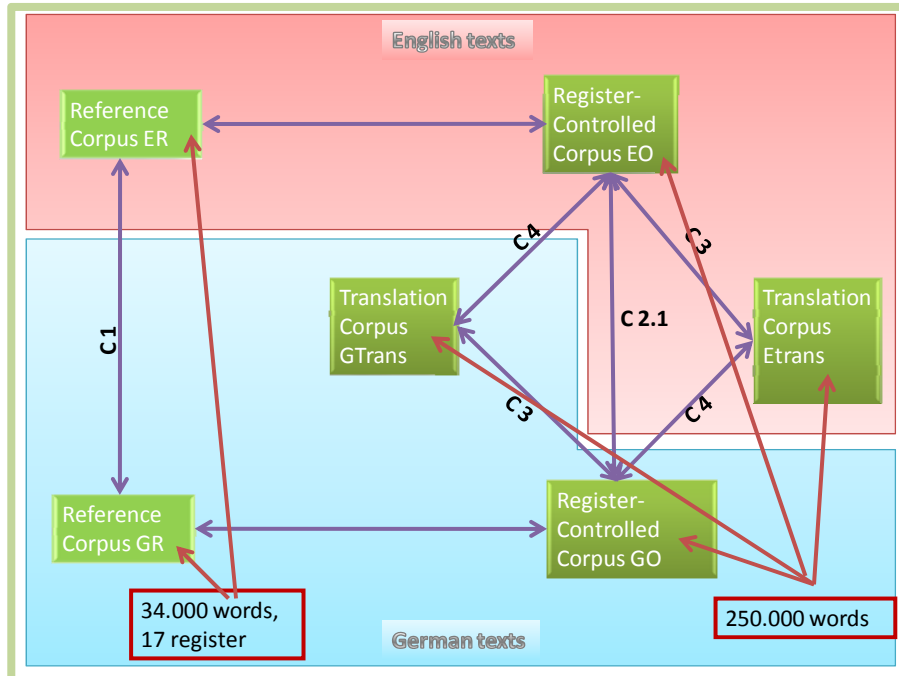


Figure 1: Structure of the CroCo-Corpus with types of contrast

Figure 1: Structure of the CroCo-Corpus with some relevant types of contrast

**1.2. Types of contrast**

Our arrangement of sub-corpora permits the following relevant types of contrast to be investigated (cf. Steiner. forthcoming):

- **Contrast C1** (*reference corpora, cross-register*) between the English Reference Corpus (ER) and the German Reference Corpus (GR), each consisting of 34 000 words (17 registers of 2000 words each). Contrasts under C1 yield a cross-register profile for original texts in the languages English and German (cf. Figure 1).

- **Contrast C2** (*register controlled*) between the registers of *Political Essays on Economics* (ESSAY), *Fictional Texts* (FICTION), *Instruction Manuals* (INSTR), *Popular Scientific Texts* (POPSCI), *Corporate Communication* (SHARE), *Prepared Speeches* (SPEECH), *Tourism Leaflets* (TOU), *Websites* (WEB) for each of English originals (EO), German originals (GO), English translations (ETrans), and German translations (GTrans), each of the sub-corpora having 31,250 words per register, each register sample comprising at least 10 texts, 250 000 words altogether for each of EO, GO, ETrans, GTrans. The translations are all translations of the corresponding samples of matching originals. Within this contrast, we can separately investigate **Contrast C2.1** (*within one register, between languages, differentiated into 8 sub-contrasts by register EO vs. GO*), and **Contrast C2.2** (*between registers, within each of the languages English and German, yielding 8 contrasts within each of the corpora EO and GO*). Figures 2 and 3 visualize these types of contrast.

**Include Figures 2 and 3 at closest possible point**

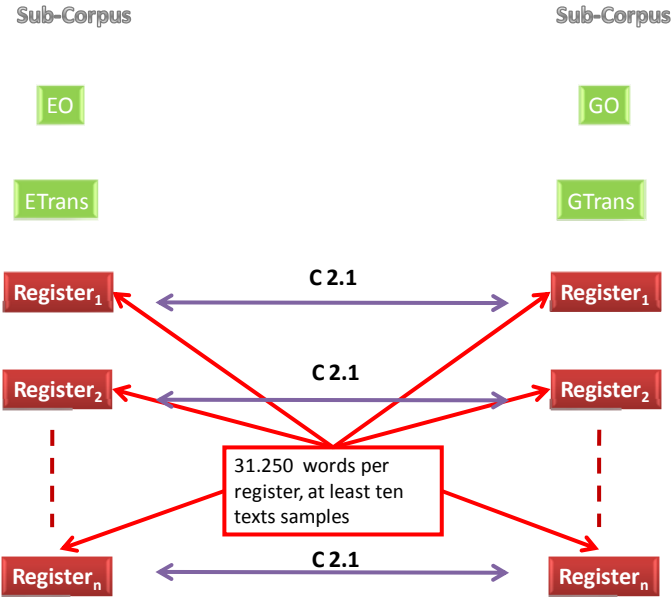


Figure 2: Contrast of Registers across Languages

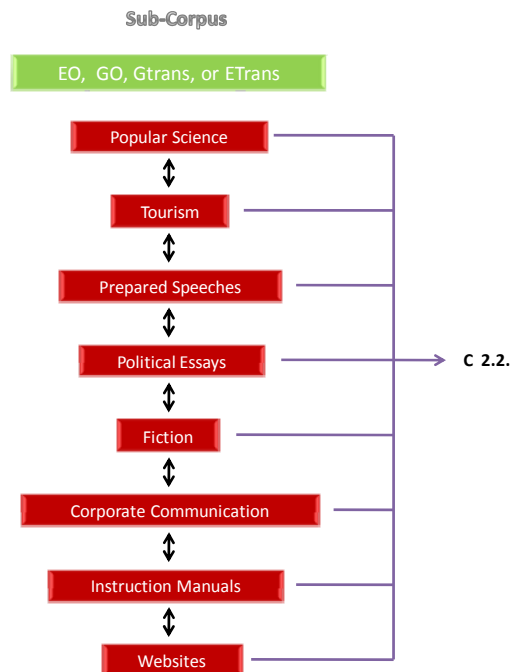


Figure 3: Contrast of Registers within Languages

- **Contrast C3** (*translations vs. originals within each of the two languages*): EO vs. ETrans and GO vs. GTrans, yielding either one contrast per language globally, or, if intersected by register, 8 contrasts between originals and translations for each of the languages (cf. Figure 1).

- **Contrast C4** (*originals and their translations across languages, i.e. EO vs. GTrans and GO vs. ETrans*); this contrast is the only one between originals and their translations. We differentiate this into **C4.1** *Sub-corpora (and texts) as wholes (without alignment)* and **C.4.2** *Aligned corpora, i.e. explicitation by translation units*, and we investigate translations between English and German in both directions (cf. Figures 4 and 5).

**Include Figures 4 and 5 at closest possible point**

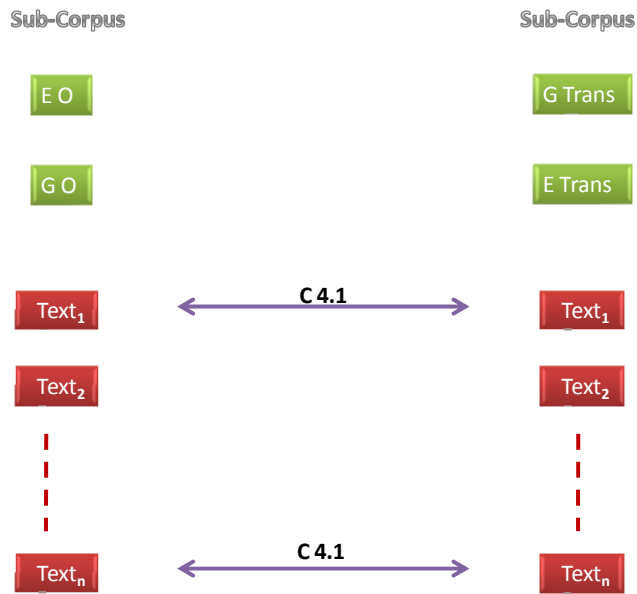


Figure 4: Contrast between Source texts and Target Texts by Text

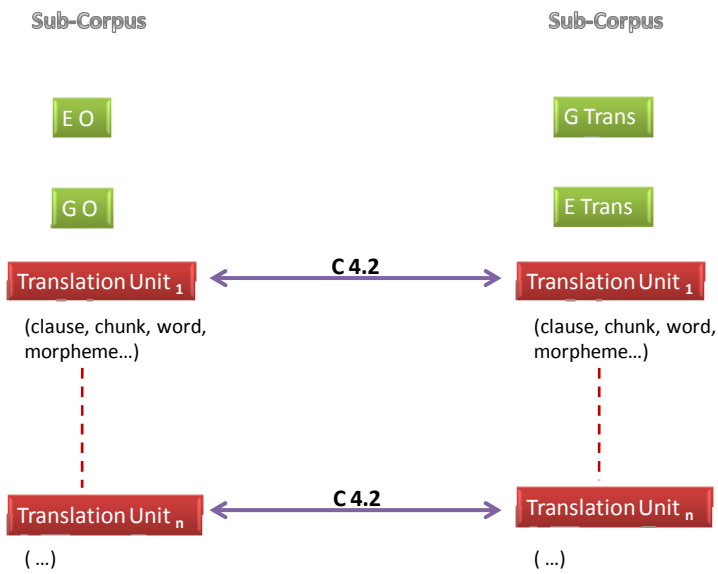


Figure 5: Contrast between aligned Translation Units

### 1.3. Overview of studies

#### 1.3.1. Types of contrast

By way of an initial overview, we can group our studies **S 1-12** below as follows:

In terms of **types of contrast**:

- S1** (cf. Steiner forthcoming) focuses on the non-aligned corpora (C1-C4.1.),
- S2** (Hansen-Schirra et al. 2007) and **S3** (Hansen-Schirra, Neumann and Vela. Ms.) have as their focus largely C4.1. and C4.2., that is original-translation (source-target) corpora, both aligned and non-aligned.
- S4** (Vela et al. 2007) contrasts registers (largely C2.2., but implicitly also C2.1., for *Fiction* and *Share*), but also originals and translations (C4.1.)
- S5** (Neumann forthcoming) investigates C2.1. and C2.2.,
- S6** (Klein 2007) focuses on C4.1. and C4.2.
- S7** (Bierster 2007) investigates C4.2.
- S8** (Kast 2007) investigates C2.1. for SHARE, to some extent C4.1. SHARE, and largely C4.2.
- S9** (Reuter 2007) investigates largely C3, but as correlates also C2.1. and C4.1.
- S10** (Schuster 2007) investigates in detail for one text-pair C4.1. and C4.2.
- S11** (Kavelius 2007) investigates C2.1. (not *CroCo*, but court decisions E-D)
- S12** (Grahn 2007) investigates C2.1., C3 and C4.1.

#### 1.3.2. Linguistic levels and operationalizations

In terms of **the linguistic levels and operationalizations** involved, the studies cover:

- S1** global text properties using word-level statistics lexical density (LD), type-token-relationships (TTR) and part-of-speech (PoS) distributions
- S2** cohesion (mainly reference and lexical cohesion); operationalized through noun-pronoun proportionalities, TTRs and LD
- S3** word level, shift of PoS in translation, operationalized through PoS-Class
- S4** lexical cohesion, operationalized through word-frequency lists, length and connectedness of lexical chains, tense oppositions, grammar, grammatical density, grammatical rank proportionalities,
- S5** lexical cohesion in the form of frequencies of lexical words, length and connectedness of lexical chains, grammar in the form of proportions between sentences, clauses, chunks and words (density), and an aspect of cohesive reference in the form of proportionalities of certain pronouns
- S6** all types of cohesion
- S7** cohesion (reference, ellipsis, substitution), example-based only
- S8** lexicogrammar, mainly syntactic functions, but also phrase type, rank, voice

**S9** cohesion (conjunctive relations); operationalized through hand-coded types of conjunctive relations

**S10** grammar, mainly syntactic function (adverbials), syntactic form (phrase type) and their behaviour (shifts) under translations (interacting with clefting, verb-incorporation, modal auxiliaries).

**S11** lexical cohesion, operationalized through repetitions, sense relations, collocations, and length and connectedness of lexical chains

**S12** text (register) properties addressee-orientation, operationalized through personal deixis, clause mood, modality, voice

### 1.3.3. Results and their relationship to explicitation/ explicitness

In terms of their results and their relationship to explicitation/ explicitness, S1 – S12 can be characterized as follows :

**S1** The results along all dimensions of contrast (C1 to C4.2.) can be taken as (very indirect) reflexes of (experiential, logical) types of explicitness. Lexical density could be related to experiential explicitness (though not logical, or even textual). High TTR might indicate certain types of strength of lexical cohesion (though not by repetition), a strong nominal orientation could be an indicator of certain types of referential density.

**Along the C1 dimension**, there is some difference in LD (0.51; E>G), but a much bigger difference in TTR (for German 21,71 > English 15,64) throughout, plus a stronger nominal orientation for German, although partly due to pronouns. In terms of POS, there seems to be a degree of dominance of nominal word classes in German (N+Adjective+Adposition) vs. verbal classes (V+Adverb+Conjunction) ( 40,21 vs. 22,53) compared to English (Nominal: 41,39, Verbal. 25,47), especially if we take into account the fact that for the pronouns, German once more has 8,45, against English having 5,46 only . There is no straightforward interpretation as to *overall explicitness* here, because the stronger verbal orientation of English may easily be counterbalanced by the stronger morphological marking in German. However, if the stronger nominal orientation of German can later be shown to be due to a higher number of arguments per predicate, this would be one type of explicitness. *Explicitation* does not enter into the picture here, as the two reference corpora are not in a translation relation.

The results along all dimensions of contrast (C1 to C4.2.) can be taken as (very indirect) reflexes of (experiential, logical) types of explicitness. Lexical density could be related to experiential explicitness (though not logical, or even textual). High TTR might indicate certain types of strength of lexical cohesion, a strong nominal orientation could be an indicator of certain types of referential density.

**Along the C2.1 dimension**, deviation from the ER vs. GR differences in **LD**-magnitude (0.51; E>G in reference corpora) and/ or even direction seem to indicate register-specific

factors. We find a difference in direction for *Essay*, *Fiction* and *Tourism*, and for *Popsci*, *Share* and *Speech* differences in magnitude (E>G by more than 0.51). In terms of TTR, some registers seem to differ from the baseline (G>E by 6,07) in magnitude (Essay, Instruction, Tourism, Web), but none in direction; PoS needs considerably more evaluation. Some individual differences are: Register differences for nouns exceed those in the reference corpora (E>G by 1,67) for all registers except Fiction. For pronouns (G>E by 2,99 in reference corpora), all registers exceed the difference except *Instruction* and *Tourism*. For verbs (E>G by 2,68), *Essay*, *Popsci*, *Web* show a larger difference than the reference corpora, but some of the other registers also a substantially smaller one.

As for an overall interpretation, whereas in terms of **lexical density** German registers and English registers vary as to which is the more *explicit*, and whereas in terms of **type token ratios**, it seems to be the German registers throughout which are more explicit, but with register-specific effects in the case of *Essay*, *Instruction*, *Tourism* and *Web*, English registers seem to be higher throughout in terms of **verbal orientation** than the German ones. Again the question remains, what precisely this implies for **explicitness**, but this will be further investigated.

**Along the C2.2. dimension**, the LD spread of variation among English registers (originals) is wider than it is for German originals. The extreme registers for both languages are *Fiction* (low) and *Tourism* (high). We can also see that the spread in lexical density *within* languages is much higher than for any register *across* the two languages. The spread of variation in TTR among German registers is numerically larger than it is for English. This time, though, the extreme registers for the two languages are *Instruction* (low) and *Fiction* (high) for English, but *Instruction* low and *Tourism* high for German. In terms of general nominal orientation, our registers are ranked almost identically in the English and German corpora. While German comes out as more nominal in a general sense, this tendency does not hold if we simply compare the register specific frequencies of nouns (without pronouns), where English usually scores higher than German (cf. S1 below for more details).

**Along the C3 dimension**, LD is globally higher for originals than for translations in both languages, but individual registers go the other way. Also, in both languages the spread of lexical density between registers is smaller in translations than in originals, pointing to interpretations in the sense of “levelling-our” (Baker). TTRs go both ways, but show very different effects between E and G (levelling-out vs. interference). In terms of PoS, translations have fewer verbs and fewer nouns than originals in both languages. They show increased *entity*- and *event* modification (Adjectives and Adverbs), as well as increased logical explicitness and nominal determination. In summary, in both languages, translations are less dense, lexically less rich (TTR), and partly more verbal than originals. They show increased *entity*- and *event* modification, as well as increased logical explicitness and nominal determination

**Along the C4.1. dimension**, lexical density (LD) is higher in English originals than in their German translations by 0.84 points, which is more than what the reference corpora would suggest, and yet more than the difference between EO and GO originals. This suggests a clear effect of the process of translation. Very significantly, in *Fiction* and *PopSci*, the lexical density of the German translations is even higher than that of the English originals, counter to the overall tendency, which indicates strong register-specific influences. In the case of *Fiction*, this could be due to a difference in LD between the originals - not so, however, in the case of *Popsci*. Quite significantly, German originals also have (slightly) higher LD than the English translations globally. TTR is higher in German translations than in EO but generally not by as much as between the two corpora of originals. Translation as a process again seems



to have an effect there. The difference between German originals and English translations is - predictably - even wider generally, though not to the same extent in all cases. In terms of POS, for the combined 8 registers of English originals and their German translations, the nominal word classes seem to score lower for English - though only due to the high percentage of pronouns and adjectives in the German translations. The verbal classes seem to score higher in English than in German:

Comparing these proportionalities with those between originals in the languages, the German translations have moved somewhat towards their English originals in comparison to German originals. Conversely, we find that the English translations have moved somewhat towards their German originals in comparison to English originals. Altogether, then, the difference in “nominal orientation” is bigger between originals than between the translations.

**S2:** C1-C4.1. proportions between nouns and pronouns are discussed as a measure of referential explicitness. The register-neutral reference corpora German (GR) include a lower proportion of nouns and a higher proportion of pronouns than English (ER). The different proportions in EO and GO are probably a reflection of the broader registerial composition of ER and GR. The frequencies in the translation sub-corpora in most cases lie between the originals and the reference corpora, moving towards the latter, possibly showing normalization. And finally, the comparison of originals and their matching translations in the respective target language reveals a strong influence of the target language. The frequency of pronouns in ETrans is even lower than both the ER and the EO percentages so that target language conventions are “exaggerated”.

TTRs are higher in German translations than in English originals, which can be largely traced to systemic factors, but also, to some extent, to translation strategies.

LD in Fiction is much higher for GTrans than for EO, which seems to be a clear indication of the effect of the translation process, given that the baseline figures for ER vs. GR show ER with the higher LD-value. As we do find a difference in Fiction between EO and GO ( $E < G$ , but not by as much as indicated in the figures reported in Table 5 of Hansen et al. 2007), the translators seem to be sensitive to register-specific profiles here as well.

C4.2. These findings (Figure 3) seem to provide evidence about explicitation/ implicitation of (pronominal) referents fairly directly. They would also be findings about explicitation on the interpersonal level (Mood/ Finite) and on Tense.

**S3:** There are clear tendencies of moving from nominal to verbal classes in translation, but varying with translation direction. A move towards verbal categories can be interpreted as explicitation into the event/verbal direction (de-metaphorization).

**S4:** There are weaker lexical chains in FICTION than in SHARE in several senses; the lexical chains in the translations are shorter in both registers (EO and GTrans in both cases). As for possible interpretations of these findings, we can assume that the referential meaning in the FICTIONAL text is more “diffused” compared to the SHARE text.

Frequencies of words do not seem to have a direct relationship to explicitness or implicitation. They may be indicators of strength of lexical cohesion, though. Length and connectedness of lexical chains are relatively direct indicators of strength of lexical cohesion, which, once more, should not be confused with referential explicitness.

Relative frequencies of past tense and non-past tenses are good indicators of registers (together with other features). Narrative texts are clearly singled out thereby. The relationship of tense selections to explicitness seems to be indirect at best.

Proportions between grammatical ranks (Table 2 in Vela et al.) seem to be a direct indication of differences in experiential density between same-register originals across the two languages, and between originals and translations. With figures available for the SHARE register only, because of the necessary hand-coding for higher-rank grammatical units, the English originals have more clauses per sentence (E 2,45 against German 1,69), but as a compensation, German has more chunks per clause (G 3,19 against English 1,99). Where the translations into English display a higher value, e.g. in the number of clauses (866 more than the German originals) the translations into German display a lower value (552 fewer clauses than the English originals). In other words, the English texts appear to be denser in terms of clauses taxis, whereas the German ones appear denser in terms of phrases per clause. The translations appear to exhibit tendencies of “normalization” in both directions, yielding something like a neutralized “middle-ground” in terms of that type of density.

There is furthermore probably some degree of association between higher lexical density and referential explicitness. As Vela et al are saying, one would expect the register of prepared speeches to exhibit a low lexical density as an indication of their spoken medium. While this register has the second lowest mean value (52.58 percent), it still has clearly higher values than FICTION. English original fictional texts have the lowest maximum value and the lowest range. They also have the lowest mean value (45.72 percent) for lexical density. The reason for these findings is probably that, on the one hand, the speeches carry more characteristics of the written mode due to their elaborated preparation process. On the other hand, the fictional texts frequently contain dialogical elements between the characters in the fictional world of the text.

**S5** The two registers (*Fiction* vs. *Share*) are clearly differentiated within their languages along all dependent variables; they are less different across the two languages; higher or lower frequencies of lexical words globally would seem to be an indicator of experiential explicitness (low frequencies of repetitions paired with high lexical density); length and connectedness of lexical chains would be one indicator of explicit and context-independent cohesion; proportions between grammatical ranks would be direct indicators of grammatical density (not in a simple relationship to explicitness); proportionalities between pronouns would be one indicator of thematic explicitness and/ or participant involvement

**S6:** In the area of *cohesive reference*, the German texts investigated here have a higher proportion of pronouns, but fewer of them cohesive (as opposed to “grammatical”) than in English. On the other hand, English “it” often is translated as either ellipsis, or fully lexical, or else a demonstrative. English possessive determiners often become definite articles in German. English demonstratives also frequently become definite articles in German. English *it/that* in clefts correspond to adverbials, but also to pronominal adverbs and similar cohesive devices. In many cases, pronominal adverbs in German translation are just an addition relative to their English originals.

Under *substitution*, only verbal substitution occurred with some frequency. In German, it shifts to either the general verb “tun/ machen” plus complementation (i.e. probably not a form of cohesive substitution) , or else to lexical cohesion, or ellipsis.

*Ellipsis* shows frequent, and different types, of changes between the two languages. In the English-to-German direction, we often get some form of explicitation, but the opposite is also attested, i.e. cases where GTrans has elliptical realizations with full English variants in EO. Translations of English tag-questions into German elliptical structures also usually yield implicitation (“nicht, oder?”).

*Conjunction* shows implicitation in the English-German direction to some extent. Of particular interest is “downranking” of cohesive conjunction as grammatical conjunction or preposition. But even more often, we find explicitation of various types in the English-German direction.

*Lexical cohesion* shows disparate phenomena, depending on register, translation direction, and type of lexical cohesion.

For reference, substitution, ellipsis, conjunction the relationship to explicitation (and, where applicable, explicitness (C2.1. and C2.2.)) is relatively straightforward, although the exact operationalization for an “implicit” relation (rather than one not given) needs to be critically examined. One type of “explicitation” is found, where and when an original encodes something grammatically, and the translation encodes it cohesively.

With lexical cohesion, we need to investigate more, whether lexical (and conjunctive) cohesion are by themselves more explicit, than the other three types, and whether within lexical cohesion, there are different degrees of explicitness dependent on the sub-type of lexical cohesion.

In summary, explicitation and simplification seem co-vary, as do implicitation and (opposite of)-simplification;

**S7:** In this example-based study, we get clear instances of explicitation whenever reference through pro-form is substituted by reference through fully lexical material. However, we also get narrowing and widening of referential focus with any change between demonstratives (articles or pronouns) and simple pro-nouns (E-D), usually happening in translations into German from English. Is increased/ narrowed deictic focus a form of explicitation?

When pronominal adverbs are translated by simple pronouns (or general nouns), usually in translations G into E, we often get implicitation/ explicitation of the types just mentioned, sometimes in addition to a change of syntactic function and therefore grammatical metaphor.

When ellipsis and substitution (usually E into G) are translated by a different type of cohesive device, the effects on explicitation are relatively obvious.

**S8:** In a substantial subclass of cases, experiential identity in Subjects is preserved under translation. Referentially empty Subjects get lost in translations, which means loss of explicitness of marking of information status (clefts in English or “impersonal passives” and “process-thematizing constructions” in German). With Imperatives, the German explicit social distance distinction gets lost in translations into English. Passives often lead to agent-implicitation. The English Subject in translations into English often receives “spatio-temporal” semantic roles, which is not the case in GO (implicitation?). With translations into German, the opposite tendency is found.

**S9:** Originals in E and G were not differentiated by explicitness/ implicitness of conjunctive relations globally in the corpus. However, there was more explicitness in German originals and in translations into English realized as punctuation. The implicitness found in English

originals often is conditioned by non-finiteness, and does not occur in German originals and also less in translations into English (ETrans). Implicitness is most often found with causal-conditional relationships. For those constellations where clear findings of difference were obtained, there is basically a fairly direct relationship to explicitness/ implicitness.

**S10:** Frequent cases of rank-shift in translation (S, CL, phrase/ group/ word), with relatively clear consequences for explicitation. These may be slightly less clear in mere cases of re-mapping between syntactic functions (e.g. Adverbials onto Subject/ Objects etc). Very frequently, English non-Adverbials became Adverbials in German. Reasons were often clefts, verb-incorporated meanings, then also other syntactic functions (Subj, Obj). In terms of word order, and as one would predict from the respective basic word orders of English and German, adverbials in German were often realized in early mid-field, as opposed to English. The effect of the mere ordering is textual, not ideational. This might point to a basic translational difficulty between the two languages: if the basic word order position of adverbials in German is in early mid-field and thus out of the focused information area (NEW), and if in English most adverbials' basic position is on either the left or right periphery of the clause, i.e. in thematically or informationally focused areas (as in Halliday and Matthiessen 2004 and elsewhere), then category change in translation seems inevitable if preservation of information structure is a high ranking goal of translation. If, however, we assumed (following Doherty 2002, 2006) English information focus to be in the centre of the clause (around the verbal complex), then the difficulty would not arise.

The overall variability of word order positions of adverbials is, predictably, higher in German. Adverbials in the German translations seem to be more frequent overall than in the English original, even though English has 56 Adv clauses vs. German 47. This could indicate a global tendency in the German towards adverbial / circumstantial explicitness, though not in the case of clauses.

**S11:** The English corpus has more *repetition* than the German. This could be interpreted as a higher explicitness, though less variability, of lexical cohesion. In terms of semantic relations (sense relations), English seems to prefer *hyponymy*, whereas German has more *meronymy*, which would seem to be neutral as to effects on explicitness. Interestingly, the English has overall substantially more lexical cohesion than the German (my interpretation, missing in thesis). This would be a striking difference and would point towards more explicitness of lexical cohesion in English.

In terms of breaks of lexical chains, English has modestly more than German, which might be interpreted as less experiential topic consistency, or at least as a less explicit marking of it.

German also has a higher frequency of composite, which may mean less explicitness compared to a more analytic composition of conceptual experiential meaning.

**S12:** This study finds more explicit addressee encoding in English texts; explicit direct imperatives in EO; more "modulation" in GO; more passives and more agent-less passives in GO. In summary, the English originals are more explicitly addressee-oriented.

### 1.3.4. Explanations: corpus architecture and sources of explanations

Finally, **on the basis of the architecture of the CroCo corpus and in view of independent sources of explanation** (language type, register, translation as a process of text production), we would assess our studies as follows:

**S1:** All the results in this study are results about explicitness, rather than explicitation (i.e. no results in the areas of C4.2.), and all these results are based on evidence on linguistically low levels or epiphenomena (LD, TTR, PoS). They have thus a wide coverage of types of contrast, but miss out on the direct contrast C4.2.. They also have a broad empirical base within the corpus, but being low-level in terms of abstraction from the data, their relationship to explicitness is very indirect.

For results in the area of C1 (slightly higher lexical density of ER, much higher TTR for GR<sup>2</sup>, some dominance of nominal word classes for GR), the explanations can only be sought in terms of the language (type) of English and German.

For results in the area of C2.1., differences which are larger, or smaller, than those observed for the Reference Corpora, or even different in direction, must be due to the register involved, because we subtract the known numerical difference observed in the reference corpora, and we know that all the texts in the sub-corpora are originals, so there cannot be influences of the translation process. The register is being kept constant in C2.1. Observe that in all cases of C2.1., we are interpreting the *magnitude* and *direction* of *differences* between ER-GR on the one hand, and registers within EO and GO on the other. It is thus not the difference C2.1. as such which matters, but the *derived difference between C2.1. and C1 (each of which are already differences)*, its magnitude and direction. The differences themselves have been summarized above ((cf. 1.3.3. under S1)).

For results in the area of C2.2., the differences which we observe must be due to the register, but this time language-internally. Also, the ranking of the registers along scales of LD, TTR and POS-proportionalities is interesting as a measure of language-internal consistency of our register selection. Differences in the *spread of variation* between registers are an interesting fact in themselves, open to various interpretations. And finally, each of the 8 registers within EO and GO shows its own difference (in terms of LD, TTR, POS) from the figures for the reference corpora, which can be taken as one indicator of the extent to which the register is, in fact, identifiable within its own language.

For results in the area of C3 (translations are less dense (LD), lexically less rich (TTR), and partly more verbal than originals; they show increased *entity*- and *event* modification, as well as increased logical explicitness and nominal determination), they can, initially, be attributed to interference, possibly additionally to register (if we are looking at individual registers), or else the translation process. If they are numerically different from what the respective differences in the reference corpora establish as a baseline, but still pointing into the direction of the source language, this numerical difference would be due to interference. If they are different in direction, then this would seem to be an influence from the translation process, yielding something like “normalization”. The “null-hypothesis” would be that there are no differences. In our results we find, depending on register, a mixture of interference and its opposite, and this fact would indicate that the register involved plays an additional role in the

---

<sup>2</sup> The substantially higher TTR values for all German corpora here are partly due to our neglecting of the results of composita-analysis in our morphological parser. We are currently considering improved figures here.

translation strategy. Furthermore, the smaller overall difference (spread) in LD for translations vs. originals in both languages is reminiscent of the assumption (Baker 1996) that translations are more similar to each other (levelling-out) than original texts. Results in C3 seem to be the most difficult to explain by recourse to one dominant factor.

For results in the area of C4.1., the Null-Hypothesis would be that the differences are exactly the differences gained from the Reference Corpora. Wherever they are different in size and/ or even direction from the observed differences in the reference corpora, they would indicate an influence of the translation process, either as interference (shining-through), or else as normalization. Where we observe differences in any of these respects for individual registers, as we do, these must be taken to be caused by the register-specific translation strategies involved. For all results under C4.1, we cannot straightforwardly attribute them to “explicitation”, because under C4.1. we are not investigating aligned corpora (for which cf. below). So, in principle even if not most likely, all difference could be due to arbitrary introductions or cancellation of meanings.

We can also globally state that, within our corpus architecture, if differences in C4.1. were just and only the differences found in the reference corpora, then there should not be any differences for the respective sub-corpora in C3 - which, as we know, is not the case.

**S2:** C1-C4.1. For proportions *nouns: pronouns* as a measure of referential explicitness, **differences between EO-GO proportions and ER-GR proportions** are likely to be a reflection of the broader registerial composition of ER and GR. The **frequencies in the translation sub-corpora** in most cases lie between the originals and the reference corpora, moving towards the latter, possibly showing normalization of the translations into the direction of the norm. Where the frequencies of translations move even further away from the reference corpora than the originals and also *not* in the direction of the source language, this must be a clear case of some property of the translation process, for example “explicitation”, as in the case of a decreased percentage of pronouns in English translations, or else excessive normalization (cf. Table 2 in Hansen-Schirra et al. 2007). And finally, the comparison of **originals and their matching translations** in the respective target language reveals a strong influence of the target language. **TTRs are higher in German translations than in English originals**, which can be largely traced to systemic factors, but also, to some extent, to translation strategies. **LD in Fiction is much higher for GTrans than for EO, which seems to be a clear indication of the effect of the translation process**, given that the baseline figures for ER vs. GR show English as higher by 0.51, and although we do find a specific difference in Fiction between EO and GO, it is much smaller than the differences observed here (Table 5 in Hansen et al. 2007).

The findings under C4.2. (Figure 3) seem to provide evidence of explicitation/ implicitation of (pronominal) referents fairly directly. They would also be findings about explicitation on the interpersonal level (Mood/ Finite) and on Tense, to the extent that increased pronoun frequencies are due to the introduction of pronominal Subjects in the German equivalents of English non-finite clauses.

**S3:** There are tendencies of moving from nominal to verbal classes in translation, but influenced by translation direction. A move towards verbal categories can be interpreted as explicitation into the event/verbal direction (de-metaphorization), unless they have as sources “empty links”, which would then indicate addition of information. If the resulting proportion between verbal and nominal classes exceeds the PoS differences between reference corpora (or corpora of originals), the main explanation would be the translation process as such.

**S4:** There are weaker lexical chains in FICTION than in SHARE in several senses. As for possible interpretations of these findings, we can assume that the referential meaning in the FICTIONAL text is more “diffused” compared to the SHARE text. The lexical chains in the translations are shorter in both registers (Vela et al 2007, Table 1), which would point to an explanation in terms of the translation process, although we will still need comparative figures for that variable for reference corpora and corpora of originals *in toto*.

Relative frequencies of past tense and non-past tenses are good indicators of registers (together with other features). Narrative texts are clearly singled out thereby, making that variable into a good indicator for (narrative) register.

Proportions between grammatical ranks (Table 2 in Vela et al.) seem to be a direct indication of differences in experiential density between same-register originals across the two languages, and between originals and translations. With figures available for the SHARE register only, because of the necessary hand-coding for higher-rank grammatical units, the English originals have more clauses per sentence (E 2,45 against German 1,69), but as a compensation, German has more chunks per clause (G 3,19 against English 1,99). Where the translations into English display a higher value in the number of clauses (866 more than the German originals in terms of C2.1.) the translations into German display a lower value (552 fewer clauses than the English originals). In other words, the English original texts appear to be denser in terms of clauses taxis, whereas the German originals appear denser in terms of phrases per clause. The translations appear to exhibit tendencies of “normalization” in both directions, yielding something like a neutralized “middle-ground” in terms of that type of density. This latter phenomenon seems to be a clear candidate for a phenomenon triggered by an (intuitive) awareness of systemic difference by the translators. We cannot invoke the psychology of language processing in translations as such here, because in that case, translations in both directions should show a shift in one and the same direction - which is not the case.

Degree and range of lexical density for different registers within a language (in this case SPEECH vs. FICTION in English) clearly would seem to reflect register. The reason for the (at first sight) surprising findings for this variable is probably that, on the one hand, the speeches carry significant characteristics of the written mode due to their elaborated preparation process. On the other hand, the fictional texts frequently contain dialogical elements between the characters in the fictional world of the text, thus moving towards spoken medium.

**S5** We can attribute all the findings exclusively to register (FICTION vs. SHARE) along the C2.2 dimension, because there are not translations involved and only one language. The findings along the C2.1. dimension could be attributed to language-type to the extent that they mirror findings of the reference corpora, and what remains beyond the differences due to the reference corpora would then be attributable to register differences across the languages.

**S6:** Findings about differences in cohesion reported here seem to reflect register specific differences for Fiction or Share globally (C4.1.), or even quite directly translational relationships under C4.2.. For those higher-level dependent variables, for which we do not yet have the baseline figures for the other corpora, we cannot filter out the independent variables in a strict sense.

Several of the findings seem to have fairly direct explanations in contrastive differences in either lexicogrammar or else cohesion:

The high German percentages of pronouns, as well as the explicitation of conjunctions, could be related to increased hypotaxis and increased finiteness in German. Changes in dependency structure are known from the contrastive grammar of English and German, as is the

translation of English clefts into German non-clefting constructions. Some of the differences in repetitions as a form of lexical cohesion have to do with differences in composita formation. Some differences in down-ranking or up-ranking of units in translation relationships have to do with known differences of the two languages in the use of grammatical metaphors, i.e. different degrees of directness of mapping between semantics and syntax.

The fairly substantive differences within cohesive reference (use of German articles instead of English pronouns, especially in a demonstrative function; limited availability of substitution in German; differences in cohesive ellipsis; tag questions; extensive use of German “pronominal adverbs”; possibly more lexical cohesion in English, where German relies on reference or ellipsis) are likely explanations for a further group of findings.

In all cases, in which the shifts observed under C4.2. are not either systemically triggered, or else triggered by target-register properties beyond lexicogrammar, we tentatively attribute observed explicitations, simplifications, implicitations and “complications” as due to some property of the translations process as such (understanding etc.).

**S7:** As far as reference is concerned, one could speculate that the general tendency of English to permit less explicit mapping between semantics and syntax helps to allow the tendencies mentioned above. In general, for a compelling explanation of those observations which are not triggered by systemic differences, we need baseline figures for these variables from other sub-corpora.

**S8:** The shifts found can be interpreted against the background of the comparative statistics within SHARE, and the reference corpora, thus isolating the independent variables of language, translation, and register. The subclass of cases in which experiential identity in syntactic *Subjects* is preserved under translation represents a null hypothesis which would state that Subjects get translated into Subjects. Where empty Subjects get lost in translations, we frequently have a comparative systemic explanation (clefts in English or “impersonal passives” and “process-thematizing constructions” in German), though by no means in all cases, as in the optional non-clefting translations of English clefts. Systemic explanations would also seem to apply to neutralizations of the German explicit *social distance* distinction (*Du/Sie*), although in these cases, there would be explicating alternatives. Similarly in the case of English actives translated as passives with agent-implication. Finally, the fact that the English Subject in translations from German often receives “spatio-temporal” semantic roles, which is not the case in the GO where the sources of these English Subjects are spatio-temporal adverbials, can be partly traced to comparative/ systemic differences. In all of the cases where the contrastive systemic difference alone does not explain the shift, we need to look at more statistics in the corresponding corpora of originals and in the reference corpora.

**S9:** The first finding, which is that globally English and German originals from “Essay” were not differentiated by conjunctive cohesion, has to be taken at face value and as a basic profile for comparison. Against that background, it is interesting that there was more explicitness in German originals and their translations into English realized as punctuation, which seems to be due to a systemically different use of punctuation as a realizer of cohesion. The fact that implicitness is most frequent with causal-conditional relationships seems to be due to processing factors.



**S10:** For many of the results, the explanations would seem to lie in language-specific information distribution. Rankshift and re-mapping of grammar-semantics configurations within one rank are often not obligatory. The same would seem to apply to at least many of the word order changes in the position of adverbials in translations English-German. This might point to a basic translational difficulty between the two languages (already said above): if the basic word order position of adverbials in German is in early mid-field and thus out of the focused information area (NEW), and if in English most adverbials' basic position is on either the left or right periphery of the clause, i.e. in thematically or informationally focused areas (as in Halliday and Matthiessen 2004 and elsewhere), then category change in translation seems inevitable if preservation of information structure is a high ranking goal of translation. If, however, we assumed (following Doherty 2002, 2006) English information focus to be in the centre of the clause (around the verbal complex), then the difficulty would not arise.

The higher word order variability in the German texts as such is predictable from systemic differences. The fact that German seems to express more of its experiential meaning in marginal adverbials than in central complements also has partly systemic explanations. In order to determine the relative share of the independent variables more precisely, we need further comparative statistics from our corpora.

**S11:** In this case, some of the explanations would clearly seem to derive from higher-level genre-differences between the two corpora in their respective legal cultures. The higher overall lexical cohesion of the English could be partly due to typology (finiteness) and should (and can) be checked against our reference corpora. The study itself also explores the possibility of a historical development, finding smaller differences in an older corpus.

**S12:** The explanations for the findings cannot be language type, because of the lexicogrammatical acceptability of ETrans (which is much closer to GO than to EO). It does seem to lie more in a) register-specific differences between museum guides E-G, combined with b) a source-text oriented translation strategy used in E-Trans.

## References

- Baker, M. 1996. "Corpus-based Translation Studies: The Challenges that Lie Ahead" In *Terminology, LSP and Translation Studies in Language Engineering*, H. Somers, 175-186. Amsterdam: John Benjamins.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan and Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Bierster, Lisa 2007 *Kohäsionsmittel im Englischen und Deutschen - ein Vergleich anhand ausgewählter Phänomene* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen (S7)
- Doherty, Monika. 2002. *Language Processing in Discourse. A Key to Felicitous Translation*. London, New York: Routledge.
- Doherty Monika. 2006. *Structural Propensities*. Amsterdam: John Benjamins.
- Grahn, Astrid 2007 *Adressatenorientierung in Museumsbroschüren - ien kontrastier Vergleich Deutsch - Englisch* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen (S12)
- Halliday, Michael A. K. and Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar*. London: Arnold (earlier versions by Halliday in 1985/ 1994).
- Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner 2007 "Cohesion and Explication in an English-German Translation Corpus", in: *Languages in Contrast* 7:2. 2007. pp. 241-265. (S2)
- Hansen-Schirra, Silvia, Stella Neumann and Mihaela Vela. manuscript. 2007. *Extracting translation shifts from parallel corpora* (S3)
- House, Juliane. 1997/ 1977. *A Model for Translation Quality Assessment. A Model Revisited*. Tübingen: Gunter Narr.
- Kast, Marlene 2007 *Variation innerhalb der grammatischen Funktion "Subjekt" bei Übersetzungen Englisch-Deutsch und Deutsch-Englisch* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen (S8)
- Kavelius, Tanja 2007 *Lexikalische Kohäsion in Gerichtsentscheidungen. Ein kontrastiver Vergleich im Sprachenpaar Deutsch/ Englisch* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen (S11)
- Klein, Yvonne 2007 *Übersetzungsspezifische Eigenschaften - eine korpusbasierte Studie am Beispiel der Kohäsion* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen (S6)
- Neumann, Stella forthcoming *Quantitative register analysis across languages*. Ms. Forthcoming in...(S5)

Reuter, Anika 2007 *Explizierung von konjunktiven Beziehungen in englischen Originaltexten und Übersetzungen am Beispiel Essay* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen **(S9)**

Schuster, Barbara 2007 *Wortstellung im Englischen und im Deutschen und ihre Relevanz für die Übersetzung - mit Schwerpunkt Adverbialbestimmungen* Diplomarbeit Fachrichtung 4.6. Angewandte Sprachwissenschaft, Dolmetschen und Übersetzen **(S10)**

Steiner, Erich. forthcoming. "Empirical studies of translations as a mode of language contact - "explicitness" of lexicogrammatical encoding as a relevant dimension." in: Siemund, Peter and Kintana, Noemi. eds. forthcoming. *Language contact and contact languages*. Amsterdam: John Benjamins (Hamburg Studies in Multilingualism Vol. 7). **(S1)**

Vela, Mihaela, Stella Neumann & Silvia Hansen-Schirra, 2007, Querying multi-layer Annotation and Alignment in Translation Corpora. In Pre-Proceedings der 4th Corpus Linguistics Conference, Birmingham, UK, 27.-30. Juli 2007. **(S4)**