Exploitation of an SFL-annotated multilingual register corpus

Stella Neumann
Institute for Applied Linguistics, Translation
and Interpreting
Saarland University
st.neumann@mx.uni-saarland.de

Abstract

This paper presents a corpus-based SFL-analysis of the English and German register "travel guide". It discusses the research design necessary to identify lexico-grammatical register features on an empirical basis. By including translations in the corpus it becomes possible to make statements not only on the characteristic features but also on intuitive translational realizations that diverge from these features. Finally, perspectives for computational applications of the presented findings are given.

1 Introduction

Describing a register requires a representative study which entails an empirical corpus-based research design. The present corpus-study aims at describing the register "travel guide" cross-linguistically in English and German. As the description is intended to yield register information for translators, translated travel guides are also included in the corpus.

Over the last decade the specific properties of translated texts have evolved as an important issue in translation studies. A number of studies dealt with the description of typical differences between translations and original texts in the same language. Baker (1993) hypothesized that these systematic differences were universal. Consequently they could be identified in the translations without taking into account the source language originals – under the precondition that the translations are taken from diverse languages. Baker suggested using corpus linguistic methodology to verify this hypothesis.

While earlier studies (e.g. Laviosa-Braithwaite, 1996) focussed on analysing raw corpora it has become more and more obvious that more abstract linguistic information has to be investigated in order to explain the specific properties of translations. Particularly the work of Teich (2001) and Hansen (2002) is seminal for this kind of research. The use of corpus linguistic methods is common to all of these studies for investigating the properties of translations.

Steiner (2002) assumes three sources for these properties: They could either be due to typological constraints of the languages involved, or brought about by the process of understanding during the translation process or finally by differing register features. The present study concentrates on the latter source.

2 Research design

Making statements on the typical features of travel guides in two languages as well as on diverging realizations in translations calls for a rather complex corpus design. The design is further complicated by the aim of giving an overview of the most important lexico-grammatical features instead of picking out just one feature. This aim restricts the choice of a computer tool for linguistic interpretation of the corpus. In the following, these characteristics of the study are discussed in turn.

2.1 Corpus design

The corpus analyzed for the present study comprises five sub-corpora necessary for identifying register features and their realization in transla-

tions. At the same time these corpora should permit distinguishing the specific register features from language typological influences.

Identifying register features: In a first step the monolingual identification of register features has to be dealt with. It is deemed impossible to determine features characteristic of a given register by merely computing frequencies in a register-controlled corpus. Unless it is contrasted to a basis of comparison, a seemingly significant feature could be as frequent as or even less frequent than in other registers, nor is it sufficient to compare the register under investigation to a neighboring register. Similarities could be characteristic of just these two registers. The specific properties of a given register only become visible if compared to a mixture of other registers serving as a tertium comparationis. The English reference corpus (E Ref) was compiled by taking texts from a sub-sample of the FLOB-corpus, a broad-ranged corpus aiming at a "general representation of text types" of British English (Johansson et al., 1978)[1], which was slightly adapted by replacing some of the fiction registers with the registers "cooking recipe", "call for tender" and "prepared speech".

The same applies to the German reference corpus (G Ref) with the sub-sample taken from the PAROLE-corpus, provided by the TELRI Research Archive of Computational Tools and Resources (TRACTOR),[2] adapted to match the modified FLOB-based English reference-corpus. The register-controlled corpora (E Ori, G Ori) were sampled from original travel guides selected by the purposive sampling method. Items were deliberately picked with a view to covering a comprehensive variety of specimens of the register "travel guide" from a range of publishing companies, authors and travel destinations.[3]

Distinguishing language-typological influences: Apart from serving to identify monolingual register features, the reference corpora in both languages also have another function. They help distinguish register-specific features from language-typological influences in the cross-linguistic comparison.

Translations: Finally, in order to gain insight into how the translators treat the register features intuitively, a sub-corpus of translated travel guides (G Trans) is taken into consideration. Again, as for the sub-corpora of original travel guides, the purposive sampling method was used to put together the sub-corpus of translated travel guides. For reasons of feasibility only one translation direction (English - German) is included. Furthermore, the difficulty involved in finding translations to match the random-sampled originals lead to the decision not to use parallel texts. Parallel texts, i.e. the matching original texts, could help verify the interpretation for findings in the translations. Under the constraints of the present study this is not possible.

Corpus size: The corpus size is designed to meet Biber's (1990, 1993) recommendations. According to his calculations ten 1,000-word samples are sufficiently representative of the lexico-grammatical features of a given corpus. To be on the safe side – especially with respect to the reference corpus representing a mixture of registers –, 15 texts samples were included in each sub-corpus. All in all, the set of five sub-corpora as depicted in Figure 1 thus contains approx. 75,000 words.

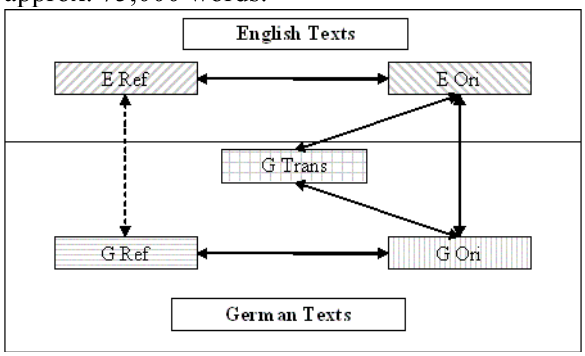

Figure 1. Corpus design

2.2 Features under investigation

While it does not take into account the most delicate specifications of every single feature, the research is intended to cover a representative range of lexico-grammatical features of the register under investigation. As the functional approach of Systemic Functional Linguistics (SFL) is deemed particularly well-suited for cross-linguistic comparison, the corpus was subjected to

[1] http://www.hd.uib.no/icame.html
[2] http://www.tractor.de
[3] For an extensive discussion of sampling methods for register studies cf. Neumann (2002)

an analysis in accordance with Systemic Functional Grammar.[4]

A brief sketch of SFL: SFL has its roots in the anthropologist writings of Bronislaw Malinowski and is oriented towards sociology. Its main representative, M.A.K. Halliday, stipulates that language cannot be investigated in an isolated way but has to be seen in its situational and cultural context. Language is thus described as a social resource enabling speaker and addressee to interact meaningfully.

Language is seen as a system of interlocking options for realizing meaning. These options are then related to each other by structures. SFL thus emphasizes the paradigmatic relationships which are represented in systemic networks (Figure 2 shows the transitivity part of the network used for the present study). Halliday (1994:xiv) calls this systemic aspect of SFL a "theory of meaning as choice".

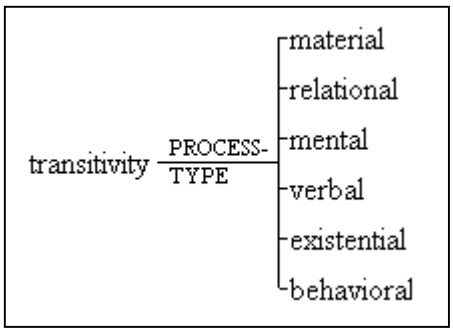

Figure 2. Systemic network for transitivity

The functional character of SFL manifests in its concentration on language in use. *Functional* also implies that the linguistic elements that realize meaning are described with respect to their function in the overall linguistic system. Finally – and most centrally – according to Halliday every language is organized around three highly generalized functional perspectives on meaning which are at work at the same time. The *ideational* metafunction refers to the character of meaning as organization of experience as well as to the logical relationships realized by language. Language's character of helping speakers and addressees to interact is covered by the *interpersonal* metafunction. The *textual* metafunction concerns the way meaning is organized to form a text.

SFL describes language as a complex semiotic system consisting of different increasingly concrete strata with lexico-grammar as the central stratum. In this context, the three metafunctions have different realizations on each stratum.

Features for the present study: The broad coverage of lexico-grammatical features in combination with the corpus design lead to a very complex interpretation task and made it necessary to limit the number of hypotheses to be investigated. Therefore more data that could for instance be yielded from part-of-speech-tagging was not included – although this should be seen as future work to make the description more comprehensive.

Prior to the analysis ten hypotheses about characteristics of travel guides were deduced from the register variables that refer to the situational context. These variables *field*, *tenor* and *mode* realize in turn the ideational, interpersonal and textual metafunction. They specify features from the grammatical systems *transitivity*, *circumstantiation*, *mood*, *modality*, *voice* and *theme*. The annotation scheme for English draws on Halliday (1994) and Matthiessen (1995). Its German counterpart is based on the description of Teich (2001) and Steiner and Teich (in print).

This design allowed the correlation of lexico-grammatical features with more abstract statements on characteristics of the register under investigation like orientation towards content versus addressee etc. In the transitivity system the frequency of the different process types - different types of goings-on that manifest by means of language – was counted. As to circumstantiation, the occurrence of circumstances that extent the core combination of process and participants in a clause was computed. Particularly the locative sub-type was of interest in this context.

While these two systems realize the ideational metafunction, the systems mood and modality can be related to the interpersonal metafunction. The mood options declarative, interrogative and imperative were analyzed as well as more delicate options regarding the verbalization of the interactants, i.e. either speaker or addressee. For modality, especially the frequency of modulation, i.e. the degree of obligation or inclination, is under investigation.

[4] Another reason for choosing this grammar theory was that SFL's focus on language in use as well as the interlocking abstract and more concrete levels correspond to the empirical methodology of a corpus-based study (cf. Neumann, 2002).

Passive is interpreted as a realization of the interpersonal as well as of the textual metafunction. Finally under the textual metafunction, the frequency of the different elements realized in thematic position is counted with a focus on spatial circumstances.

Section 3 gives an overview of the main findings that were gained on the basis of this annotation scheme.

2.3 The Computer tool

The complex research design in combination with a set of features that proved to be rather abstract restricted the selection of a computer tool considerably. The choice fell on O'Donnell's (1995) *Systemic Coder*. This tool facilitates the manual linguistic analysis of a corpus. It allows the definition of a feature hierarchy supports the marking of segments and prompts the relevant categories. Coder organizes the features in a systemic network, hence in form of an inheritance hierarchy. This reduces the annotation effort noticeably: Only the features relevant to the respective stage of annotation are presented. Coder then statistically records the features selected by the annotator. The corpus, enriched with linguistic information, is saved in XML-format and is at disposal for processing in other applications. Coder's review function supports querying for single features or feature combinations.

Although compared to an analysis which is not computer-aided the Systemic Coder speeds up the analysis, the annotation effort is still enormous. Therefore the corpus size chosen for the present study is the maximum that can be handled by one annotator.

3 Findings

The research design made it possible to identify a set of register characteristics. At an earlier step of the study the lexico-grammatical features under investigation were deduced as operationable indicators of the hypotheses that were related to the register variables. The results[5] gained from the analysis of these lexico-grammatical features again were generalized to the level of the underlying register variables. This resulted in register profiles of English and German travel guides as well as a cross-linguistic comparison of these profiles. The consideration of translations produced a generalization that was, again, put in relation to the monolingual profiles.

To identify register features the value for each feature in the register-controlled corpus is compared to the respective value in the reference corpus. After applying the chi-square-test to check significance of the results the value can be described as either a positive or a negative feature depending on whether it is significantly higher or lower than the reference value.

3.1 Monolingual characteristics

English travel guides: English travel guides can be characterized as being rather fact-oriented. These ideational features outweigh interpersonal characteristics which lose importance when compared to the reference corpus.

This becomes apparent in the significantly more frequent use of relational processes in travel guides compared to the other registers included in the study. Material processes occur as often as in the other registers and are thus neutral to the register "travel guide". Mental and verbal process types are underrepresented in travel guides. Figure 2 illustrates these findings.

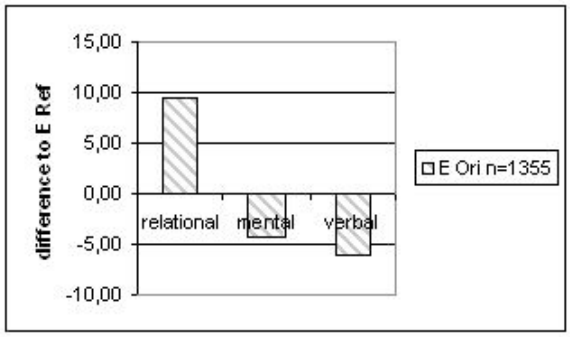

Figure 3: English transitivity structures

The extensive use of circumstances appears to be typical of this register, giving travel guides a more nominal character in comparison to the reference corpus. This underpins findings by Nilsson (2001) who concludes a descriptive character of travel guides from the frequent occurrence of complex NPs in combination with rather simple sentence structures. With respect to the sub-types of circumstantiation, location is the only positive feature in travel guides. The other sub-types are

[5] The results have to be seen in the context of the limitations of the present study. It was not possible to repeat the annotation by a second annotator.

either insignificant or constitute a negative feature. The frequency of relational processes might be correlated with locative circumstances thus relating inanimate things like buildings or places of interest to their spatial situation. In his study of transitivity in topographic procedures Matthiessen (1998) calls this "rest in space".

While interrogatives do not occur, imperatives are significantly less frequent than in the reference corpus and thus represent a negative register feature. This lack of instructions and recommendations in combination with the concentration on expository description by means of relational processes points to a fairly neutral agentive role of the author. He or she takes on the role of giving information rather than that of giving advice.

The verbalization of the interactants was analyzed to make statements on whether travel guides are rather more oriented towards the addressee or towards the content. This feature can be characterized as register neutral. It occurs in travel guides as often as in the other registers. Nevertheless there is a significant shift among the more delicate variants of interactants: The speaker is verbalized less frequently, which makes the direct address of the reader the preferred option. This feature points in the direction of a consultative style, but the significantly frequent use of passive constructions again shifts the balance towards an impersonal mode of expression.

With respect to thematic element, the character of English as a fixed word order language restricts register-specific variation to a certain degree. Still, a significant positive deviation to the register corpus can be observed with respect to spatial circumstances in thematic position. By realizing a spatial circumstance as thematic element before giving the reader some expository information the author follows the chronological order in which the reader moves.

German travel guides: The overall characteristics stated earlier for English travel guides also apply to their German counterparts, albeit to a higher degree. German travel guides strongly rely on experiential characteristics. Interpersonal features recede in favor of ideational characteristics.

These characteristics manifest in a significantly higher value for material processes, while mental and verbal are significantly less frequent than in other registers. That means in travel guides what is felt or spoken is neglectable compared to what is actively done, a fact that is little surprising. Circumstances play an even more important role in German than in English travel guides. This high frequency is one symptom of a nominal style. Particularly the locative type of circumstances is clearly characteristic of travel guides, as the author aims to provide an orientation in space for the reader.

With respect to mood options, imperatives are significantly less frequent in travel guides than in other registers, even more so than in English travel guides. Interrogatives are altogether missing. Thus declaratives are overwhelmingly frequent. As imperative meaning is neither conveyed indirectly in the form of modality it can be said German travel guides do not show a consultative style. Judging from the overall results for mood, the relationship between author and reader can be characterized as one of exchanging information in a neutral way. The author does not give advice from a higher position which would manifest for example in a frequent use of imperatives.

The results regarding verbalization of interactants show a significantly negative value pointing to a concentration on factual information rather than building up a relationship between author and reader. When taking into account the more delicate variants that discriminate between speaker and addressee, a redistribution of the reference values becomes visible. While in the reference corpus mostly the speaker is verbalized, there is a shift towards addressing the reader in travel guides. The value for passive rises significantly in comparison to the registers sampled in the reference corpus. While this feature points to the fact-orientedness of travel guides it may at least partly also be attributed to the function of a varied mode of expression. These last two features – when supposing the function of varied style for passive – can be interpreted as at least a slight turning towards addressee orientation.

The analysis of the thematic element further substantiates statements made earlier in connection with circumstantiation. Parallel to the

reader's orientation in space, the authors of German travel guides very often put a spatial circumstance in thematic position, thus helping the reader to find his/her position first before s/he goes on to tell him/her what s/he can do or see at the place described. This supports Enkvist's (1991) thesis of a "stop-look-see strategy", which he claims is used universally in travel guides.

3.2 Cross-linguistic comparison

The registers in both languages prove to be rather similar. Although for six out of ten hypotheses the cross-linguistic differences are significant, the compared results point in the same direction. The differences show in the degree to which a given feature is distinct in the respective language.

The feature circumstantiation, for instance, was found to be characteristic of travel guides in both languages, featuring a nominal style in this register both in English and in German. In cross-linguistic comparison the difference between both values is still significant. Figure 3 gives an overview of the frequency of circumstances found in the register-controlled and reference corpora.

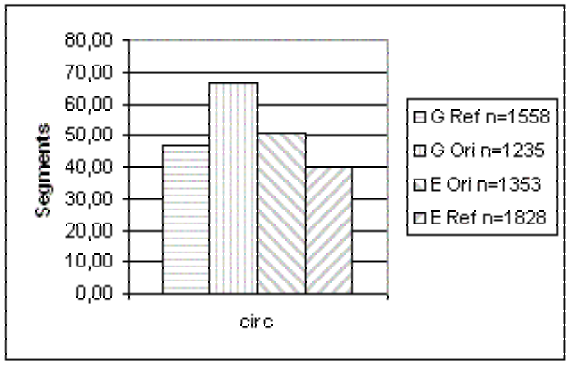

Figure 4. Overview circumstantiation

Here a language-typological explanation comes into play. German can be characterized as a language which strongly relies on ideational meaning, while English seems to be more oriented towards the addressee (cf. House, 1997). Considering the given typological constraints the frequent use of circumstantiation still constitutes a positive feature of both English and German travel guides.

3.3 Diverging realizations in translations

How did translators intuitively deal with the cross-linguistic differences and similarities of the register under investigation? Did they adhere to the target language register characteristics? The results of the study show that the values for translated travel guides mostly lie between the values pertaining to originals in both languages. There are some interesting divergences. These refer to features that are realized in a way which is contrary to both the source and the target language register characteristic.

First, the feature "spatial circumstance" shows no significant difference in cross-linguistic comparison and thus should not pose a problem for translators. Still, the value for this feature in translated travel guides deviates significantly from the source language value.

Secondly, translators verbalize interactants significantly more often than authors of originals in both languages.

Finally, passive constructions occur significantly less frequently in translations than in originals in both languages.

The first of these three features may be an indication of the translator's intention to make the text more explicit. He/she might want to make sure that the reader finds his/her way in space. The other two features show that translators intuitively make their texts more reader-friendly by increasing the direct address of either speaker or reader as well as by using less impersonal passive constructions.

Generally speaking, it can be said that the translated travel guides are more addressee-oriented than originals in both languages and therefore contain more interpersonal traits than the originals in both languages.

This takes us to a possible application of the present study for translation purposes. Translators could use the linguistic information compiled in the present study when deciding how to translate a certain feature – mainly of course if they whish to make their translations more in line with the target language register characteristics. A translator could thus apply the analysis described here to the travel guide s/he is translating. But as this procedure is excessively time-consuming, a computational solution would have

to be realized in order to make this kind of register-specific proof-reading accessible for translators.

4 Computational perspectives

A possible solution could involve integrating register-specific linguistic information in a translation memory. This could be achieved by feeding a register description like the description of travel guides introduced here to a translation memory. The translation process could then be combined with linguistic annotation of the text according to the register features.

This annotation could then be compared with the register description. For the integration of register-specific information in a translation memory the combination of linguistic annotation with the functionalities of a translation memory has to be worked out.

This integration in a translation memory would make sense in the context of a multilingual register lexicon: This lexicon should not be restricted to terminology but should also include grammatical data as well as more abstract information about typical correlations of certain features. The multilingual register lexicon raises questions regarding multi-layer annotation and representation of complex linguistic information for querying.

Finally, the register-specific training of a parser would represent a spin-off of this kind of research.

5 Conclusion

Not only does the register study presented here give an overview of the main lexico-grammatical features of travel guides and their more abstract function. It also proves that intuitive register knowledge is not sufficient for translating in line with the register characteristics of the source or the target language.

The research design of the present study has the following characteristics:

- The corpus size allows general statements on the register under investigation.
- The corpus design with the opposition of a register-controlled corpus to a reference corpus permits the identification of register-specific features against the background of a basis of comparison.
- The study is based on Systemic Functional Linguistics, a linguistic theory that is suitable for cross-linguistic comparison.

This results in a register profile of travel guides in both languages as well as in specific statements on how translators realized texts belonging to this register.

Extensive information about characteristic features of travel guides is of use not only for translators but also for authors of original travel guides.

The present study can also serve as a template for further register studies. This would however require an extension of the size and composition of the two reference corpora: It is questionable to what degree a corpus consisting of 15 registers corresponds to a representative overview of the registers in the given languages. The reference corpus might serve as a monitor corpus that can be extended by each register analyzed in the way described here. As to their quantitative representation, the quantity of the register samples is capable of improvement. For further use of the present research design an extension of the reference corpora is therefore a precondition.

Another perspective for future work is the extension of the analysis to different levels. This could for instance include part-of-speech tagging.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. Mona Baker, Gill Francis, Elena Tognini-Bonelli (eds.): *Text and Technology: In Honour of John Sinclair.* Benjamins, Amsterdam, Philadelphia:233-250.

Douglas Biber. 1990. Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing,* 5/3:257-269.

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

Nils Erik Enkvist. 1991. Discourse Type, Text Type, and Cross-Cultural Rhetoric. Tirkkonen-Condit, Sonja (ed.): *Empirical Research in Translation and Intercultural Studies*. Narr, Tübingen:5-16.

M.A.K. Halliday. 1994. *Introduction to Functional Grammar*. 2nd edition. Arnold, London.

Silvia Hansen. 2002. *The Nature of Translated Text.* Ph.D. Thesis. Universität des Saarlandes, Saarbrücken.

Juliane House. 1997. *Translation Quality Assessment. A Model Revisited.* Narr, Tübingen.

Stig Johansson, Geoffrey N. Leech, Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Department of English, University of Oslo, Oslo.

Sara Laviosa-Braithwaite. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation.* Ph.D. Thesis. UMIST, Manchester.

Christian M.I.M. Matthiessen. 1995. *Lexicogrammatical Cartography.* Intern. Language Sciences Publishers, Tokyo.

Christian M.I.M. Matthiessen. 1998. *The Transitivity of space in topographic procedures*. Macquarie University, North Ryde. (unpublished draft)

Stella Neumann. 2002. *Die Beschreibung von Textsorten und ihre Nutzung beim Übersetzen.* Ph.D. Thesis. Universität des Saarlandes, Saarbrücken.

Tore Nilsson. 2001. *Noun Phrases in British Travel Texts: A Corpus-Based Study.* Ph.D. Thesis. Department of English, Uppsala University, Uppsala.

Mick O'Donnell. 1995. From Corpus to Codings: Semi-Automating the Acquisition of Linguistic Features. *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.* Stanford University, California:120-124.

Erich Steiner. 2002. Grammatical metaphor in translation – some methods for corpus-based investigations. Hilde Hasselgard, Stig Johansson, Bergljot Behrens, Cathrine Fabricius-Hansen (eds.): *Information Structure in a Cross-Linguistic Perspective.* Rodopi, Amsterdam, New York:213-228.

Erich Steiner and Elke Teich. In print. German: a metafunctional profile. Alice Caffarel, James R. Martin, Christian M.I.M. Matthiessen, (eds.): *Systemic functional typology.* Benjamins, Amsterdam.

Elke Teich. 2001. *Contrast and commonality between English and German in system and text.* Habilitationsschrift. Philosophische Fakultät II, Universität des Saarlandes, Saarbrücken.