10

Translation, corpus linguistics and cognition

Stella Neumann and Tatiana Serbina

10.1 Historical background: The corpus approach in Translation Studies

Empirical approaches to translation started to gain ground in the mid-1980s. A first wave of process-based studies (e.g. Königs, 1987; Krings, 1986) appeared around that time and were soon followed by first papers proposing a corpus-based approach to translation (e.g. Johansson & Hofland, 1994; Toury, 1995/2012/1995). Particularly inspired by Baker’s (1993) paper advocating the creation and analysis of translation corpora, a corpus-oriented approach quickly gained ground in Translation Studies, yielding a flurry of reports of typical properties of translated texts (Hansen, 2003; Kenny, 1998; Laviosa-Braithwaite, 1996; Mauranen & Kujamäki, 2004b; Olohan & Baker, 2000; Teich, 2003, etc.), giving rise to what has been referred to as corpus-based Translation Studies (CBTS). Until recently, process- and product-based approaches were carried out more or less independently of each other, although both approaches were obviously concerned with translators’ behaviour.

In earlier CBTS, cognition was not explicitly addressed as a source of potential explanations of corpus findings, not least because corpus linguists were traditionally at least agnostic towards the role of cognition in language. As Anderman and Rogers (2008, p. 13) point out, the underlying idea of corpus linguistics was initially to study language by observation rather than through introspection or experimentation. Discussing the notion of universals in translation, Mauranen and Kujamäki (2004a, p. 2) refer to a distinction in general linguistics between “universals which can be traced back to general cognitive capacities in humans, and those which relate linguistic structures and the functional uses of languages” and then go on to link them to different characteristics of translated language, namely cognitive translation processes, social and historical determinants of translation, and the typical linguistic features of translations. Claims like these suggest that corpus-based translation scholars were perhaps not primarily interested in but at least aware of cognitive aspects of translation, even if they did not refer to cognition explicitly. Instead, they variously referred to translators’ behaviour during the translation process, subconscious processes or understanding. Steiner (2002, p. 216), for example, discusses three sources of explanation for the specific properties of translations observed in corpora, namely contrastive differences in the language pair, differences between registers, and individual differences in the process of understanding during translating. He also refers to “internal ‘fatigue’” (Steiner,
One way of operationalizing such references to understanding is explained by Hansen-Schirra:

Cognitively speaking, the decoding phase during translation, which includes comprehension, understanding, and conceptualization of the source text, triggers the phenomena introduced above [translation properties]. Translators transfer their own understanding and interpretation into the target texts, which make them more explicit and easier to understand. In case of ambiguities or complex structures, this results in resolving, unpacking, and deconstructing [sic] the source text structures (including defective and unexpected text passages) and translating them with unambiguous, clearer, and optimized ones.

Chesterman (2010, p. 43) explicitly argues that the most obvious cause for typical features of translations found in corpora is a cognitive one, “something in the mind of translators that affects the way they process texts simultaneously in two languages”. This is not altogether implausible, as corpora provide a window on systematic patterns of translators’ behaviour.

The goal of this chapter is not to give a comprehensive overview of CBTS in general but to focus on the interplay between corpus linguistic and cognitive approaches in Translation Studies (see also Rodríguez-Inés, 2017), assuming that both approaches are novel in their interest in the empirical facts of translation. The chapter is organized as follows. First, we very briefly introduce some sources of corpus-based approaches in Translation Studies and some central notions of corpus research. Section 10.2 will be concerned with a detailed discussion of various aspects of corpus-based research in Translation Studies and its interaction with cognitive approaches. Finally, in Section 10.3, we will summarize more recent developments in CBTS and conclude with some recommendations for further reading in Section 10.4.

A precursor of the corpus-based approach in Translation Studies can be found in Even-Zohar’s (1979) polysystem theory (see Baker, 1993; Kruger, 2012; Øveras, 1998). It emphasizes a matter-of-fact perspective on the empirical reality of texts and consequently strongly rejects a focus on particularly valued texts. Even-Zohar’s theory characterizes the individual literary work as part of systems that capture traditions, genres, etc. Translations are seen as being part of, and potentially playing a central role in, the literary system. Linking the approach to Holmes’ outline of Translation Studies, Gideon Toury (1995/2012) went on to work out what an empirical or—in his words—descriptive approach to translation would have to cover. Specifically, he discussed potential laws that should be empirically investigated. This was, in turn, picked up by Baker (1993), who linked Toury’s programmatic claims as well as small-scale studies of traditional literary corpora, such as Vanderauwera (1985), to Sinclairian corpus linguistics (Sinclair, 1991).

Initially, the term “corpus” referred to a collection of texts used for some kind of text analysis, typically in literary studies. With the advent of corpus linguistics (see Johansson, 2008 for a historical overview), the concept was narrowed down to large, electronic collections of texts compiled in a principled way to achieve balance and, ideally, representativeness of some population of texts, which can be used for systematic empirical analysis. Translation corpora, in a more specific sense, are corpora containing translated texts, i.e. products of the translation process (see Anderman & Rogers, 2008 for the sources of corpus linguistics in Translation Studies). For the purposes of comparison, quite often they also include non-translated texts, either the source texts of the included translations or originals in the same language as the translated texts.

In a parallel development, Stig Johansson, an early proponent of corpus methods in linguistics, initiated the compilation of some of the best-known translation corpora (see later). His main interest was linguistic rather than translation-related, focusing primarily on contrastive
aspects for which he used translations as a tertium comparationis. Similarly, Fabricius–Hansen (1999) demonstrated how using corpora could help shed light on translation-related linguistic questions. Although Fabricius–Hansen in particular based her study on assumptions about the (incremental) processing of complex sentences, these approaches are not directly linked to cognitive aspects of translation.

One main reason for collecting translation corpora is to investigate the particular linguistic traits making translations distinguishable from non-translated texts (for collections of CBTS see, e.g., Fantinuoli & Zanettin, 2015; Ji, 2016; Kruger et al., 2011; Laviosa, 1998; Mauranen & Kujamäki, 2004b; Oakes & Ji, 2012). However, the compilation of a sufficiently large corpus is costly and subject to copyright issues. As a consequence, corpora compiled for individual studies are usually fairly small. In addition to corpora collected for the purpose of individual corpus investigations, a number of well-known corpora have been used in more than one study. These include the Translational English Corpus\(^2\) (TEC; Baker, 1996), the English–Norwegian Parallel Corpus (ENPC; Johansson & Hofland, 1994), the German–English CroCo Corpus (Hansen–Schirra et al., 2012), the Dutch Parallel Corpus (Macken et al., 2011), the MeLLANGE Learner Translator Corpus\(^3\) (Kunz et al., 2010), the interpreting corpus EPIC (Monti et al., 2005) and the ZJU Corpus of Translational Chinese\(^4\) (ZCTC; Xiao & Hu, 2015).

As corpora are by definition samples of the overall population of all texts, carefully considering design criteria in terms of balance and representativeness (coverage of different authors and publishers, spoken and written language, a range of genres/registers, and all this in relation to translation) is of great importance. The actual compilation according to the specified criteria may involve compromising on certain criteria due to unavailability of texts. The optimal size of a corpus is not an entirely straightforward question and largely depends on how vigorously aspects of balance have been taken into consideration in the previous steps. A well-balanced corpus may reflect the distribution of linguistic features, even if it is comparatively small (see Biber, 1990). Likewise, a very large corpus may turn out to be uninformative if texts are sampled mainly from the same sources or if the distribution between parts of the corpus is imbalanced. Usually, corpus size is given in number of words; however, this may be inappropriate, especially for translation-related research questions, as translations have been shown to deviate from their source texts in length (Frankenberg–Garcia, 2009). Moreover, they display differences in length depending on the translation direction (Neumann & Hansen–Schirra, 2012). This is likely to be a more general phenomenon in translation, resulting in difficulties in determining the appropriate size in number of words.

Compilation of modern corpora usually involves a range of manual or automatic preprocessing steps to ensure computer readability of the data, including encoding of specific features such as formatting, labelling errors, tokenization of the stream of symbols into words, punctuation marks, etc., automatic enrichment with linguistic annotation of word classes (so-called part-of-speech tagging) and occasionally also other types of linguistic information. If the corpus includes matched originals and translations, it will usually also include sentence alignment, and occasionally alignment of smaller units (on the importance of analysing aligned units see Steiner, 2012). Various types of comparison used in Translation Studies (see Hansen–Schirra & Teich, 2009) have their advantages and disadvantages. Comparing original texts and translations in the same language (in so-called comparable corpora) ensures the comparability of linguistic features. However, this makes it difficult to specify the exact factors that could explain the observed phenomena in translations: they could be triggered by the specific linguistic facts of the corresponding source-text unit, they could be due to some general feature of the source or target language (including their registers), or they could be brought in by the translator without any traceable reason in the texts or languages. This problem is resolved by parallel corpora, i.e.
corpora of matching source and target texts, especially if the texts are aligned at least at sentence level. However, parallel corpora introduce the challenge of having to determine comparable features that capture phenomena which are formally and/or functionally equivalent. Corpora also differ as to whether both translation directions in a language pair are covered. Coverage of both directions is very useful as a first indication of how pervasive a certain translation phenomenon is: if some assumed universal feature of translation can only be found in one translation direction, this already calls into question the universal character of the feature. However, bi-directional corpora stratified into registers raise the question of comparability of registers. It is quite likely that corresponding registers in a language pair will display some incomparability. In such cases, claims about directionality effects can be called into question.

Although the definition of corpus in corpus linguistics is quite specific, the ways such corpora are used still cover a fairly wide methodological space. Tummers et al. (2005) point out that there is a range from what they call corpus-illustrated research through descriptive operationalization to explanatory research using bivariate or multivariate statistical analysis. Some earlier translation-related studies can arguably be classified as corpus-illustrated research, in which the corpus is used as a resource for examples to underpin the (theoretical) linguistic discussion. Most translation-related studies fall into the descriptive category; while some of them include statistical significance testing, others focus on the comparison of frequency counts, often simply represented by bar charts. Only a few recent studies use multivariate statistics (e.g. Delaere et al., 2012; Hu et al., 2019; Ji, 2016; Lapshinova-Koltunski, 2017; Oakes & Ji, 2012).

In a stricter methodological sense, corpus-based approaches can be categorized as observational studies. Experimental studies involve control over unwanted factors and deliberate manipulation of some variable assumed to have an effect on some phenomenon that is to be explained (Butler, 1985, p. 65). Observational studies, by contrast, draw on data outside the researcher’s control, where it is impossible for the researcher to manipulate a variable in order to measure its effect on the dependent variable. This has implications particularly for control over confounding factors, which would allow robust explanations of phenomena in terms of the manipulated independent variables. Corpus linguists cannot manipulate independent variables because they are not able to control the generation of the data they are analysing. Rather than investigating the effect the independent variable has on the dependent variable, the corpus linguist investigates correlations between the different variables. What may be of help in some cases is metadata: the more information we have about the author(s), the text production circumstances, etc., the more possible it will be to filter the texts accordingly, thus simulating the manipulation of independent variables. But in many cases, the corpus linguist will simply face a different task than the experimental researcher in comparing two measurements of the same linguistic variable in different data sets and interpreting the results in light of the characterization of the data sets.

Laboratory experiments are sometimes criticized for the extent to which potentially confounding factors are excluded. In extreme cases this may lead to laboratory artefacts which cannot be replicated under authentic conditions. In contrast, less controlled experiments allowing more authentic conditions will yield results that are difficult to link to one particular explanation, because various factors could lead to the observed behaviour. A major advantage of corpus-based investigations is the authenticity of the data, including translational data. However, this is not to say that corpora are necessarily neutral in this respect. Establishing what is authentic or, more generally, appropriate data is at the corpus compilers’ discretion. They may choose to consider certain types of language use as inadequate, incorrect or not accepted in a given culture (see Mauranen, 2004, p. 73) and consequently only to include esteemed texts or texts judged to display a certain quality. In the context of CBTS, for example, a relevant question is whether to
include non-professional translations, translations of texts written by non-native speakers, etc., as all these factors potentially affect the translation product. The less restrictive corpus compilers are, the more likely the corpus will be to reflect the empirical reality of the translations the consumer is actually exposed to. However, variation in the collected texts will be reflected by unexplained variance in the corpus study if the corpus does not contain relevant information on the translator and details on the translation process. For instance, providing information about the experience and training of the translators of corpus texts will give access to a comparison of translation products based on notions such as expertise. If such information is not available, corpus texts will still reflect differences in the translators’ expertise, but the variance in the data cannot be linked to this potential explanation. Regardless of these design considerations, a main characteristic of corpus studies is that they involve a certain amount of idealization (McEnery & Wilson, 2001, p. 77): The search will retrieve exactly those patterns that match the query. If the linguistic phenomenon is characterized by high variability, it is likely that the corpus analysis will not retrieve all relevant cases.

### 10.2 Core topics

Existing (published) translations are rarely used as material to investigate the cognitive processes the translator engages in, since they reflect the result or final product of cognitive processing taking place during translation task execution. The actions that lead up to this result, i.e. the translator’s behaviour reflecting the processing, are no longer observable in the product. Although it is possible that a translation contains particular linguistic traits that are identifiable as a result of, say, increased cognitive demand during translating, it is also possible that this demand is not apparent in the product because the final translation is the most likely one of a given source-text segment despite having been produced in more than one attempt, i.e. in a prolonged process (see Section 10.2.3). Alternatively, corpus findings might hint at increased cognitive demand during task execution, e.g. in the form of simplified grammatical structures, but process-based experiments do not retrieve behavioural correlates of increased demand. Instead, they suggest automatized translation behaviour (Heilmann et al., in press). Halverson (2015, p. 323) points out that “[t]he basic consensus emerging from this methodological discussion is that while corpus data provides evidence of aggregate patterns that must be explained, this type of data alone does not provide direct evidence of online cognitive processes”.

It therefore appears plausible that cognition during translating is usually investigated with the help of behavioural measures such as eye movements and keystroke logging, which could be used as a window to cognitive activity. Nevertheless, as will be shown later, corpus studies are still useful for generating or testing assumptions about cognitive processes during translating. The discussion of the interaction between CBTS and cognition will be structured into three areas. The first area discusses the use of corpora as a means of making and testing assumptions about cognition in the above sense, that is, without necessarily committing to theoretical claims about cognition. The second area examines the use of corpora against the theoretical background of cognitive linguistics. Approaches in this framework do not necessarily adopt a corpus method as the default method to test claims about mental processing. The third area deliberately bridges the gap between experimental and corpus-based methodologies by triangulation or by developing integrated research designs. In the next sub-sections, we will discuss these three areas in more detail, concentrating on the potentials and limitations of each approach as a way of gaining insight into translation as a cognitive activity.
10.2.1 Corpus-based inferences about cognition based on eliminating other sources of explanation

The first of the three areas, i.e. corpus-based inferences about cognition, includes studies that draw on cognition or some related concept such as the (unconscious) process of understanding to explain instances of language use which cannot be explained by other factors like contrastive differences or register characteristics. Since corpora do not give access to the actual behaviour during task execution, inferences are thus made by eliminating the other potential explanations.

The question of whether the specific phenomena observed in translations are actually universal, i.e. whether they occur irrespectively of the translator, the languages involved, the specific linguistic feature, etc., has triggered much debate in CBTS (see among others Becher, 2010; Bernardini & Zanettin, 2004; Chesterman, 2010; Zufferey & Cartoni, 2014). However, the sheer fact that machine learning approaches achieve high accuracy in classifying translated versus non-translated texts (see, e.g., Baroni & Bernardini, 2006; Volansky et al., 2015) suggests that there must be something peculiar about translations that computers can pick up on with the help of certain—usually fairly shallow—features (for more detail see Evert & Neumann, 2017). To some extent, this could be due to the fact that features of translations are probabilistic in nature and more specifically conditional rather than deterministic; therefore, simple frequency counts are likely to be too concrete (Toury, 2004). At the same time, as Toury points out, claims that are too general run the risk of not being informative and therefore have to be made as specific as possible. This, in turn, cannot be operationalized by investigating individual features. It does not do justice to the complexities of translation, as “there is probably no single variable affecting translation which cannot be enhanced, mitigated, maybe even offset by the presence of another” (Toury, 2004, p. 25). Probabilistic concepts that have “the possibility of exception” built into them (Toury, 2004, p. 29, emphasis in original) are to be preferred to the notion of universals. Therefore, it appears advisable that one should be careful when labelling the observed peculiarities as translation universals for as long as we do not have compelling evidence of the universal character of those peculiarities. In view of such considerations, we will refer to phenomena such as explicitation of information implicit in the source text, adaptation of features to the norms of the target language (normalization), simplification of linguistic structures (Baker, 1996) and shining through or interference of the source language (Mauranen, 2004; Teich, 2003) as properties or features of translations rather than as universals.

The main task of corpus-based studies is, then, to examine the corpus findings and link them to plausible sources of explanation. It seems that, among the potential explanations, the ones that are most immediately related to the translation process are the ones translation scholars are actually interested in. Baker (1993, p. 243), for example, makes a distinction between features linked to “the nature of the translation process itself” and features resulting from the “confrontation of specific linguistic systems”. She explicitly characterizes the assumed features of translation as “a product of constraints which are inherent in the translation process itself” (ibid.). While this is a far cry from making claims about cognitive processing, it at least closes in on the translation process during which the translator displays a particular behaviour. In corpus analyses, isolating explanations linked to the translation process itself can happen in different ways. The most indirect way is to address the suspected feature right away and design the study in such a way that it precludes any other explanation. A more direct way involves the explicit discussion of various sources of explanation and closing in on cognitive aspects by eliminating other explanations. Lastly, studies might also work the other way around, i.e. rather than eliminating other explanations, rule out cognitive aspects as an explanation.
In an early corpus study, Olohan & Baker (2000) link the assumed translation property of explicitation to what they call subconscious processes\(^5\) in the title of the paper. The concept of unconscious processes or unintentional language use is not explained, but we might reasonably assume that the authors are interested in indications of mental processes. In a sense, they even strengthen their assumptions about mental processes by specifically referring to unconscious phenomena. Unfortunately, Olohan & Baker (2000) do not go into more detail about the extent to which their findings are actually evidence of such mental processes. This evidence can be arrived at indirectly by eliminating other potential explanations. Contrastive differences can be ruled out, since the Translational English Corpus contains English translations from a wide range of source languages. They report statistically significantly more occurrences of the optional complementizer \textit{that} in translations than in comparable original English texts and come to the conclusion that “inherent, subliminal processes” must lead translators to spell out the implicit grammatical information in their target texts. In a follow-up study, Olohan (2002, p. 166) corroborates Olohan and Baker’s (2000) findings with further features and comes to the conclusion that explicitating optional syntactic elements “may be considered subliminal or subconscious, rather than a result of deliberate decision-making of which the translator is aware”.

In a study of interference in translation, Mauranen (2004) explicitly addresses the simultaneous activation of source- and target-language systems in the brain during translating as a form of bilingual processing. However, she does not link the results of her analysis of the Corpus of Translated Finnish, which contains Finnish translations where the source language is specified, to conclusions about processing of the two languages. A special case of elimination of factors not linked to the translation process is a priori ruling out certain factors as irrelevant or straightforward.

Addressing different possible explanations in a more direct way, Hansen-Schirra et al. (2007) examine various combinations of subcorpora in the CroCo Corpus and features allowing them to eliminate language contrasts, register differences and understanding as sources of explanation, respectively. Their paper also covers features for which contrasts or register differences are identified as the explanation (findings on pronominal reference are explained by contrastive differences), thus allowing the authors to link findings on lexical density to what they call the translation process, i.e. understanding (Hansen-Schirra et al., 2007, p. 225). It should be noted that, like the other papers cited earlier, they also remain vague about whether they actually refer to cognitive processing. Zufferey and Cartoni (2014) examine a set of different factors in their study of the potential explicitation of connectives, but this set does not include cognition as a factor.

A more complex research design in combination with multivariate statistical techniques can add reliability to the interpretation of corpus findings. The computation of occurrences of 27 lexicogrammatical features in the bi-directional CroCo corpus allows Evert and Neumann (2017) to question the assumed parallel activation of both languages during translation as suggested by Mauranen (2004). They show that translations into German systematically display more similarities with their English source texts than translations and source texts in the opposite direction. This leads to the conclusion that the cognitive assumption of parallel activation cannot be the sole explanation for shining through because of the detected directionality effect of this phenomenon. If the cognitive explanation of parallel activation of both language systems were the main—or only—explanation of shining through, the effect should be virtually identical in both translation directions. While parallel activation appears to be a plausible explanation for translators’ behaviour in general (see empirical evidence reported in Balling et al., 2014), other explanations such as social factors have to be considered, too. This shows that assumptions about cognition can be tested with a corpus-based approach. In principle, this result could also be
achieved in an experimental study; however, the necessary size in terms of participants and their specification (coverage of both translation directions) virtually precludes experimental testing.

### 10.2.2 Corpus-based investigations of cognitive linguistic concepts

According to Halverson (2013, p. 66), who has been one of the main proponents of introducing cognitive linguistic explanations (e.g. Goldberg, 2006; Langacker, 2008) into Translation Studies, “the most striking consequence of adopting a cognitive perspective is that it places human cognition and human agency at the centre of the causal picture”. Halverson (2013) and Tabakowska (2013) both suggest that the explanatory power of a cognitive linguistic theory can provide a better understanding of the cognitive factors playing a role in the translation process.

Over the last few decades, cognitive linguistics has been increasingly applied within Translation Studies to provide theoretical explanations for various translation phenomena (e.g. Halverson, 2017; Szymańska, 2011). In cognitive linguistic approaches, linguistic phenomena are explained by means of general cognitive processes, namely categorization, analogy, chunking, schematization and association (Langacker, 2008, p. 16). Entrenchment, one of the central concepts of cognitive linguistics, is primarily related to the process of chunking. It is assumed to take place when “through repetition or rehearsal, a complex structure is thoroughly mastered to the point that using it is virtually automatic and requires little conscious monitoring” (Langacker, 2008, p. 16). The link between cognitive entrenchment of a linguistic construction and its use as observed in corpora is established through the so-called “corpus-to-cognition principle” (Schmid, 2010, pp. 101–102): frequency of use identified by means of corpora is assumed to give insight into the level of entrenchment of the analysed linguistic patterns in the speakers’ minds. Stefanowitsch and Flach (2017, pp. 102–103) argue that there are two possible perspectives on the relationship between corpora and cognition: corpus-as-output and corpus-as-input. According to the former perspective, the corpus is seen as a sample of linguistic usage, which reflects the cognitive representations of linguistic phenomena, while the latter perspective assumes that the corpus is a sample of linguistic exposure, which forms the cognitive representations over the course of life-long language acquisition. On a more general note, the authors point out that entrenchment is a theoretical construct and as such, cannot be measured directly by any method. Instead, the measurement of entrenchment depends on its operationalization. Since this cognitive concept is closely related to the frequency of usage and the frequency of exposure, corpus linguistic quantitative measures appear to be possible means of operationalization (Stefanowitsch & Flach, 2017).

Halverson (2003, 200) applies the concept of entrenchment to the study of translations, arguing that the repetition of certain translation routines leads to their entrenchment and to easier activation during the translation process. For her model of the translation process, Halverson combines the theoretical framework of cognitive grammar (Langacker, 2008) with models of semantic representation of lexical material in bilinguals (e.g. de Groot, 1992). According to Halverson’s model, source-text (ST) material—the lexical/phonological representation of a linguistic item—activates the relevant parts of the cognitive semantic network available to the speaker, which is shared for languages she/he is familiar with. Depending on the number of shared semantic elements between the corresponding structures in the source and target languages (SL and TL), the phonological representation in the TL, which is required to produce a translation, is assumed to be more or less readily available. Halverson argues that, even if the SL items do not directly activate any particular TL elements, the more often the translator actively searches for the appropriate translation solution, the more entrenched the corresponding connection between the two nodes in the semantic network becomes (Halverson, 2003, p. 215), making the selected translation strategy more automatic.
Halverson further suggests that some of the nodes in the cognitive semantic network are more central or salient than others. Halverson (2003) refers to the gravitational pull hypothesis to provide a cognitive explanation for the corpus linguistic findings discussed in the literature, such as translation properties. In its latest version, the gravitational pull hypothesis is divided into three possible sources of translational features, namely “source language salience (gravitational pull), target language salience (magnetism), and link strength effects (connectivity)” (Halverson, 2017, p. 15), the last referring to entrenched translation solutions.

From the perspective of conceptual structures, Halverson (2007) links the concept of translation shifts to the construal operations related to general cognitive processes. Depending on the construal operation, the same situation is conceptualized in a different way, resulting in different linguistic structures. Halverson (2007, pp. 118–119) identifies three general reasons for translation shifts, based on cognition, convention and context. Szymańska (2011) uses the notion of construction to study translation shifts: since constructions represent multiple formal and functional levels of analysis and display a certain range in size and abstractness, they are believed to be suitable for explaining such multi-layer translation concepts as equivalence and translation shifts. While it may be possible to identify corresponding constructions that share some formal and functional features (Leino, 2010, p. 131), translators may encounter “constructional resistance”, that is, “the difficulty (sometimes impossibility) of finding a TL construct representing all the functional properties of a ST construct (semantic, pragmatic, discourse, stylistic) while keeping the formal similarity as well as producing a ‘natural’ TL expression” (Szymańska, 2011, p. 127). Constructional resistance, which occurs because constructions—complex bundles of formal and functional features—are language specific, may therefore lead to translation shifts. Tabakowska (1993, 2013) views constructions as scene conceptualizations: during translation, potential differences in the conceptualization of the same situation and its linguistic realization in different languages may lead to translation shifts (Tabakowska, 1993, pp. 73–77).

In these accounts, constructions thus serve as translation units: assuming that a complete inventory of constructions is available, constructions may be used in corpus-based studies to systematically classify various translation shifts at different levels of analysis (Szymańska, 2011, p. 216). Recent corpus research on translation shifts drawing on the notion of construction, though bound by the range of identified constructions, has shown the potential of combining this theoretical framework with multivariate statistical approaches. Using the comparable and parallel parts of the CroCo corpus (Hansen-Schirra et al., 2012), Serbina (2015) investigated shifts to and from argument structure constructions on several levels of abstractness. The results indicate that one of the reasons for translation shifts may be different frequency distributions of the corresponding constructions in English and German, which, according to the corpus-to-cognition principle, reflect different levels of entrenchment in the speakers’ minds. To add the process perspective on the construction shifts, the corpus-based study was complemented by a small-scale translation experiment (see Heilmann et al., in press, for more comprehensive experimental testing).

In another corpus-based study adopting the framework of construction grammar, Kruger and van Rooy (2016) focus on reported-speech constructions in different contact varieties. The study draws on cognitive assumptions to explain the over- or under-representation of the features of the investigated constructions. For instance, the frequent presence of the complementizer that in indirect speech (see Olohan & Baker, 2000, discussed in Section 10.2.1) in translated English and second-language writing is interpreted in terms of the assumed reduction in processing effort and the potentially higher level of entrenchment of this variant in the minds of bilingual speakers of English and Afrikaans.
Kruger and De Sutter (2018) also investigate the use of the complementizer that. They focus on the predictors for the use of explicit that, namely cognition, risk avoidance and contrastive differences. The study takes into account multiple factors considered as operationalizations of cognition and risk-avoidance-related sources of explanations. Kruger and De Sutter (2018) apply a multifactorial statistical analysis to compare the effect of these factors on the presence of the complementizer in different varieties. The influence of contrastive effects is included in the research design indirectly by analysing two pairwise comparisons between the standard variety and the two language-contact varieties. While corpus-based work applying this statistical approach appears promising for assessing the relative importance of various groups of factors on the omission of that, the authors observe that to clearly separate cognition and risk-avoidance factors, it is necessary to carry out more controlled corpus-based and experimental studies (Kruger & De Sutter, 2018, pp. 279–280).

A constructional account of translations with a focus on frame semantics (see, e.g., Fillmore, 1982) is presented in Ćulo (2013). The study investigates two cases of constructional mismatch for the language pair English–German. One of these constructions is the unagentive subject construction, that is, an extreme case of non-agentive subjects (see Section 10.2.3). Ćulo (2013) reports an analysis of the CroCo corpus according to which this construction could be translated into German using another topicalization construction. This constructional shift may be accompanied by a frame-semantic shift. Building on this analysis, Ćulo (2017) proposes a frame-semantic and constructional model of translation.

The overview of research presented in this sub-section has shown that cognitive linguistics allows assumptions to be made about the cognitive reasons for various phenomena observed in translated language, thus taking into account the important role of cognitive processes during the process of translation. As discussed in the cognitive linguistic literature, corpus data—conceptualized as usage data both reflecting cognition and (re-)forming it—could be used as a link to cognition, provided the appropriate operationalization of the cognitive theoretical constructs is adopted. However, this overview has also indicated that the triangulation of corpus linguistic methods with experimental work can provide further insights into the analysed research questions.

10.2.3 Combined corpus-based and experimental studies

The third broad area in the interaction between CBTS and cognition entails complementary process- and product-based analyses of translation data (see also Hansen-Schirra, this volume). One way of combining product and process analyses is by sequentially (or cyclically) investigating the same linguistic phenomenon using a corpus linguistic and an experimental research design. Both approaches can also inform each other by using the one to generate hypotheses, which are then tested with the help of the other. This strand of research often uses corpus data to study the final products of translations, drawing on available large parallel or comparable corpora (see Section 10.1). Thus, corpus-based research is complemented by experimental data, which provides an additional process-based perspective not present in such corpora. Typically, process data is used to test hypotheses about cognition. Therefore, in this approach, corpus data is only indirectly linked to the study of cognition, either by preparing the ground for experiments that aim at testing hypotheses related to cognition, or by testing experimental evidence on a larger quantitative basis using corpora.

Vandepitte’s research group has been focusing on the analysis of non-prototypical agents, i.e. non-agentive subjects, combined with proto-agents requiring predicates. A translation experiment performed using the software Morae indicated slower translation speed for sentences
containing this type of agent (Vandepitte & Hartsuiker, 2011). In addition, a corpus study has shown that a combination of non-prototypical agents and the verb give is likely to be changed in translations from English to Dutch (Doms et al., 2016). The experimental research preceding the corpus analysis is considered to be evidence for existing restrictions on the combination of non-human agents and action verbs in Dutch—this evidence is taken as a starting point for the subsequent corpus study.

S. Hansen (2003) is an early example of a more direct triangulation of the product- and process-based perspectives. She combines a detailed analysis of comparable and parallel corpora with an experimental pilot study. The corpus analyses study the translation properties of normalization and anti-normalization (i.e. avoidance of typical features) and establish the influence of the source language on the features of the translated language. The experiment tests the process of translating the linguistic features that displayed specific patterns in the corpus analysis.

Alves et al. (2010) combine corpus and experimental research to study the phenomenon of grammatical metaphor, more specifically the process of de-metaphorization, from both product and process perspectives. The authors specifically address the advantages of triangulating corpus-based and experimental methods by noting, on the one hand, that corpora provide authentic data but make it difficult to test assumptions about cognition, while, on the other hand, experiments provide a more direct perspective on the process, making the connection between the observed behavioural data and cognitive processes somewhat more straightforward. A combined process-product analysis could, therefore, benefit from the advantages of both methods and help overcome their individual disadvantages.

In Serbina et al. (2017), analyses of experimental data on the shifts between nouns and verbs are preceded by a study of the proportions of these word classes in the CroCo Corpus. Data from a translation experiment in the English–German translation direction shows that shifts from verbs to nouns are most frequent: these results are in line with contrastive tendencies observed in the corpus analysis, suggesting that translators follow the norms of the target language. Keystroke logging analyses additionally show that these shifts are typically performed in one step without additional intermediate solutions. Moreover, cognitive effort, particularly when operationalized through the measure of fixation count, increases when a shift either to nouns or to verbs is present compared with cases where the word class is not changed.

Halverson (2017) tests the latest version of the gravitational pull hypothesis (see Section 10.2.2) in a cyclic approach drawing on frequency distributions in the English–Norwegian Parallel Corpus (ENPC) and the British National Corpus (BNC) as well as on results reported in the previous literature, corroborating these in a sentence generation task. The results of these initial analyses are then reviewed again using a combination of corpus and keystroke logging data.

An approach that further integrates product and process analyses treats translation process data collected with the help of keystroke logging as a corpus. While the studies described earlier use two different types of data to combine product- and process-based perspectives, treating keystroke logs as a corpus allows researchers to analyse the same data using different methods. The idea of creating a corpus consisting of translation process data was introduced in Pagano et al. (2004) using data stored in the Corpus on Process for the Analysis of Translations (CORPRAT), a corpus which can be manually annotated and semi-automatically queried using the software LITTERAE, designed to analyse translation units stored as translation process data (Alves & Vale, 2009, 2011).

A recent study by Kajzer-Wietrzny et al. (2016) also uses data collected in an experiment for analysing the source and the final target texts from a process perspective by focusing on different phases of translation. However, it is of less interest in the present context, as the study does not have a strong focus on cognition.
The Keystroke Logged Translation Corpus (KLTC), introduced in Serbina et al. (2015), is also based on data collected during translation experiments. This corpus can be used to analyse intermediate versions of individual words and sentences, defined as “variants of the unfolding texts produced at certain points in time during the translation process” (Serbina et al., 2015, p. 102). Further pushing the boundaries between corpus-based and experimental methods, this corpus is enriched with part-of-speech annotation of these intermediate versions even if they do not make it to the final translation product. Similarly to the corpora described in Section 10.1, querying for this type of annotation in a corpus consisting of the keystroke log files (Alves & Vale, 2009, 2011) allows researchers to generalize beyond individual observations. Appropriate querying strategies for this new type of corpus still require further development.

10.3 Recent developments and future directions

Until now, the large number of individual papers in CBTS somehow does not seem to have brought a major breakthrough in terms of fully illustrating the empirical facts of translation. To some extent this is due to the inconclusiveness of the results, which, in turn, can be partly explained by limitations of some of the studies. So, one might come to the conclusion that perhaps the approach has not lived up to its promise after all, were it not for the machine learning findings mentioned in Section 10.2.1, which suggest that there must be something about translations that makes them so easy to spot.

So, possibly, the methodological approach needs to be refined. Often, studies investigate individual linguistic features in one language pair and usually in a fairly limited data set in one genre/register. However, translations are pushed into different directions by different features, and different contexts will involve different expectations as to the extent to which translations are allowed to be identifiable as such (House, 1997). Moreover, excluding the (aligned) source texts will make it utterly impossible to decide whether a feature of the target text is specific to translation without checking whether the feature was present in the source text (see Steiner, 2012). Also, it may be necessary to use more strictly quantitative, explanatory designs, as suggested by Tummers et al. (2005).

In recent years, a number of CBTS studies have worked towards overcoming these limitations. A number of corpus studies have employed more complex designs that include different registers, different translation directions and sometimes also different features, using advanced statistical techniques, and thus yielding a more differentiated picture of the multifactorial make-up of translations (e.g. Delaere, 2015; Evert & Neumann, 2017; Kunz et al., 2017; Serbina, 2015; Vandevoorde, forthcoming, etc.).

An important recent development in CBTS that calls into question the conventional conceptualization of the translation process in empirical translation research is concerned with a closer investigation of the influence of editing/proofreading on the translation workflow. Bisiada (e.g. 2013) reports findings from a corpus of translations submitted to a journal and the final versions of the translations after revision by the journal editor. It turns out that many of the phenomena described by CBTS scholars as translation properties appear to be introduced by the editor rather than by translators (for indirect evidence in newspapers see also Delaere, 2015). These findings are further strengthened by Kruger (2017), who reports statistically significant linguistic differences in a corpus of originals before and after editing, which suggest that editors introduce the features that have also been described as properties of translation, and plausibly conjectures that edited translations should show similar effects. Although these approaches are, strictly speaking, not concerned with cognitive aspects of translation, they are of great relevance to cognitive approaches to translation because they relativize the role of the translator’s cognition.
in relation to the published translation, in which those aspects that might require the highest cognitive demand on the part of the translator are not present because they are changed during editing.

A potential development towards the integration of process- and product-based translation research could emerge from the CRITT TPR database (Carl et al., 2016). At this stage, this is a database containing translation and post-editing experiments. However, the way it is structured and analysed bears some striking similarities to corpus linguistic methods. While at this stage it is impossible to search the database in the same way a corpus can be searched, it contains the relevant information and could be exploited in a similar way.

It seems appropriate to conclude that CBTS has become mature in recent years: there is a trend towards more complex research designs involving statistical analysis. In line with a general trend in corpus linguistics, this also involves a tendency to take into consideration cognitive explanations. Halverson (2015, p. 311) even claims that “the adoption of some cognitively oriented explanatory models is driving the integration of process and product perspectives on translational phenomena”. Also, the analysis of process data as a corpus will probably see some additional expansion.

Depending on the details of the research design, it is also possible that a very large corpus with rich metadata can be used to discover behavioural patterns that might be interpretable with respect to cognitive explanations. Generally speaking, the use of larger, but still balanced, corpora should level out idiosyncratic aspects of the individual translators and the specific situations in which they produced their translations, so that patterns of translators’ behaviour become visible. Only such more complex designs—ideally with more meta-information on the details of the production context—can reduce the gap between high-level theoretical assumptions and limits of the empirical research design.

Hansen-Schirra (2017, p. 238) discusses links between psychological mechanisms and properties of translation: she suggests that priming (of the source language) could lead to shining through, whereas monitoring or inhibition (of the source language) could lead to normalization. This appears to be a very promising avenue for future research, as it could provide an explanation for the mechanisms that lead to these features. However, as Evert and Neumann (2017) show, the linguistic characteristics of translations are not exhausted by cognitive explanations.

Notes
1 Although Toury never carried out any corpus analyses, he discussed quantitative analyses and specifically the role of probabilities for understanding specific characteristics of translations far beyond what was typically done in Translation Studies in Toury (2004).
3 Retrieved 16 March 2018 from http://corpus.leeds.ac.uk/mellange/ltc.html
4 Retrieved 3 April 2018 from www.lancaster.ac.uk/fass/projects/corpus/ZCTC/
5 Olohan and Baker indeed use the term “subconscious” despite its psychoanalytical connotations. In what follows, we will use the more neutral term “unconscious”.

Further reading

Using the concept of “literal translation”, Halverson demonstrates the combination of product- and process-based methods for theorizing translation grounded in cognitive linguistics.


This publication provides a recent overview of both empirical translation and interpreting studies. The papers apply a broad range of methodologies, often combining corpus-based analyses with process-based methods including eye-tracking, keystroke logging and functional MRI experiments as well as surveys.


The papers in this volume apply state-of-the-art quantitative methods, some of which are frequently used in cognitively oriented corpus linguistics, to translation corpora.


This volume contains a range of papers taking a cognitive linguistic perspective on Translation Studies. It includes both theoretical and empirical work in the area.

**References**


Neumann and Serbina


Neumann and Serbina


