DE GRUYTER
MOUTON

Silvia Hansen-Schirra, Stella Neumann, Erich Steiner

CROSS-LINGUISTIC CORPORA FOR THE STUDY OF TRANSLATIONS

INSIGHTS FROM THE LANGUAGE PAIR ENGLISH-GERMAN

TEXT, TRANSLATION, COMPUTATIONAL PROCESSING

DE
G

Silvia Hansen-Schirra, Stella Neumann,
Erich Steiner

Cross-Linguistic Corpora for the Study of Translations

Insights from the Language Pair English-German

In collaboration with
Oliver Čulo, Sandra Hansen, Marlene Kast,
Yvonne Klein, Kerstin Kunz, Karin Maksymski
and Mihaela Vela

DE GRUYTER
MOUTON

MIX
Papier aus verantwortungsvollen Quellen
FSC
www.fsc.org
FSC® C016439

ISBN 978-3-11-026029-8
e-ISBN 978-3-11-026032-8
ISSN 1861-4272

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Silvia Hansen-Schirra, Stella Neumann, Erich Steiner

Cross-Linguistic Corpora for the Study of Translations

Text, Translation, Computational Processing

Edited by
Annely Rothkegel and John Laffling

Volume 11

Acknowledgements

This book reports on the outcomes of the project *Sprachliche Eigenschaften von Übersetzungen – eine korpusbasierte Untersuchung für das Sprachenpaar Englisch-Deutsch* ('Linguistic properties of translations – a corpus-based investigation for the language pair English-German') nicknamed CroCo (for Cross-linguistic Corpora) by a team fond of herding exotic fauna. It was funded by the German Research Foundation (DFG) as projects no. STE 840/5-1, STE840/5-2 and HA 5457/1-2.

We are greatly indebted to an anonymous reviewer for a detailed and very constructive report which helped us to clarify many points and to improve the structure of the current volume. The book would furthermore not have made it through the production process without Karin Maksymski's thorough and patient formatting work. Our heartfelt thanks go to her. We gratefully acknowledge our proofreaders' efforts: Paula Niemietz and Sarah Signer worked reliably and very fast.

Last but not least, we would like to thank the members of the CroCo team who collaborated with us on this volume: Oliver Čulo, Sandra Hansen, Marlene Kast, Yvonne Klein, Kerstin Kunz, Karin Maksymski, and Mihaela Vela, as well as several generations of diligent student assistants who helped compile and analyze the CroCo Corpus.

We are finally grateful to our editors at de Gruyter Mouton for supporting us in producing the volume. Needless to say all remaining errors and misconceptions are ours entirely.

The authors
Aachen, Germersheim, Saarbrücken
July 2012

Table of contents

I Texts – The CroCo resource

II Global findings

III Case studies

IV Computational applications

V Generalizations, Conclusions and Outlook

Erich Steiner

1 Introduction

1 Topic

Our topic *Cross-linguistic Corpora for the Study of Translations: Insights from the language pair English-German* covers at least two major sub-domains:

On the one hand, we describe a corpus architecture, including annotation and querying techniques, and its implementation. The corpus architecture is developed for empirical studies of translations, and beyond those for the study of texts that are in some sense inter-lingually comparable, that is to say for texts of similar registers. The compiled corpus, *CroCo*, is a resource for research and is, with some copyright restrictions, accessible to other research projects.

On the other hand, we present empirical findings and discuss their implications for translation as a possible contact variety for the language pair English-German. Beyond our main focus on translation, though, our interest in the longer run is in language comparison and language contact more generally. The text property which is the focus of attention is *relative explicitness* of texts under comparison, and *explicitation* as a possible relationship between source texts and their translations in particular. *Explicitation* has often been assumed to be a specific property of translated texts, alongside possible other properties, such as *simplification*, *normalization*, *levelling out*, *sanitization*, *interference* and *shining through*. It is one of the motivations of the work reported on here to find out whether and to what extent the assumption of such properties can be supported through empirical work, and if so, whether these properties are interesting as influences on language contact phenomena.

Most of the research was undertaken as part of the DFG-Project *CroCo*, a corpus-based investigation into linguistic properties of translations for the language pair English-German.[1]

2 Motivation and goals

The long-term goal of our research is a contribution to the study of translation as a contact variety, and beyond this to language comparison and language contact more generally with the language pair English-German as our object

[1] German Research Foundation (DFG) project no. STE 840/5-1, STE840/5-2 and HA 5457/1-2. For current information cf. http://fr46.uni-saarland.de/croco/.

languages. This goal implies, in our methodology, a thorough interest in possible specific properties of translations, and beyond this in an empirical translation theory.

The methodology developed is not restricted to the traditional exclusively system-based comparison, where real-text excerpts or constructed examples are used as mere illustrations of assumptions and claims, but instead implements an empirical research strategy involving structured data (the sub-corpora and their relationships to each other, annotated and aligned on various theoretically motivated levels of representation), the formation of hypotheses and their operationalizations, statistics on the data, critical examinations of their significance, and interpretation against the background of system-based comparisons and other independent sources of explanation for the phenomena observed. It is our belief that over the past couple of years sufficient progress has been made in corpus technologies and in extracting information on the data to render such an endeavor promising.

3 Theoretical foundations and state of the art

Theoretical foundations of the developments outlined here are to be found

- in the more textually-oriented and linguistically-based strands of translation studies (3.1),
- in models of linguistic variation and register (3.2),
- in the area of corpus design and implementation, and corpus technology more generally (3.3),
- in studies of language comparison and contact, with a focus on language-specific ways of encoding meaning (3.4).

This introduction aims at an outline of the theoretical foundations on the most general level only, because individual chapters will review their own locally relevant state of the art. However, there are some theoretical foundations which form a sort of macro-background for our overall enterprise, and it is this general background which will be sketched here.

3.1 Translation studies

There is a tradition of assumptions *in the more textually-oriented and linguistically-based strands of translation studies* about specific properties of translated texts. According to such assumptions, translations are characterized by specific textual properties; they constitute a "text-type", or "register", of their own (cf. Frawley 1984; Blum-Kulka 1986; Sager 1994; Toury 1995; Baker 1993, 1996; House 1977, 1997, 2002, 2008; Steiner 2001a, 2001b; Teich 2001, 2003; Hansen 2003; Neumann 2003; cf. Fawcett 1997: 100 and Laviosa-Braithwaite 1998 for overviews). These assumptions, and some hypotheses deriving from and specifying them, have been subjected to some initial empirical testing, but nothing approaching an accepted answer to the question embodied in it has been found to date. Furthermore, where some properties of translated texts have been tentatively identified so far, no consensus is in sight as to whether such properties might be mainly due to the specifics of the translation *process*, and in that sense universal to translations, or whether they must rather be explained by recourse to contrasts between the linguistic systems involved and/or by contrasts between the text types, or registers, of the source and target texts and specific translation strategies deriving from those.

Translation studies and linguistics have produced a body of work on language pair-specific and sometimes direction-specific translation problems and translation procedures which provides valuable initial insights on implications of language contrast for translation (Vinay and Darbelnet 1958/1995 in their comparative stylistics of English-French; didactically motivated explorations for the language pair English-German [Friederich 1977; Purser and Paul 1999; Königs 2000], more linguistically founded work by Doherty throughout the 1990s culminating in Doherty 2002 and 2006, and differently House 1977, 1997, both for English-German mainly, or Fabricius-Hansen 1996, 1999 for the language triangle English-German-Norwegian). These studies contribute significantly to our understanding of language-pair specific processes and relationships in translation, without, however, foregrounding the question of whether there are "universal" properties of translated texts. Neither are they methodologically empirical in the stricter sense. By "in a stricter sense" we mean, initially, based on a somewhat larger quantity of data, sampled with some technique aiming at representativeness, and using categories of data which allow a transparent relationship to research questions formulated, and also repeatability of the analysis by different researchers at different places and times.

More recent years have seen the emergence of empirical investigations into universal properties of translations (Baker 1996; Laviosa-Braithwaite 1998; Olohan and Baker 2000; Kenny 1998, 2001; Olohan 2004; cf. House 2008 for a critical overview), where the assumed properties were of the type *simplification*, *normalization*, *levelling out*, *sanitization*, *disambiguation*, *conventionalization*, *standardization*, *avoidance of repetition* and in particular *explicitation* (cf. various contributions in Mauranen and Kujamäki 2004; Saldanha 2008; Englund Dimitrova 2005; and for an earlier summary Klaudy 1998). The property of *explicitness* and the process of *explicitation* will be defined and operationalized in some detail in chapter 4. About the other properties, we would like to say a bit more at this point. *Simplification* usually refers to increasing "readability" of a text, for example by simplifying a type of linguistic structure, e.g. in terms of number of constituent elements of some linguistic unit. Other measures include increased and more explicit punctuation, decreased lexical density or decreased type-token-ratios. *Normalization* refers to a process within which a (translated) text approximates or even exaggerates some norm of the target register it is translated into, always in terms of some selected textual/linguistic feature. Normalization also often means the avoidance of some syndrome of marked features or structures in target texts. *Levelling out* is always predicated of sets of texts, for example when we hypothesize that a set of translated texts, when compared to a set of non-translated texts of a given language and a given register will be composed of texts which are more similar to each other in terms of some (set of) linguistic features, in other words, the range of variation among translations are assumed to be smaller than for otherwise similar original texts. *Sanitization* as a property is assumed to be given when translations avoid affectionally strong language, in particular stigmatized language, relative to original texts. *Shining through* in the sense of Teich (2003: 209–218) means an interference in a translation from its source language, but often in terms of proportionalities and frequencies, rather than simply in terms of individual structures or lexical items as in cases of simple "interference".

We shall meet these, and other, assumed properties of translations as phenomena to be tested throughout our study (especially chapters 5ff.), even though usually our emphasis is on investigating explicitness and explicitation. As far as the assumption of universal properties of translations is concerned, though, our general stance is probably close to the cautious and skeptical attitude adopted in House (2008: 10–12): Much of what is all too loosely postulated as a "translation universal" may well turn out to be either a general property of language (use), or it may be specific for some given combination of languages, it may be specific to one direction between two languages, it may be strongly dependent on register or genre, it may be sensitive to language-change phenomena. In any case, whatever there may be of translation universals, it could be restricted to a highly general level only: one such highly general "universal" may be the fact that each translation necessarily represents an attempt at optimizing conflicting constraints posed by the ideational, interpersonal and textual functional dimensions of encoding – which would be a universal so general that its predictive power would be very limited, unless it were reformulated as much more specific instantiations of that general assumption – something which we believe to be possible in principle. However, and maybe slightly more "universalist" than the stance adopted in House (2008), if it could be shown that an assumption about de- and re-metaphorization in translation-oriented psycholinguistic processing of the type made in Steiner (2001a: 170ff., 2001b: 15ff.), Hansen (2003: 118–125), and summarized again here in chapters 7 and 14, is valid, then this could be the source of a property shared by all translated texts relative to non-translated ones, even though the kind and extent of explicitation would be strongly sensitive to language-pair specific and direction-specific factors. We have begun investigations of such de-/re-metaphorization processes in process-oriented experiments (Alves et al. 2010) in which we focus on interactions between variable translation-units (in a processing sense) and degrees of metaphorization, where "metaphorization" is always to be understood as "grammatical metaphor" in the sense of Halliday's "Functional Grammar" (Halliday 1985: 319; Halliday and Matthiessen 2004: 586; see also chapter 7).

However, independent of whether or not any of the assumed properties of translated texts are general across more than two languages, genres, registers, we see their particular research potential in their relationship to feature- and property-based approaches, to contrastive linguistics, language contact studies and issues to do with processing. Languages and texts can usefully be contrasted in terms of properties; they can be assumed to influence each other in terms of such properties. Dynamic processes such as language change and language processing can be modeled on properties – and we would hope to be able to interface with empirical research traditions currently being developed in these areas, some of which we shall address below, and again towards the end of this book.

The line of argumentation positing properties of translated texts, even though we discuss some aspects of it critically here, represents progress towards an empirical research methodology, as well as an increased focus on properties of translated texts due to the translation process. While it thus has paved some of the way for our own goals, some of it suffers, in our view, from impoverished linguistic modeling: its essentially corpus-driven, rather than theory- or model-driven, methodology and the linguistically low level at which phenomena are operationalized make it very difficult to address higher and more theoretically meaningful linguistic levels, lexico-grammar, semantics and text/discourse in particular. It is therefore also no coincidence that within this line of research, the valuable insights of language typology and typologically-based linguistic comparison are not exploited in explanations of the phenomena observed.

So far, then, we are claiming that on the one hand, the linguistically more informed studies of translations mentioned above would gain from a more empirical methodology, and from taking the process of translating as a mode of text production more seriously as a source of explanation. On the other hand, existing and methodologically more empirical studies of translations would need much more of an influence of linguistic models of variation and register, and of studies of language comparison and contact, with a focus on language-specific ways of encoding meaning, in order to be able to make a contribution not only to our awareness of isolated and theoretically sometimes arbitrary features which characterize translations, but rather to our understanding of translations as texts, and to translations as a possible contact variety between languages. Both research strands could gain substantially from devoting more explicit attention to the areas of corpus design and implementation. In these areas we hope to be able to make a contribution, and we would like to start with computational design and implementation of corpora, before turning to the linguistic basis for the modeling to be suggested here.

3.2 Models of linguistic variation and register

In terms of general awareness of tools and architecture in corpus technologies, we are, like many other projects, indebted to *models of linguistic variation and register* (cf. Biber 1988, 1995; Biber, Conrad, and Reppen 1998) and to work on languages in contrast (cf. the SPRIK project in Oslo, for example Johansson and Oksefjell 1998). As part of this legacy, we have attempted to integrate statistics for the evaluation of the significance of results where appropriate (cf. Biber, Conrad, and Reppen 1998; Butler 1985; Oakes 1998).

As for models of linguistic variation and register, we obviously need an understanding and some modeling of how, and along which dimensions, texts can be classified as similar or different. A "lean" variant of such a model is the notion of "register" as used in Biber's work, or in Biber et al. (1999). A richer and theoretically more committed variant is the notion of "register" in its original theoretical context in Systemic Functional Linguistics (cf. Halliday, McIntosh, and Strevens 1964: 87–88; Halliday and Hasan 1989; Matthiessen 1993). Translation studies have a substantial history of using this notion (cf. House 1977, 1997: 196; Hatim and Mason 1990; Hansen 2003: 23; Neumann 2003: 16; Steiner 2004b: 11), and we have used it in various degrees of theoretical commitment (for an advanced example cf. Neumann 2008). In a "lean" version, register theory can be seen as not much more than some form of text typology, and quite a few of our studies use it just in this "lean" version. In a more theoretically-committed version, the dimensions of variation of this typology systematically link up with the linguistic system and its multi-functional grammar on the one hand, and with the context of culture on the other. The modeling translation within this overall architecture can be seen in Matthiessen (2001), Teich (2001) and Steiner (2001a).

3.3 Corpus design and implementation

In the area of corpus design and implementation, we have imported and further elaborated techniques from multi-layer corpus architectures, annotation, treebank technologies and information extraction on data in such corpora. A fundamental characteristic of our methodology is that we are not working on raw corpora, but on multi-layer annotated corpora (with and without alignment), bridging the gap between the formulation of hypotheses on higher levels of linguistic structure and their operationalizations in instantiated texts (cf. Hansen 2003; Teich 2003; Neumann 2008).

On a more technical note, existing corpus tools have been used – ranging from automatic to semi-automatic to computer-assisted manual annotation and alignment (cf. Lüdeling and Kytö 2008, 2009 for an overview). These include some tools that are language-independent, but the trade-off for the high degree of flexibility is a low degree of automation. Other tools enabling automatic or interactive annotation require language-specific training, which raises the question of comparability across multilingual annotations (cf. Neumann and Hansen-Schirra 2003).

The multi-layer annotation and alignment of the CroCo Corpus allows us to view the annotation in aligned segments and to pose queries combining different layers. 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. For this purpose, two technical requirements must be met: the exploration of the integrated data (i.e., simultaneous viewing of the different levels and searches across levels) and integrated processing, e.g. for the discovery of correlations across layers. These requirements are met by using stand-off annotation at each layer on the one hand (cf. McKelvie et al. 2001) and alignment of base data across the layers on the other (Bird and Liberman 2001). Developed for multi-layer annotation in XML, the XML Corpus Encoding Standard (XCES) guarantees exchangeability and consistency since predefined XCES Schemas, DTDs and XSLT scripts can be used (Ide, Bonhomme, and Romary 2000). For efficient querying, the annotation and alignment information can be stored in a relational database (cf. Cassidy and Harrington 2001), which allows the integration of hierarchical annotation layers. Chapters 6–11 will show that empty alignment links, crossing alignment lines as well as the combination and exclusion of annotation tags are important for the linguistic exploitation of the CroCo Corpus. The results of such combined queries can then be interpreted in terms of linguistic properties of translated text.

3.4 Studies of language comparison and language contact

Let us now turn to *studies of language comparison and language contact*, with a focus on language-specific ways of encoding meaning:

Language contact is the situation in which languages, or rather, instantiations of language systems through their speakers, influence each other synchronically in shared socio-semiotic contexts (classical accounts include Weinreich 1953; Thomason and Kaufman 1988; Oesterreicher 2001; a more recent account is given in Siemund and Kintana 2008). This is complementary to the historical axis, along which genetically related languages are in contact through time. Language contact applies to varieties within languages, as it does to different standard languages. Major topics of research are (cf. Thomason and Kaufman 1988: 65–100):

- the interplay between synchronic contact and genetic inheritance
- linguistic vs. socio-cultural constraints on interference
- analytic frameworks for contact-induced language change (linguistic levels of change; borrowing vs. interference through shift; predictive power of the frameworks, external vs. internal explanations)
- language maintenance
- normal vs. exceptional transmission (creolization, pidgins)

In an attempt to generalize on the strength and on linguistic levels of language contact, a *borrowing scale* is postulated, ranging from lexical borrowing only through slight structural borrowing, moderate structural borrowing and finally to heavy structural borrowing. Most studies to date have focussed on lexical items and/or grammatical structures, rather than on features or properties of the linguistic systems and instances (discourses, texts) involved, although both perspectives have often been acknowledged as relevant (cf. also Heine 2008: 37 in his Figure 1 on contact-induced linguistic transfer).

Multilingualism is usually predicated either of individuals, or of linguistic communities as socio-cultural formations, or else of discourses/texts (for a representative and comprehensive survey cf. Auer and Wei 2007). In the first sense, studies of multilingualism are often carried out as studies of language development/acquisition of several languages in one speaker (bilingualism, trilingualism, etc.). In the second sense, they are targeted at linguistic communities and are methodologically situated in sociolinguistics. A terminological distinction which reflects this division is that between *bilingualism* as referring to the individual, and *diglossia* as referring to communities. In the third sense, there are a few strands of research into *multilingual text production* (cf. Matthiessen 2001; Steiner and Yallop 2001; Teich 2001; Steiner 2004a, 2004b, 2005a, 2005b, 2005c), cross-cultural pragmatics (House 1997, 2002) and information structure across languages (Hasselgård et al. 2002; Fabricius-Hansen and Ramm 2008), in which *multilingualism* is treated as a property of discourses which are assumed to have interesting and typical properties compared to *monolingual* discourses (cf. several contributions in Franceschini 2005, in particular von Stutterheim and Carroll 2005). If we say that *discourses* are multilingual, then we imply that they show special *discourse properties* of *directness vs. indirectness*, *orientation towards self vs. other*, *orientation towards content vs. interaction*, *explicitness vs. implicitness*, *routine-orientedness vs. ad-hoc formulation* (as e.g. in House 1997: 84), ultimately to be realized in lexico-grammatical phenomena such as interference, borrowing, code-/language switching, special metafunctional orientations in terms of ideational, interpersonal, or textual biases, directness, density, explicitness, and others. These discourses thus instantiate specific contact varieties, or registers. In our own research, we regard translations as an important venue of influence in language contact (cf. Frawley 1984; Baker 1996 for translations as a special text type or even code). But this venue of influence is additional to, and different from, more traditional venues of contact through borrowing or interference. It is less obvious, the resulting varieties are superficially close to native ones, and it applies intra-lingually, across registers, as much as it does inter-lingually.

Investigations of *multilingualism* are meaningful on all of the levels mentioned above, provided the empirical claims that are being made by the ascription of the property to individuals, communities, or discourses are clear. Furthermore, in the case of discourses, it must be clear whether empirical claims are made about properties on the level of text/discourse, or else on the level of lexico-grammar – or about both of them. A multi-functional and feature-based perspective will usually encompass the discourse-oriented perspective, at least as an important component, and certainly as a prominent object of study. *Multilingualism* of discourses can be assumed to be a property which is both a result of, and an environment for, language contact and change.

In a first attempt to characterize our own research efforts relative to the substantial tradition of research briefly characterized so far, it will be obvious that they rely for their modeling to some extent on Systemic Functional Linguistics (cf. Halliday and Hasan 1976; Halliday 1978; Halliday and Martin 1993; Halliday and Matthiessen 2004). We have additionally drawn on comparative and typological perspectives with some functional leanings (cf. Hopper and Thompson 1982; Hawkins 1986; Thomason and Kaufman 1988; Biber 1995; Simon-Vandenbergen and Steiner 2005; Traugott and Dasher 2005) and on insights from certain strands in translation studies, contrastive linguistics and cross-cultural pragmatics (Doherty 1996, 2002, 2006; Fabricius-Hansen 1996; House 1977, 1997, 2002). In