

TRANSLATION CORPORA IN CONTRASTIVE RESEARCH, TRANSLATION AND LANGUAGE LEARNING

*Josef Schmied**

ABSTRACT: This article looks at the role of translation corpora in all fields of comparative language studies. Over the last decade, corpus linguistics has expanded into a new, powerful and easily accessible methodology, which has brought new impulses to many older sub-disciplines of linguistics. Thus, translation corpora have revitalised all comparative or cross-language studies, since they can be used profitably in contrastive linguistics and translation studies as well as in language teaching and learning. Translation corpus studies are particularly popular among non-native speakers of English, as they offer a sound basis for language analysis that does not depend on introspection. Because they combine a qualitative and a quantitative perspective, they are particularly interesting for gradient phenomena, like the auxiliary-catenative-full verb cline that provides the empirical test field in much of this contribution. The opportunities offered by translation corpora are illustrated using examples from the Chemnitz English-German translation corpus, mainly in three case studies of auxiliary *help*, catenative *appear/seem*, and modal *may/might*.

KEYWORDS: translation corpora, contrastive grammar, language learning, auxiliaries, modality, catenatives.

RESUMO: O presente artigo analisa o papel de corpora de tradução em todos os âmbitos da pesquisa lingüística contrastiva. Na última década, a Lingüística de Corpus se

* Chemnitz University of Technology, Germany.

transformou numa metodologia nova, poderosa e de fácil acesso, que impulsionou a maioria das disciplinas tradicionais da lingüística. Assim, os corpora de tradução revitalizaram todas as áreas de estudo comparativas, já que podem ser usados, com proveito tanto nos estudos lingüísticos contrastivos quanto na tradução e no ensino e aprendizado de língua. Estudos baseados em corpora de tradução são bastante difundidos entre os falantes não-nativos de inglês, uma vez que constituem uma fonte confiável para a análise da língua, que não depende da introspecção. Por agruparem uma perspectiva tanto quantitativa quanto qualitativa, são particularmente interessantes no caso de fenômenos que indicam gradação, como a escala verbo auxiliar-de ligação-pleno, que fornece o campo de teste empírico para a maior parte deste trabalho. As possibilidades oferecidas por corpora de tradução são ilustradas com o uso de exemplos extraídos do Corpus de Tradução Inglês-Alemão Chemnitz, que focalizam em especial três estudos de caso: o uso de help como auxiliar, os verbos de ligação appear/seem e o modal may/might.

UNITERMOS: *corpora de tradução; gramática contrastiva; aprendizado de língua; verbos auxiliares, modais e de ligação.*

1. Recent developments in translation corpora

1.1 Definitions

Similar to most introductions to corpus linguistics (e.g. Biber/Conrad/Rippen 1998, Kennedy 1998, McEnery/Wilson² 2001 or Meyer 2002) we define a corpus, in traditional terms, as a text collection that is the basis for linguistic analysis and, in more restricted terms nowadays, as computer-readable (i.e. in electronic form) and maximally representative. The latter criterion is obviously most problematic, since it suggests ideals about sampling and stratification that are often difficult to meet (cf.

TRADTERM, 10, 2004, p. 83-115

Schmied, 1990 or Biber, 1993). Thus, an ideal translation corpus would have to contain several text types that are typically translated,¹ and possibly even spoken language in authentic texts from interpreted discourse. However, that is rarely the case.

After monolingual corpora had been applied successfully to descriptive language research (for the state of the art in grammar, cf. Biber et al., 1999 and in lexicography, see Ooi, 1998), multilingual corpora were compiled in the 1990s (cf. Aijmer/Altenberg/Johansson eds., 1996) and have so far yielded some interesting results. In some of these text collections there are direct translations in usually two (rarely more) languages; occasionally there are also comparable authentic, original texts juxtaposed that serve the same function in the source and target cultures or belong to the same text type and register, but are not really equivalent in semantic details. This setup creates an interesting system of related texts, source texts (ST) and target texts (TT) in L1 (like English) and L2 (like German), that can be used for various types of analyses (Fig. 1). Despite all efforts, the however slight lack of semantic compatibility makes a comparison of source texts (STE vs. STG) difficult for use in contrastive linguistics. The establishment of language-specific natural patterns, however, is facilitated, providing a good basis for the establishment of target-language norms against which translations could be measured. The source and the target texts are, hopefully,²

¹ The idea put forward in a discussion on the linguist list that a translation corpus could contain only translations (of an original that is not included in the corpus) seems rather unusual. Finding appropriate translations and obtaining copyright for using both versions is usually not easy. It should also be pointed out that the translation industry is very one-sided, i.e. that translations are very text-type specific and occur much more often from English into German than from German into English, according to the Index Translationum, published by UNESCO.

² The discussion of equivalence or the *tertium comparationis*, which preoccupied translation specialists in the 1960s and 1980s, has since fallen into discredit in favour of more adaptive, dynamic approaches (cf. Tymoczko 1998: 653), as well as approaches focusing more on actual language use than on language structure (cf. House, 2000). This means a shift of interest from the source to the target text (cf. Garcia 2002: 396).

equivalent in meaning and this allows the researcher to look for structural similarities and differences at the same time (STE vs. TTG and STG vs. TTE, the results of which should ideally be reciprocal, if they depended only on the two languages and not at all on the translation process as such). Finally, a comparison of translated (target) texts (TTG and TTE) can reveal instances of or even tendencies in ‘translationese’, which may consist of interference from the source language or deviations from the target-language norm (cf. Schmied/Schäffler, 1996). Of course, unconscious over- or underusage of certain structures may be counter-balanced by translators’ awareness of target-language preferences. For example, indirectness or tentativeness (as in S1E) are said (and can be proved) to be ‘typically English’ and are thus rightly omitted in some German translations.

(S1E): ... that it *may* be useful to illustrate the modernity of the vocabulary of the subject itself (Hobsbawm, 1991: 25)

Finally, it should be made perfectly clear again that all corpora have their special values: monolingual corpora (including source and parallel text) are used to investigate naturalness; monodirectional translation corpora increase the awareness for translation strategies and norms; bilingual bidirectional corpora help users to investigate translationese and language learners to analyse systematic differences.

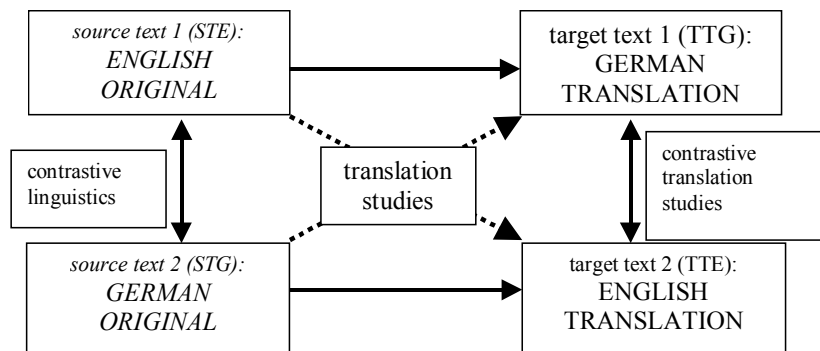


Fig. 1: Model of a multilingual (original, parallel and translation) corpus network (adapted from Johansson/Hofland, 1994:26; also cf. CEXI corpus in Bernardini, 2002: 178)

Unfortunately, the terminological distinction between parallel texts (STE vs. STG above) and translation texts (STE vs. TTG and STG vs. TTE above) is not maintained consistently in the literature on corpus linguistics and translation (cf. Fig. 2). For instance, Baker (1995: 230-5), in her discussion of the use of corpora in translation studies, uses a different classification. She makes a distinction between ‘parallel corpora’, which we would call translation corpora (STE vs. TTG and STG vs. TTE in Fig. 1 above), ‘multilingual corpora’, which would be parallel corpora in our terminology (STE vs. STG), and so-called ‘comparable corpora’, which contain authentic texts in language A as well as translated texts in the same language A (STE and TTE or STG and TTG). In our view (also in Lauridson’s, 1996 and House’s, 2000), in a bilingual corpus of translations from language A to language B and vice versa, the respective source texts form a parallel corpus (i.e. they fulfil the same functions and belong to the same text category) and the texts of language A and language B constitute respectively what Baker calls a comparable corpus. However, a comparable corpus can also be a ‘reference corpus’, which can be described as a (larger) ‘monitor corpus’ against which a purpose-built corpus is weighed as far as representativeness is concerned (cf. below).



Fig. 2: Multilingual corpora terminology

The term ‘parallel corpora’ seems attractive because it suggests alignment, but, of course, all translated corpora need to be aligned for computer-based analysis, otherwise they would not be very helpful for the type of analysis exemplified below.

Finally, the term ‘monitor corpus’ needs clarification. It has been popularised mainly by Sinclair (e.g. 1991: 24ff) as a ‘large

and up-to-date selection of current English available' (ibid. 25). Today the World Wide Web is obviously the appropriate 'monitor corpus' if we use suitable tools and precautions (cf. Schmied, 2004).

1.2 The English-German translation corpus in the Chemnitz Internet Grammar

The Chemnitz Internet Grammar (cf. Schmied, 1999 and Schmied/Haase, 2003) is a language learning tool that allows users to work on major areas of English grammar in an inductive or deductive way. The inductive way, from examples to rules, includes a search engine that allows users to extract their own example sentences from an English-German translation corpus.³ All user movements, inputs and test results are stored in logfiles, which can be analysed for indirect feedback from the users (in addition to the direct feedback via email). The first version⁴ of the Chemnitz Internet Grammar (used in Fig. 3 and Fig. 4 below) displays all the sentences that match the search criteria in English (and German, if required) and some basic statistics that illustrate the distribution across text types.⁵ This is considered

³ The text-types range from political documents 'written as if spoken' to academic writing in natural sciences and humanities, from tourist brochures to economic and social EU documents; unfortunately, literary texts could not be included for copyright reasons (cf. Schmied 1994). To date, over 1.5 million words from texts mainly from the 1990s have been included in German and English (the first date after a quotation refers to the publication of the original, the second to that of the translation).

⁴ A new version was installed in September 2003, which has a modernised interface and an expanded translation corpus and allows the user to skip from inductive to deductive learning mode within each chapter. I wish to thank the German Research Association (DFG) for their support over the last five years and my collaborators, esp. Ellen Gorlow, Christoph Haase, Naomi Hallan, Isabel Heller, Diana Hudson Eittle, Gerard Keohane, Tobias Lehnert and Katrin Voigt, for many interesting discussions.

⁵ The following tables and figures can only illustrate the possibilities in a few simple examples that emphasise the text-type-specific distribution of language forms. Further statistical operations may include relative comparisons across lexical features and all relevant significance tests.

important since many language phenomena are text-type specific (like the adverb *thus* in Fig. 3).





| | | | | | | | | | | | | |
|------------------------------|---|-----|-----|-----|-----|-----|-----|-----|-----|------|---------------------|---------------|
| 314 soc | Thus , both sides of the equation have to be looked at if we are to answer the question whether these more varied and flexible forms of working time and participation in work represent an opportunity to reconcile efficiency with both a higher quality of life and more widely spread working opportunities. | | | | | | | | | | | |
| 315 soc | Thus , there will be a need for action on two levels: a series of measures to restore growth and a parallel action to address the structural barriers to job creation. | | | | | | | | | | | |
| 316 soc | Thus , particularly in the context of a shift towards quality production processes, more flexible and even reduced working time can be an integral part of the process of change. | | | | | | | | | | | |
| 317 soc | Thus , there was a solid base on wich to develop the legal and practical arrangements to facilitate the movement of people. | | | | | | | | | | | |
| 318 scot1 | Thus , there is a wide choice of adventurous mountainous landscapes to be enjoyed. | | | | | | | | | | | |
| 319 wls3 | Cardiff’s story goes back, at least, to the Romans and thus gives it 2000 years of history. | | | | | | | | | | | |
| Relative Statistics | | | | | | | | | | | Absolute Statistics | |
| Domain | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% | Percentage | Numbers |
| Public Speeches and Articles |  | | | | | | | | | | 0.15 % | 7/ 4625 |
| EU Documents |  | | | | | | | | | | 1.33 % | 94/ 7066 |
| Tourist Brochures |  | | | | | | | | | | 0.04 % | 2/ 4368 |
| Academic Texts |  | | | | | | | | | | 1.76 % | 216/ 12266 |

Fig. 3: Search results for *thus* in the Chemnitz Internet Grammar (screenshot of last few examples and statistics)

The graphic visualisation displays relative (in colour bars) and absolute statistics. Thus, the English connector *thus* is clearly a formal phenomenon, since its occurrences can be found almost exclusively in EU documents and academic texts. But this is not especially surprising since they constitute the larger proportions of the corpus. This has to be born in mind for the correct interpretation of the raw figures of occurrences in simple cross-tabulations. The contrast becomes manifest when a percentage of cases is calculated on the basis of corpus sentences in each text type. This percentage shows that *thus* occurs in over 1 per cent of all sentences in more formal texts, but much more rarely (below 0,2 %) in the more popular text types, political speeches and articles and tourist brochures.

A bilingual search displays a more complex picture (Fig. 4). When all the sentences where English *so* is rendered as German

so are displayed, we see that this particle has two very different functions in both languages; it is a conjunct, a logical clause connector (in example 450), and an adverb, modifying adjectives (in example 451). This may be one reason why the text-type pattern is not as clear as *thus* above.

450 york

Northern shoppers love a bargain **so** expect great value for money.

Die Kundschaft im Norden Englands erfreut sich immer eines preiswerten Einkaufs. **So** dürfen auch Sie das gleiche erwarten.

451 york

Wildflower meadows, drystone walls, sweeping valleys and fast flowing rivers provide a breathtaking backdrop to the stone villages **so** characteristic of the Yorkshire Dales.

Blumenwiesen, Bruchsteinmauern, geschwungene Täler und schnelle Flüsse bilden den atemberaubenden Hintergrund für die Natursteinhäuser, die **so** charakteristisch für die Yorkshire Dales sind.





| Relative Statistics | | | | | | | | | | | Absolute Statistics | |
|------------------------------|---|-----|-----|-----|-----|-----|-----|-----|-----|------|---------------------|-----------|
| Domain | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% | Percentage | Numbers |
| Public Speeches and Articles |  | | | | | | | | | | 1.05 % | 49/4625 |
| EU Documents |  | | | | | | | | | | 0.50 % | 36/7066 |
| Tourist Brochures |  | | | | | | | | | | 0.68 % | 30/4368 |
| Academic Texts |  | | | | | | | | | | 2.73 % | 336/12266 |

Fig. 4: Search results for English *so* rendered as German *so* in the Chemnitz Internet Grammar (screenshot of last few examples and statistics)

Of course, the effective use of a translation corpus depends on its query system, which must be intuitively easy and consistent. However, it also depends on the search skills of its user, for searching a translation corpus requires some knowledge about language variation and typology, since capturing the complete morphological variation (esp. of the German equivalents with their complex inflectional endings) can be quite demanding. The simple search program allows for the Boolean operators *or* and *not* and the usual asterisk * wildcard for letters as well as (parts of) words. Thus the Standard English morphological variation with 3rd person singular present tense *-s* and past tense/past participle *-ed* and present participle *-ing* have to be taken into account. The use of wildcards (like *help**) may prevent undercollecting (forgetting *helping*), but it may also favour overcollecting (including the

plural noun *helpings*). The issue is made much more complicated by German ablaut forms (as in *hilft*, *half*, *geholfen*) and conscious and unconscious omissions in translation in general.

Despite these caveats, a translation corpus with its integrated search options provides a sound basis for certain types of empirical linguistic research.

1.3 The importance of translation corpora as a discovery procedure

The advantage of corpora as databases is that they allow the comparative researcher to access and analyse vast quantities of authentic natural data in a very short time. It is not only much faster, but also much more consistent than other techniques such as introspection and the ‘manual’ analysis of parallel texts (see above). Corpus studies are particularly popular among European and now also Japanese non-native speakers of English, since they offer a sound non-introspective basis for language analysis. If intuition cannot be relied on due to lack of continuous exposure, a key-word-in-context view of hundreds of examples makes patterns ‘visible’, at least for the experienced analyst (but cf. 4.1 below).

Since corpora combine a qualitative and a quantitative perspective, they are particularly interesting for gradient phenomena, like the auxiliary-catenative-full verb cline. Although natural language patterns (and their grammaticality) are difficult to objectivise, they include at least two aspects: frequency and acceptability. On the one hand, frequency allows at least tentative hypotheses of patterns, although acceptability, the second major criterion for ‘rules’, does not always follow suit. On the other hand, infrequent patterns or other special cases may make the borderlines of phenomena manifest, at least more easily than introspection by non-native speakers (cf. *help* in the sense of ‘contribute’ in negative cotexts in S2 below).

(S2) the failure of one such effort **helped** bring down Nikita
Khrushchev deren Scheitern z.B. zum Sturz von Nikita
Chruschtschow **beigetragen** hat (Crosby, 1986/91)

TRADTERM, 10, 2004, p. 83-115

Translation corpora offer an additional advantage in that a cross-linguistic comparison can be used as a discovery procedure, since semantically equivalent phenomena can be chosen by translators who do not want to render a phenomenon in the formal equivalent in the target-language. It can, for instance, be shown (Schmied, 1998) that the prototypical equivalent to the English preposition *with*, German *mit*, was chosen by translators in just half of the cases, since many more structurally different equivalents were available. Thus a quantitative corpus approach allows us to pursue unifying and diversifying strategies, i.e. we can go for the prototypical solution as well as the non-prototypical solution. For example, we can see that the standard translation of German *beitragen* is *contribute*, but a structurally different ‘typically English’ construction would use *help* as an auxiliary (cf. 2.1 below).

The great asset of translation corpora is that their analysis can be reversed. Therefore, the overlap between *help* and *helfen* can be approached from both sides, by searching for forms of *help* that are rendered by a form of *helfen* and by forms of *helfen* that are rendered by a form of *help*. This approach does not only assist in discovering translation errors, but also to measure the overlap between closely related languages like German and English.⁶

Unfortunately, current introductions to corpus linguistics, from Kennedy (1998) to Meyer (2002), do not appreciate this new opportunity. Kennedy hardly mentions translation corpora, although he has some related project descriptions (e.g. 1998:42). Meyer (2002:22-24) raises expectations in a chapter on ‘Contrastive analysis and translation theory’, but does not go beyond two illustrative project descriptions either. McEnery/Wilson (2001) deal neither with corpora and translation studies nor with corpora and language learning (but they mention applications in language teaching and give examples from the teaching of linguistics).

⁶ The resulting lexical web can also be used as a new basis for a thesaurus, which could complement the traditional deductive attempts made since Roget, since they are developed inductively on an empirical basis.

Today the WWW lists several websites devoted to 'Parallel translation corpora around the world'. Comparative studies based on translation corpora are carried out for instance for English and German (at Chemnitz and Dublin, e.g. Schmied, 1994 and Kenny, 1999), Portuguese (at Oslo/Porto/Lisboa, e.g. Maia, 2000 and Frankenberg-Garcia, 2002), Spanish (at Leon), Italian (at Bologna, e.g. Zanettin, 2002 or Bernardini, 2000 and 2002), Norwegian (and German, at Oslo, cf. Johansson, 2002 and Johansson, this volume), Swedish (at Lund, e.g. Johansson, 1996), Finnish (at Jyväskylä) and Polish (at Lodz), but also for other language pairs.⁷

2. Contrastive research

2.1. The revival of contrastive studies through translation corpora

The new corpus-linguistic approach to contrastive linguistics was proclaimed by Sajavaara (1996). He also recorded the failure of the earlier structuralist contrastive linguistics and called for a socio-psycholinguistic basis, in other words, for a cognitive, process-oriented approach that does not ignore the setting (*ibid.*: 21). With translation corpora, specialists in contrastive linguistics and the related new approaches to typology not only have access to a vast array of examples (and do not have to invent or collect them impressionistically anymore), but can also access gradable phenomena or language-specific preferences. This is particularly useful for closely related languages like English and German, with almost identical structural inventories but preferences that may differ typically. The 'related'

⁷ Although many case studies are available, few translation corpora are distributed freely or can be consulted via the WWW. This is mainly due to copyright restrictions. The English-German translation corpus described here is available as part of our InternetGrammar at www.tu-chemnitz.de/InternetGrammar; the same applies to the Portuguese corpus COMPARA at www.linguateca.pt.

international conferences (e.g. Rábade/Suárez eds., 2002) bear witness to the expansion of translation-corpus applicability in the field.

2.2. Analysing language-specific phenomena: auxiliary ‘help’

Despite the many structural features shared by English and German due to their common West Germanic origin, the auxiliary-catenative-full verb cline has expanded considerably over the last few hundred years and obviously continues to do so at present, as in the case of *help*. *Help* is a catenative, i.e. a verb that takes non-tensed clauses as complements (cf. Huddleston/Pullum, 2002, esp. 1194-1245). Syntactically, there are 4 types of catenatives (Table 1). Type 1 catenatives are far more common than type 2. In fact, *get* (in passive meanings as in *it got repaired*) is the only case of type B2, and *help* and *dare* (apart from a few idioms like *make do* or *let go*) are the only two verbs without an object, occurring more often with than without infinitive anyway.⁸

| Catenative type | type 1 | Type 2 |
|-----------------|---------------------------------|---------------------------|
| A) infinitive | + to : <i>hope</i> | – to : <i>help</i> |
| B) participle | + -ing : <i>remember</i> | + -ed : <i>get</i> |

Table 1: Types of English catenatives according to complements

As *help* and *dare* are followed by a pure infinitive like auxiliaries, they can be assigned to two very different types of syntactic structures. Auxiliaries are considered verbal modifiers, thus the main verb comes after them. But type 1 catenatives are considered main verbs and the following full verb with or without infinitive is considered a complement. This makes ‘dropping the

⁸ Since our corpus is too small for this analysis, the WWW had to be consulted. The proportion in the domain .uk is about 5 to 1 (cf. Schmied, 2004).

infinitive particle *to* after *help*’ a major syntactic change, which can be interpreted as grammaticalisation.⁹ In contrast with *dare*, which is said to move away from auxiliary status, *help* seems to be moving towards it. As in other cases of grammaticalisation, this syntactic change is accompanied by a semantic change away from the traditional meaning ‘support’ towards the new meaning ‘contribute’.

A corpus-linguistic analysis can contribute to this discussion in many ways:

- 1) A bilingual search for the forms of English *help* (*help/helps/helped*) shows (Table 2) that they occur very often when they are NOT rendered in German by any of the forms of *helfen* (*hilft/hilst/hilfe/helfen/helft/half/halfst/halfen/half/geholfen*).¹⁰ This surprising result can be easily verified by calculating the reverse relationship, i.e. the proportion of German *helfen* rendered in English by *help*, which shows the same overlap **and** that the total number of *helfen* is only 12 more, namely 57. Thus the relationship between the two forms is not reciprocal: English *help* covers the same functions as German *helfen* – and many more (75 %).
- 2) The text-type comparison reveals that the non-equivalence occurs most often in EU documents: in 62 out of 66 cases altogether, i.e. in 94 % of the occurrences, the stereotypical equivalent is not used.
- 3) The unexpectedly low figure for the correspondence of related terms illustrates the wide variation that occurs in freely translated text passages.

⁹ In the historical development of English over the last 500 years, the distinction between auxiliaries and full verbs was made structurally obligatory, i.e. verb forms that take only bare infinitives as complements are central modal auxiliaries; thus *dare* is considered a borderline case since it can be used with or without *to*.

¹⁰ In the English search we cannot exclude (without part-of-speech tagging) the noun *help*, which can be excluded in the German search (*Hilfe*). Thus the lexical overlap (including nouns and adjectives; cf. S5 below) is actually a little larger than suggested by the figures in Table 2.

| Text Type | English <i>help</i> | | German <i>helfen</i> | NOT German <i>helfen</i> |
|------------------------------|---------------------|---------|----------------------|--------------------------|
| EU Documents | 66 | 28.82 % | 4 | 62 |
| Academic Texts | 60 | 26.20 % | 10 | 50 |
| Public Speeches and Articles | 73 | 31.88 % | 21 | 52 |
| Tourist Brochures | 30 | 13.10 % | 10 | 20 |
| Total | 229 | 100 % | 45 | 186 |

Table 2: The English forms of *help* (not) translated as *helfen* in our translation corpus

Of course, translators legitimately try to vary their language and may choose verbal synonyms (like *unterstützen* ‘support’ and ‘helps’ as *hilfreich* in S3) or other constructions with related adjectives (like *hilfreich* ‘helpful’ in S4) and nouns (*Hilfe*/‘help’ in S5). Some of the cases also had to be covered by German *beitragen/contribute* (in S6), which renders the new, expanded meaning of *help* that correlates with the new syntactic pattern mentioned above.

(S3) A wider and more comparable information system supports comparative analysis and **helps** in determining the direction of future developments.

Ein umfassenderes und vergleichbareres Informationssystem unterstützt eine vergleichende Analyse und ist **hilfreich** bei der Bestimmung der Ausrichtung zukünftiger Entwicklungen. (EU 1991)

(S4) This should provide a useful framework and, at the same time, **help** point out the direction we’ll be going. Dies soll ebenso einen sinnvollen Rahmen abstecken wie als **hilfreiche** Orientierung dienen, welche Wegerichtung wir einschlagen werden. (John Murphy 1991/92)

(S5) Perhaps the best way of interpreting Scotland’s wild places is with the **help** of the Countryside Ranger Service. Die beste Art, Schottlands unberührte Natur zu erkunden, ist mit **Hilfe** des Countryside Ranger Service. (Scottish Tourist Board, 1993)

(S6) Such economic growth can **help** cut off the oxygen of terrorism.

Ein derartiges Wirtschaftswachstum kann dazu **beitragen**, dem Terrorismus den Boden zu entziehen. (Speech John Major, 1994)

The reverse analysis shows that, except for a few cases in EU documents and political speeches/articles, the vast majority of German ‘beitragen’ is not (yet) rendered as ‘help’. The prototypical ‘contribute’ is much more common in general, probably because German *helfen* is more colloquial than English *help*.

| Text Type | German <i>beitragen</i> | | NOT English <i>help</i> | <i>contribute</i> |
|------------------------------|-------------------------|---------|-------------------------|-------------------|
| EU Documents | 54 | 62.79 % | 42 | 24 |
| Academic Texts | 11 | 12.79 % | 10 | 7 |
| Public Speeches and Articles | 21 | 24.42 % | 12 | 6 |
| Total | 86 | 100 % | 64 | 37 |

Table 3: The German forms of *beitragen* a) **not** translated as *help* and b) translated as *contribute** in our translation corpus

There are also many cases where semantic elements are combined in completely new lexemes, as in S7, where *helped to assure that ... would not* is rendered as *verhinderten* (i.e. ‘prevented’) and the German translation adds the much more explicit ‘and thus to compensate the wood cutting of thousands of trees’ to the English original.

- (S7) They **helped** to assure that seedlings would not grow into trees to replace the thousands cut down in answer to European needs in the islands and elsewhere.
 Sie verhinderten, daß nachsprießende Sämlinge sich zu Bäumen auswachsen und damit den Holzeinschlag ausgleichen konnten, dem Tausende von Bäumen zum Opfer gefallen waren, um den Holzbedarf der Europäer auf den Inseln und anderswo zu befriedigen. (Crosby, 1986/91)

Similar contrastive cases can be found in many constructions. For many syntactic comparisons, however, a tagged corpus is necessary, esp. when it comes to determining whether German is really ‘giving us a “tighter fit” between surface form and semantic representation’ (Hawkins 1986:122). The heated – more theoretical than empirical – debate on this issue could be complemented by quantitative analyses of raising constructions,

WH-extraction, pied piping and NP deletions of a stratified English–German translation corpus.

Multilingual corpora, and particularly translation corpora, are, therefore, definitely a way forward in contrastive linguistics, although more corpora need to be compiled, more (syntactic and semantic) tags need to be inserted and many more detailed studies have to be carried out before we can come to a more comprehensive statement about the contrasts between German and English, for instance.

3. Translation studies

Over the last decade, Baker (1993, 1995, 1999 and 2002) has contributed to the scholarly discussions on the relationship between corpus linguistics and translation studies. Since then, the number of corpus-linguistic approaches to translation has exploded, as can be seen in the related international conferences on the subject (cf. Maia/Haller/Ulrych eds., 2002). Although translation corpora can be used for more theoretical or more practical types of translation studies, the following examples will concentrate on the latter.

The main advantage of a translation corpus for the translator is that it is much more varied – and much more demanding than a dictionary or a translation memory system; the following examples are intended to illustrate that.

3.1 The translation corpus as a dictionary

Like parallel texts, translation corpora can be used for finding translation equivalents. The quantitative dimension is obviously the great advantage here, because if the word is frequent enough across text types it can be taken as a prototypical translation. However, if the frequency is low, the native speaker's intuition is still needed to verify the query results. This is illustrated by the translations of *overabundant*, which are not prototypical (in S8) or not natural (in S9).

TRADTERM, 10, 2004, p. 83-115

(S8) The humans recruited two other species to assist in the slaughter: horses, which they rode, and dogs (greyhounds), which helped in locating and running down the **overabundant** species.

Für diese Schlachtung nahmen die Menschen die Dienste zweier anderer biologischer Arten in Anspruch: der Pferde, die sie als Reittiere, und der Hunde (greyhounds), die sie dazu benutzten, die **überzählige** Spezies aufzuspüren und zu Tode zu hetzen. (Crosby 1986/91)

(S9) The objective prism spectra of Sk-69 202 thus led to the same conclusion as the construction of synthetic spectra of SN1987A by Lucy (1987); his conclusion was that SN 1987A was somewhat **overabundant** in helium. Die Objektivprismenspektren von Sk -69 202 führen uns also zu der gleichen Schlußfolgerung wie Lucys (1987) Berechnung synthetischer Spektren der SN 1987A; seine Folgerung war, daß SN 1987A etwas helium**überhäufig** war. (Paul Murdin, 1990/1991)

Different translation equivalents can also be used for disambiguating meanings or for finding standard collocations. Thus the German *Haushalt* is obviously mainly a political term in our corpus and it clearly has two equivalents in English: *household* and *budget(ary)*, the latter again exclusively in political texts (Table 4). Thus the text type is almost enough for disambiguating the two meanings, which can also be rendered as *home* and *fiscal* in English, respectively.

| Text Type | German <i>Haushalt</i> | | English <i>household</i> | English <i>budget</i> |
|------------------------------|------------------------|----|--------------------------|-----------------------|
| EU Documents | 80.41 % | 78 | 8 | 50 |
| Academic Texts | 4.12 % | 4 | 4 | |
| Public Speeches and Articles | 12.37 % | 12 | 1 | 9 |
| Tourist Brochures | 3.09 % | 3 | 2 | |
| Total | 100 % | 97 | 15 | 59 |

Table 4: Occurrences of 'Haushalt' in German and its two main equivalents in English texts

Finally, we can resume the long-running debate on the meanings of *run*. Whereas German school- and textbooks still refer to it mainly in the 'literal' contexts of *rennen*, this equivalent hardly ever occurs in the 134 cases in our corpus. *Run* is overwhelmingly the function verb as in *run into problems/trouble/difficulties, headwinds, barriers; run the risk* or part of idiomatic expressions such as *in the long/short run; run down (houses), run out of (food), run over (by a bus)*. The most common German equivalents are *(ver)laufen* and *(durch)föhren*, as in *(railway) lines **run** through idyllic green countryside and family-**run** guest houses*. *Run* is the typical example of a lexeme with little etymological core meaning, where the specific meaning and thus the translation is clearly determined by the co(n)text, a good case for using a translation corpus as a dictionary – if it is large enough.

3.2 Increasing diversity and adding variation: appear and seem

The issue of lexical repetition is judged differently in European languages and styles. How much variation is considered necessary obviously depends on individual word meanings (apart from personal choice). Generally, in English, structural parallels may be considered helpful, in German variation is preferred. Such preferences have to be extracted from a parallel corpus, and a conscious adaptation contributes to the naturalness of the translated texts.¹¹

The English catenative verbs *appear* and *seem* are considered equivalent to the German *(er)scheinen*; but the overlap seems smaller than assumed. In both languages, the choice is much broader, and in German it covers many adverbs like *scheinbar, wahrscheinlich, wohl, offenbar* and *offensichtlich* (S10) and other occasional idiomatic expressions (S11):

¹¹ This is based on the assumption that the aim of the translation is a completely natural translation which is fully integrated into the target culture. Thus, it would be a 'covert translation' in terms of House (1997).

- (S10) Porto Santo was too dry for sugarcane, but Madeira **seemed** ideal, and in all likelihood sugarcane was growing there before the middle of the fifteenth century. Porto Santo war für den Zuckerrohranbau zu trocken. Madeira hatte **offensichtlich** das ideale Klima: Höchstwahrscheinlich wuchs das Zuckerrohr hier bereits vor der Mitte des 15. Jahrhunderts. (A.W. Crosby 1986/91).
- (S11) The Mystery Spot remains largely a mystery; in fact without further clues it **seems likely** always to remain so. Der geheimnisvolle Fleck bleibt weiter ein Geheimnis, und wenn wir nicht zusätzliche, ihn betreffende Beobachtungen erhalten, **kann es gut möglich sein**, daß er immer ein Geheimnis bleiben wird. (Paul Murdin 1990/1991)

The collocation *seems + likely/probable* (S12) is considered redundant in German and occurs only once out of eleven possible cases. In English, it is quite common and it can be interpreted as a kind of modal harmony (cf. Huddleston/Pullum, 2002: 179f).

- (S12) This Old World insect may have lived in the islands before the coming of the Europeans, but it **seems** more **likely** that the invaders brought hives of bees from Iberia. Dieses der Alten Welt entstammende Insekt könnte auf den Inseln schon vor Ankunft der Europäer heimisch gewesen sein, weit **wahrscheinlicher** ist aber, daß die Eindringlinge ihre Bienenstöcke von der Iberischen Halbinsel mitbrachten. (A.W. Crosby 1986/91).

Thus *appear/seem* form an obviously complex web (or semantic field) with verbs like *(er)scheinen*, adverbs and adjectives such as *offensichtlich/offenbar*, *wahrscheinlich* or *vielleicht* and nouns such as *Wahrscheinlichkeit*. Fig. 5 shows that in many cases German adverbs (*anscheinend*; *scheinbar*; and particularly *offensichtlich/offenbar*) are chosen instead of the prototypical verbs *scheinen* and *erscheinen*.

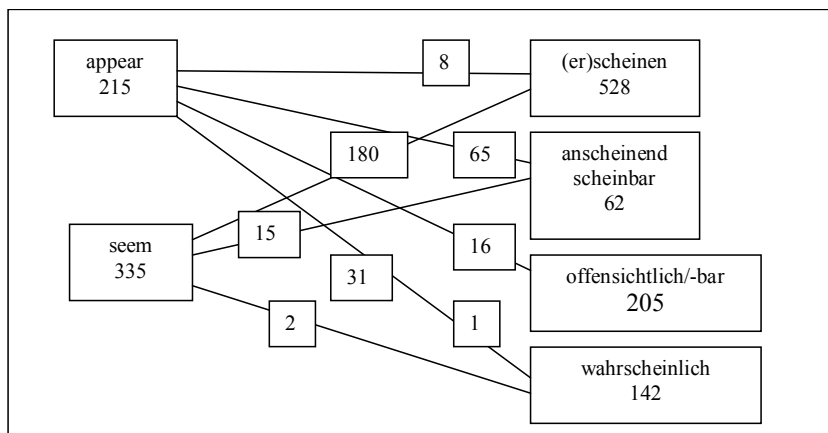


Fig. 5: The network of *appear/seem* and their German equivalents

Occurrences are indicated in the boxes on the lines connecting two corresponding lexemes, e.g. *seem* corresponds 180 times to a form of the verb *scheinen* or *erscheinen*; *appear* only 8 times. There are many other equivalents of *(er)scheinen* not displayed in the diagram.

4. Language learning

The present dissatisfaction with language teaching has grown out of the recognition that its results have often left much to be desired. The combination of rules to be learnt (by heart), invented illustrative examples and insufficient contextualisation lead to serious problems when students try to apply their 'knowledge'. A corpus-linguistic approach provides an alternative because it is based on authentic natural language in co(n)text, since concordances can show language patterns as well as variation according to collocation and colligation, text-types and contexts.

Of course, developing the *Sprachgefühl* ('feel for the language') is crucial for advanced language learners, and intuitive hypotheses are best set against hard evidence of actual language use, as documented in modern corpora.

TRADTERM, 10, 2004, p. 83-115

4.1 Data-driven language learning

Through Tim Johns' publications (e.g. 1986 and 1993) and his well-known homepage (<http://web.bham.ac.uk/johnstf/timconc.htm>), data-driven learning (DDL) has become almost synonymous with corpus-based inductive learning. Exposing the learner to automatically generated usage examples from translation corpora can make learners aware of natural language patterns, so that they develop their own *Sprachgefühl*, an intuitive feel for 'rules', which do not have to be formulated explicitly. The possibilities of querying a translation corpus independently make DDL a special form of autonomous learning, although the teacher can, and probably has to, support the learners at least initially, by providing the data in a palatable form, so that the students learn to be able to see 'something'. The possible array of sentences and tasks allows for many learner-centred activities. Later on, the options available for students to create their own concordances and statistical tables make DDL an advanced version of explorative learning. Thus DDL combines most of the latest keywords in language teaching and learning.¹² This has been pointed out by scholarly contributions on language teaching, teacher education and language learning (e.g. Mukherjee, 2002).

Ideally, students can develop their own queries, submit them to the appropriate translation corpus, create suitable statistical tables and draw conclusions on language patterns in general as well as infer correct answers to immediate language problems, e.g. in tests. For many students, however, DDL is not as simple as often assumed. How to formulate search queries on the basis of pattern hypotheses, how to verify or falsify hypotheses and how to read statistical tables are by no means clear to many language learners, and a lot more empirical research is needed to understand the psycholinguistic and cognitive basis of DDL (cf. Schmied/Haase, 2003).

¹² The usual advice 'identify – classify – generalise' seems to work only for the well-presented standard examples and the proposal to experiment with subcorpora (Aston, 2002) may not always be feasible.

4.2 Avoiding false friends (and discovering translation weaknesses)

In the early stages of language learning, so-called false friends constitute a learning problem, i.e. words in the target language that look or sound similar to words in the mother tongue but do not have exactly the same meanings (cf. CIDE 1995: 503 and for German *ibid*: 343). CIDE is the only dictionary that warns learners systematically about such dangers. Since most of these phenomena are either due to special borrowing or a long common etymology, it is not surprising that French and German have the longest lists of false friends, i.e. where a familiar French/German word 'suggests' a wrong meaning for an English word. Infamous examples are German *Strand* ('beach') and English *strand* with the same spelling, German *Stuhl* ('chair') vs. English *stool* with the same pronunciation or *slip* ('briefs', 'pants') with both. A translation corpus usually shows clearly that they are not common equivalents. Most of these lexemes, however, do not occur often enough to make a systematic study possible; only the most common cases can illustrate the intricacies of the relationships.

A quick analysis shows that none of the 193 cases of English *actual/actually* has *aktuell* as a suitable translated equivalent; instead, a fascinating variety is used, including *tatsächlich*, *in Wirklichkeit*, *sogar* etc. However, the four cases (all in science texts) where *aktuell* does occur turn out to be rather weak translations (as in S13, where *tatsächliche* would be the better adjective), which makes the translation corpus a possible discovery method for predictable cases of interference and even false friends.

(S13) In other words the **actual** or literal 'mother tongue', i.e. the idiom children learned from illiterate mothers and spoke for everyday use, was certainly not in any sense a 'national language'.

Mit anderen Worten, die **aktuelle** oder prosaische »Muttersprache«, d.h. das Idiom, das Kinder von ungebildeten Müttern lernten und für den Alltagsbedarf sprachen, war zweifellos in keiner Hinsicht eine »Nationalsprache«. (Hobsbawm 1990)

The problem may be even more complex with ‘little words’ like particles and adverbs. The adverb *also* exists in both German and English, but a search query with the same form in both languages shows convincingly that they do not have very much in common. Table 5 reveals that their overlap is minimal, English *also* is much more frequent and the distribution of English *also* and German *also* across text types is quite different (e.g. German *also* is more academic).

| Text Type | English <i>also</i> | | German <i>also</i> | | Overlap |
|------------------------------|---------------------|--------------|--------------------|--------------|-----------|
| EU Documents | 552 | 44.84 % | 62 | 19.75 % | 3 |
| Academic Texts | 368 | 29.89 % | 213 | 67.83 % | 9 |
| Public Speeches and Articles | 106 | 8.61 % | 29 | 9.24 % | 2 |
| Tourist Brochures | 205 | 16.65 % | 10 | 3.18 % | 1 |
| <i>Total</i> | <i>1231</i> | <i>100 %</i> | <i>314</i> | <i>100 %</i> | <i>15</i> |

Table 5: The occurrence of English and German *also* as an example of non-equivalence of false friends

A closer examination of the sentences with *also* in both languages makes it obvious that there are no real correspondences; either the English *also* occurs as *auch*, *ebenfalls* etc. in the German (as in S14) or the German *also* is actually the translation of *so*, *thus* etc. in the English text (in S15). The few cases that cannot be explained in this way may again be weak translations (or cases where subtle, underlying causal meanings are made explicit through an additional *also* in German).

(S14) The livestock that provide these farmers with meat, milk, leather, and power **also** provide them with the means to raise grains and vegetables and fiber in plenty on the same plots of ground that their fathers’ fathers cultivated.

Dieselben Tiere, die den Bauern mit Fleisch, Milch, Leder und Muskelkraft versorgen, liefern ihm **also** auch das Produkt, das es möglich macht, kärglichen Ackerböden reiche Ernten an Getreide, Gemüse und Faserpflanzen abzugewinnen. (A.W. Crosby 1986/91).

(S15) **Also** like the leptons, all quarks have spin 1/2, so they are fermions.

TRADTERM, **10**, 2004, p. 83-115

Wie die Leptonen haben alle Quarks den Spin 1/2, sind
also Fermionen. (P.C.W. Davies & J. Brown. 1988/1992)

Finally we consider the most famous case, where an English word has expanded its usage out of all proportion, namely *do*. A comparison of *do/does/did/done* and *tun/tust/tut/getan/tat/taten/ mach*/gemacht* (Table 6) reveals the expected disproportion, since the cases of *do* as an auxiliary obviously outnumber those as a full verb, which is the only meaning of the German equivalents *tun* and *machen*.

| Text Type | Do | | tun/machen | |
|------------------------------|------|---------|------------|---------|
| EU Documents | 191 | 14.10 % | 7 | 7.53 % |
| Academic Texts | 814 | 60.07 % | 72 | 77.42 % |
| Public Speeches and Articles | 293 | 21.62 % | 12 | 12.90 % |
| Tourist Brochures | 57 | 4.21 % | 2 | 2.15 % |
| Total | 1355 | 100 % | 93 | 100 % |

Table 6: A comparison of English *do* forms and German *tun/machen*

4.3 Modal ‘may/might’

The final and most complex example is drawn from English modal auxiliaries. Admittedly, this is again a typologically interesting new development compared to German, which has kept the more traditional forms of expressing modal concepts, subjunctive and modal adverbs. Of course, linguistic expressions for modality cover a vast formal spectrum from auxiliaries to full verbs and from (past) tenses to adverb, adjectives and nouns (cf. Huddleston/Pullum, 2002: 173f). For language learners this is extremely important if they want to achieve a near-native level of English. The simple juxtaposition of the translation options of a single auxiliary (as done with *might* below) may therefore be an important means of raising awareness about the subtleties involved.

Of course, one modal has to be seen as a member of the whole modal system. *May/might* are often seen in contrast with *can/could*, which also cover the wide spectrum from deontic to

epistemic modality. 'Prototypically, epistemic modality concerns the speaker's attitude to the factuality of past or present time situations while deontic modality concerns the speaker's attitude to the actualisation of future situations' (ibid.:178). In the case of *might*, this means that it can be either a weak prediction or a permission. Although deontic usages are usually (historically) seen as the core, and epistemic usages as an extension into another domain of human interaction, the latter are more frequent today, at least as far as *might* is concerned. In English the usages of *may/might* and *can/could* are rather similar. The main reason for the higher frequency of *can/could* in our corpus (Table 7) is that this pair of verbs is more often used in deontic and dynamic meanings (ibid.). Table 7 also shows that *can/could* is more informal, since it occurs relatively more often in the less formal text-types, viz. tourist brochures and public speeches/documents, but occurring also in EU documents.

| Text Type | <i>may/might</i> | | <i>can/could</i> | |
|------------------------------|------------------|---------|------------------|---------|
| EU Documents | 161 | 16.72 % | 483 | 22.24 % |
| Academic Texts | 641 | 66.56 % | 1123 | 51.70 % |
| Public Speeches and Articles | 97 | 10.07 % | 312 | 14.36 % |
| Tourist Brochures | 64 | 6.65 % | 254 | 11.69 % |
| Total | 963 | 100 % | 2172 | 100 % |

Table 7: Occurrences of *may/might* and *can/could* across text-types¹³

Generally, *might* has developed most remarkably into the epistemic direction. This also explains why the prototypical German equivalent (*könnte/möchte*) is used in fewer than half of the (285) cases.

The following examples illustrate the most important typological ways in which *might* can be translated into German:

- a subjunctive in S16,
- a subjunctive plus an adverb or adjective (*vermutlich*) in S 17,

¹³ The figures have not been corrected for the (few) unwanted 'exceptions', like *May* (month), *might* (strength) or *can* (of tin).

- an inverted word order indicating a conditional clause relationship in S18,
- a non-finite clause that does not include this semantic option, as in S19,
- a simple adjective, such as *vermutlich*, which also expresses *guess*, in S20,
- an adjective (suffix) in S21,
- a noun including a modal component (*Behauptung*/ 'claim') in S22,
- the finite verb in a subordinate clause that could but does not use the subjunctive in S23, and
- no translation at all; even double modality, in *might* as well as in *suggest*, may be ignored, as in S 24

(S16) Thus **might** all the forces be unified, each force merely manifesting but one aspect of a single supersymmetric superforce.

So **wären** alle Kräfte vereinigt, und jede Kraft würde lediglich als ein besonderer Aspekt einer einzigen supersymmetrischen »Superkraft« erscheinen. (Davies/Brown. 1988/1989)

(S17) There it **might** have died, had not Schwarz and his then collaborator Joel Scherk spotted that it could be used in an altogether different, and much more exciting context.

Es **wäre** vermutlich vollständig erloschen, wenn Schwarz und sein Mitarbeiter Joël Scherk nicht gezeigt hätten, daß sie in einem ganz anderen und viel aufregenderen Zusammenhang angewandt werden konnte. (Davies/Brown. 1988/1989)

(S18) The fact that there **might** be a systematic curvature of space on a cosmological scale raises the interesting question of the topology of the universe.

Sollte eine systematische Krümmung des Weltalls in kosmologischem Maßstab existieren, stellt sich die interessante Frage nach der Topologie des Universums. (Davies/Brown. 1988/1989).

(S19) That left only natural increase as a means by which the Crusaders **might** have solved their manpower problems.

Als einzige Methode, ihre personelle Unterlegenheit zu beseitigen, blieb den Kreuzfahrern das natürliche Bevölkerungswachstum. (A.W. Crosby 1986/91)

- (S20) The Guanches sensibly surrendered all flat and open country (and therefore, one **might** guess, most of their grain fields and their flocks)

Die Guanchen gaben, sobald sie die Macht der berittenen Soldaten kennengelernt hatten, das flache und offene Gelände vollständig auf – und damit **vermutlich** auch die meisten ihrer Getreidefelder und Viehherden. (A.W. Crosby 1986/91)

- (S21) Scandinavians, like other northwest Europeans, are among the world's champion milk digesters, which perhaps had effects that **might** not be readily apparent. Dieser Vorteil, den sie mit anderen Nordwesteuropäern teilen, hatte unvorhersehbare Folgen (Simoons 1978: 964 f.) (A.W. Crosby 1986/91)

- (S22) Thus Einstein was led to the idea that gravity **might** be nothing more than geometry – a distortion in the geometry of space.

So kam Einstein zu seiner **Behauptung**, daß Gravitation ihrem Wesen nach nichts anderes als eine Verzerrung der Geometrie des Raumes, darstellt. (Davies/Brown. 1988/1989)

- (S23) Ever since 1947, when manuscripts dating from the first century A. D. were found in caves overlooking the Dead Sea, there has been speculation that Jesus **might** have been connected with the group that produced these documents.

Seit man im Jahre 1947 in den Berghöhlen von Qumran oberhalb des Toten Meeres Manuskripte aus dem 1. Jahrhundert n. Chr. fand, reißen die Spekulationen nicht ab, wonach Jesus mit jener Gemeinschaft in Verbindung stand, die diese Texte verfaßte. (Kee 1990/1993)

- (S24) This **might** suggest one underlying anxiety on which the question is based.

In der Frage steckt eine grundlegende Befürchtung. (John Wilson 1963/1984)

The intricacies of one aspect of modality can best be appreciated in the mother tongue, but the principles can similarly be illustrated using translations in the other direction, of course. Since one of the most common (about 10%) German equivalents of *might* is *vielleicht* (S25), another way of comparing equivalents is to see how many of the German *vielleicht* cases correspond to *perhaps*, which was about 37 %.

(S25) Commercial man **might** be a social but he could never be a wholly political being.

Der homo oeconomicus wäre **vielleicht** noch ein soziales, keinesfalls aber ein wirklich politisches Wesen. (J.G.A. Pocock 1993/1993)

Although the advanced language learner has to be aware of stylistic and collocational variation, *perhaps* and *vielleicht* are prototypical equivalents. Thus qualitative investigations have to be complemented by quantitative ones.

5. Conclusion

5.1 Limitations

For lexical and particularly collocational analyses, even the largest translation corpus pushes the possibilities to the limits. For making comparisons of more than the most frequent lexemes and their collocations, the World Wide Web has to be used as a 'supplementary corpus' (cf. Schmied, 2004). Then we have to bear in mind that the WWW is not a neatly stratified corpus (since it is still strongly biased towards the written, public and formal language variation), and it does not represent world-wide usage either (since the content of web pages is very development- and culture-specific).

Thus individual lexemes (such as *appear*, *seem* or *might* analysed above) have to be seen in relation to the entire field of options expressing modality. Table 8 includes full verbs (*appear* and *seem*), modals (*may* and *might*) and adverbs (*maybe*, *perhaps*, *probably* and *likely*), which overlap in their usages. Their occurrence on the WWW is compared in absolute and relative

terms.¹⁴ This makes it manifest that *may* and *might* are, respectively, used almost 3 and 6 times more often than *maybe*, whereas the other options are clearly in the middle range.¹⁵

| total sites | relationship | appear | seem | may | might | maybe | perhaps | probably | likely |
|-------------------|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| UK 9.25M | absolute: | 1,640,000 | 1,420,000 | 2,600,000 | 1,040,000 | 1,050,000 | 1,660,000 | 1,860,000 | 1,940,000 |
| | relative: | 17.73% | 15.35% | 60.22% | 29.19% | 11.35% | 17.95% | 20.11% | 20.97% |
| | int. share: | 9.19% | 7.96% | 31.22% | 15.13% | 5.88% | 9.31% | 10.43% | 10.87% |
| | int. factor: | 1.6 | 1.4 | 5.3 | 2.6 | 1.0 | 1.6 | 1.8 | 1.8 |
| Australia 3.9M | absolute: | 598,000 | 460,000 | 2,600,000 | 1,040,000 | 389,000 | 558,000 | 654,000 | 761,000 |
| | relative: | 15.33% | 11.79% | 66.67% | 26.67% | 9.97% | 14.31% | 16.77% | 19.51% |
| | int. share: | 8.47% | 6.51% | 36.83% | 14.73% | 5.51% | 7.91% | 9.26% | 10.78% |
| | int. factor: | 1.5 | 1.2 | 6.7 | 2.7 | 1.0 | 1.4 | 1.7 | 2.0 |

Table 8: Occurrences of full verbs (*appear* and *seem*), modals (*may* and *might*) and adverbs (*maybe*, *perhaps*, *probably* and *likely*) expressing modality in UK and Australian sites

Many of the analyses illustrated here are limited to the formal surface level. A deeper-level analysis on the basis of meaning and syntactic categories would have been much more interesting, but this can only be done with a semantically or syntactically tagged corpus. The analysis of specific discourse meanings in an untagged corpus is only possible in exceptional cases. Table 9 gives such an example using the fact that the discourse marker *well* (introducing a new thought or turn in direct or indirect spoken English) is usually followed immediately by a comma (whereas the common adverb *well* is not). However, a ‘manual’ verification shows that several cases would be misclassified in such an automatic approach (e.g. because a comma may also occur after the phrase *as well*), although the distribution remains similar.

¹⁴ The relative parameters include the occurrence of sites in proportion to the total number of WWW sites in the domain (measured in terms of occurrences of *the*), their intra-site share (in per cent of all the search phrases) and their intra-site factor (with the least frequent phrase taken as 1.0).

¹⁵ Obviously, a differentiation into meanings would be useful, but impossible, here. For a discussion of the methodology, opportunities and limitations of using Google output for comparative lexical studies cf. Schmied (2004).

| Text Type | Well, | % | verified |
|------------------------------|-------|-------|----------|
| EU Documents | 2 | 5.56 | 2 |
| Academic Texts | 27 | 75.00 | 16 |
| Public Speeches and Articles | 4 | 11.11 | 2 |
| Tourist Brochures | 3 | 8.33 | 1 |
| Total | 36 | 100 | 21 |

Table 9: Distribution of the discourse marker *well* across text-types

5.2 Future prospects

I hope to have shown that contrastive corpora have contributed everywhere to a more empirical descriptive approach in several branches of applied linguistics or comparative linguistics, especially contrastive linguistics and translation studies, as well as language teaching and learning. They are valuable in proceeding from a more item-and-text-based analysis to a broader co-and-context-based analysis. Of course, corpus-linguistic searches and statistics should not be seen as ‘an end in itself through empty and unnecessary quantitative investigations’, as Tymoczko (1998: 1) points out, but rather as a good starting point for comparative research, which may have to be complemented by specialised ad-hoc monitor ‘corpora’ depending on the specific issue of analysis. However, the few examples presented here suffice to illustrate the old advertising slogan of corpus-linguistic grammar, because the keywords in context can really show ‘more grammar than meets the eye’.

References

- AIJMER, K.; ALTENBERG, B.; JOHANSSON, M. (eds.). (1996) *Languages in Contrast*. Lund: Lund U. P., Lund Studies in English p. 88.
- _____. (1996) ‘Text-based contrastive studies in English. Presentation of a project’. in AIJMER, K./Bengt Altenberg/Mats Johansson eds., p. 73-85.
- ASTON, G. (2002) ‘The learner as corpus designer’. Kettemann, Bernd/ Georg Marko (eds.) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, p. 9-25.

- BAKER, M. (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in Mona Baker/Gill Francis/Tognini-Bonelli(eds.) *Text and Technology: In Honour of John Sinclair*, Amsterdam & Philadelphia: John Benjamins, p. 233-50.
- _____. (1995) 'Corpora in Translation Studies. An Overview and Suggestions for Future Research', *Target* 7(2): p. 223-43.
- _____. (1999) 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics* 4(2): p. 281-98.
- _____. (2002) 'Corpus-based Studies within the Larger Context of Translation Studies'. *Génesis: Revista científica do ISAI* 2: p. 7-16.
- BERNARDINI, S. (2000) *Competence, Capacity, Corpora*. Bologna: CLUEB.
- _____. (2002) 'Educating translators for the challenges of the new millennium: the potential of parallel bi-directional corpora'. in: Maia/Haller/Ulrych eds., p. 173-86.
- BIBER, D. (1993) 'Representativeness in corpus design'. *Literary and Linguistic Computing* 8: p. 243-57.
- BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S. FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- BOWKER, L. (2001) 'Towards a methodology for a corpus-based approach to translation evaluation'. *Meta* 46, p. 345-64.
- CIDE. (1995) = PROCTOR, Paul. (ed.) *Cambridge International Dictionary of English*. Cambridge: C.U.P.
- FRANKENBERG-GARCIA, A. (2002) 'COMPARA -language learning and translation training'. in: Maia, Belinda/Johann Haller/Margherita Ulrych (eds.), p. 187-98.
- GARCIA, N. R. (2002) 'Contrastive linguistics and translation studies interconnected: the corpus-based approach'. *Linguistica Antverpiensia. New Series* 1, p. 393-406.
- HOUSE, J. (2000) 'Concepts and methods of translation criticism: a linguistic perspective'. *Arbeiten zur Mehrsprachigkeit B 10*. Hamburg: Sonderforschungsbereich 538 Mehrsprachigkeit.
- _____. (1997) *Translation Quality Assessment. A Model Revisited*. Tübingen: Narr.
- HUDDLESTON, R./Geoffrey K. P. (2002) *The Cambridge Grammar of the English Language*. Cambridge: C.U.P.
- JOHNS, T. (1986) 'Microconcord: a language-learner's research tool'. *System* 14/2: p. 151-62.

TRADTERM, **10**, 2004, p. 83-115

- _____. (1993) 'Data-driven learning: an update.' *TELL & CALL* 1993/2, p. 4-10.
- JOHANSSON, S. (2004) 'Multilingual Corpora: Models, Methods, Uses'. (this volume).
- JOHANSSON, S.; Knut H. (1994) 'Towards an English–Norwegian parallel corpus'. Fries, Udo/Gunnel Tottie/Peter Schneider. (eds.) *Creating and Using Language Corpora*. Amsterdam/Atlanta: Rodopi, p. 25-37.
- KENNEDY, G. (1998) *An Introduction to Corpus Linguistics*. London and New York: Longman.
- KENNY, D. (1998) 'Corpora in Translation Studies', in Mona Baker ed. *Routledge Encyclopedia of Translation Studies*, London and New York: Routledge, p. 50-3.
- KENNY, D. (1999). 'The German-English Parallel Corpus of Literary Texts (GEPOLT): A Resource for Translation Scholars'. *Teanga* 18: p. 25-42.
- KÜHLWEIN, W.; THOME, G.; WOLFRAM WILSS, H. (1980) *Kontrastive Linguistik und Übersetzungswissenschaft*. München: Fink.
- LAURIDSON, K. (1996) 'Text corpora and contrastive linguistics: which type of corpus for which type of analysis'. Aijmer, Karin/Bengt Altenberg/Mats Johansson eds., p. 63-71.
- MAIA, B.; HALLER, J.; ULRYCH, M. (Eds.). (2002) *Training the Language Services Provider for the New Millennium*. Porto: Faculdade de Letras, Universidade do Porto.
- MCENERY, T.; WILSON, A. (2001, 1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- MEYER, Charles F. (2002) *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.
- MUKHERJEE, J. (2002) *Korpuslinguistik und Englischunterricht. Eine Einführung*. Frankfurt am Main: Peter Lang.
- OOI, V. (1998) *Computer Corpus Lexicography*. Edinburgh: Edinburgh U.P.
- RÁBADE, L. I.; SUÁREZ, S. M. D. (Eds.). (2002) *Studies in contrastive linguistics. Proceedings 2nd International Contrastive Linguistics Conference*. Santiago, October, 2001. Santiago: Universidade de Santiago de Compostela.
- SAJAJAARA, K. (1996) 'New challenges for contrastive linguistics'. in Aijmer, Karin/Bengt Altenberg/Mats Johansson (eds.), p. 17-36.
- SCHMIED, J. (1990) 'Corpus linguistics and the nativization of English'. *World Englishes* 9, p. 255-68.
- _____. (1994) 'Translation and cognitive structures'. *Hermes. Journal of Linguistics* 13: p. 169-81.

- _____. (1998) 'To choose or not to choose the prototypical equivalent.'
SCHULZE, Rainer, ed. *Making Meaningful Choices in English. On Dimensions, Perspectives, Methodology, and Evidence*. Tübingen: Gunter Narr, p. 207-22.
- _____. (1999) 'Applying contrastive corpora in modern contrastive grammars: The Chemnitz Internet Grammar of English'. HASSELGARD, H.; OKSEFJELL, S. (Eds.) *Out of Corpora. Studies in honour of Stig Johansson*. Amsterdam: Rodopi, p. 21-30.
- _____. (2002a) 'A translation corpus as a resource for translators: the case of English and German prepositions'. Maia, Belinda/Johann Haller/Margherita Ulrych eds., p. 251-69.
- SCHMIED, J. (2002b) 'Prototypes, transfer and idiomaticity: an empirical analysis of local prepositions in English and German'. RABADE, L. I.; SUAREZ, S. M. D. (Eds.), p. 947-59.
- _____. (2004) 'New ways of analysing ESL on the www with WebCorp and WebPhraseCount'. Renouf, Antoinette ed. *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.
- SCHMIED, J.; HAASE, C. (2003) „Grammatik lernen im Internet: Die Chemnitz Internet-Grammatik'. KEITEL, E.; BOEHNKE, K.; WENZ, K.; (eds.) (2003). *Neue Medien im Alltag: Nutzung, Vernetzung, Interaktion*. Lengereich: Pabst Science Publishers, p. 109-26.
- SCHMIED, J.; SCHÄFFLER, H. (1996) 'Approaching translationese through parallel and translation corpora'. PERCY, C.; LANCASHIREI.; MEYER, C. (Eds.) *Synchronic Corpus Linguistics*. Amsterdam: Rodopi, p. 41-56.
- SCHMIED, J.; SCHÄFFLER, H. (1997) 'Explicitness as a universal feature of translation'. Ljung, Magnus, ed. *New Ways in Corpus Linguistics*. Amsterdam: Rodopi, p. 21-34.
- SINCLAIR, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford U.P.
- TYMOCZKO, M. (1998) 'Computerized Corpora and the Future of Translation Studies', *Meta* 43: p. 652-9.
- ZANETTIN, F. (2000) 'Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis', in Maeve Olohan (ed) *Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects*, Manchester: St. Jerome, p. 105-18.