R.R.K. Hartmann
University of Exeter

The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography

Abstract

The paper is intended to demonstrate the practical applicability of the theoretical notion of 'contrastive textology' (Hartmann 1980) to bilingual lexicography. By means of a systematic analysis of parallel texts from corresponding genres in particular pairs of languages it is possible to generate matching words and their collocations which can be codified as translation equivalents in bilingual dictionaries. Promising work has been done to develop computer-aided techniques for utilizing such parallel text corpora in the search for lexical equivalence Examples from English and German.

1. Introduction

Enormous strides have been made in the last few years in applying the findings of text linguistics and corpus technology to the field of lexicography. The former (sometimes under the heading of 'combinatorics') covers such phenomena as 'collocation' within sentences and 'cohesion' between sentences as well as the 'pragmatic' embedding of discourse in context, with interesting ramifications into sociolinguistics. The latter (under the title of 'corpus linguistics') is concerned with techniques for compiling and exploiting textual databases in an effort to document the whole range of linguistic structure and extra-linguistic knowledge, which overlaps with the territory of artificial intelligence.

The aim of this paper is to relate these developments to a particular problem in bilingual lexicography, viz. translation equivalence. Starting with the theoretical idea of 'contrastive textology', it traces some of the possibilities and practical problems of computer-aided parallel text analysis for the benefit of compilers (and users) of bilingual dictionaries.

2. Contrastive textology

In a thought-provoking paper on the theme of data-gathering, Bujas (1975) asked what is needed in the efficient updating of a bilingual dictionary for a language pair like English and SerboCroat. To answer the question, he had to employ a large number of student helpers to manually excerpt texts from newspapers and magazines and check their appropriacy for a revised edition of the dictionary. Today much of this work can be done by relying on

existing text corpora or by computer scanning and concordancing (cf. Flowerdew & Tong 1994).

However, while the use of text corpora is fairly firmly established in monolingual general–purpose, pedagogical and terminological lexicography, much remains to be done in bilingual lexicography. The stumbling block here is the problem of translation equivalence, which requires an interlingual approach.

My own book *Contrastive Textology* (Hartmann 1980) was intended as a programmatic plea for a systematic combination of contrastive analysis and discourse analysis. Its double purpose was the improved description of the linguistic facts (at the level of the text) of pairs of languages and improved problem–solving in practical domains such as translation, foreign–language teaching and bilingual lexicography.

Examples of problems awaiting solutions are the following: What are the means available in different languages for anaphora and other forms of cross–reference in text? What are the signals that delimit successive discourse blocs? What are the factors that determine register and genre ranges in different languages? What shifts are required in the translation of texts from one language to another? (cf. Hartmann forthcoming)

Examples of different approaches to these problems include 'comparative stylistics', 'contrastive rhetoric', 'cross–cultural discourse grammar', 'comparative discourse analysis', and many others (cf. Péry–Woodley 1990). For the field of bilingual lexicography, the idea of collecting and comparing 'parallel texts' seems particularly promising – see below.

3. Translation equivalence

The traditional notion of equivalence was to relate words to their counterparts as corresponding formal units in parallel linguistic systems, a view that was strengthened by the apparent ease with which bilingual dictionaries can supply ready–made lexical equations for insertion into the appropriate portion of a text (cf. Zgusta 1984).

However, the semantic abstraction that is built into the lexical inventory of the dictionary has deprived each of these words of their natural context, and the translator must compensate for the lack of contextual information from his/her own bilingual discourse competence, particularly in that most intractable area of 'culture–specific' vocabulary. More recent research (cf. Hartmann 1985 and 1992a, Hatim & Mason 1990) has stressed the approximative nature of these equivalence creation processes.

From this vantage point, the contrastive textologist will want to go beyond the mere comparison of given parallel texts as translation products and search instead for the actual code–switching operations that allowed the competent translator to 'find' a suitable target–language equivalent in the first place. This is of direct relevance to bilingual lexicography: the dictionary maker needs not only to codify the results of past translation acts – however

they may have been achieved – , but also to have an awareness of the techniques that can be used to bring about such translation equivalence.

4. Dictionary equivalents

The coverage of lexical equivalents in the bilingual dictionary is a hit–and–miss, trial–and–error task capable of empirical observation and systematization (and thus improvement). I have consulted a number of monolingual and bilingual dictionaries (Hartmann 1992b) to check how they treat a range of 14 regionally marked lexical items in British English and Austrian German and their equivalents – see Fig. 1.

The conclusion I came to was that the coverage of the translation equivalents in bilingual dictionaries, while in general quite reasonable and no worse than that of even more specialized monolingual dictionaries, is based on an element of chance which we should attempt to reduce in future.

5. Parallel text corpora

One solution to the problem of systematizing the discovery of translation equivalents, as suggested by the proponents of the various approaches to contrastive textology, lies in the comparison of so–called parallel texts, i.e. bits of discourse from corresponding varieties or text types in the two languages in question. If we knew, or so the argument goes, what the semantic ranges and collocational restrictions of words were in the textual contexts of one language, then we could match them in parallel texts from the other language.

This is exactly what John Laffling (1991) attempted. He built up a corpus of parallel texts of party political manifestoes in English and German, in other words, the political programmes of the British Labour Party and the German Socialists, the British Conservatives and the CDU–CSU in Germany, and the Greens in both countries. By means of an algorithm which computer– matched words and phrases in these parallel texts, he managed to extract from them the naturally occurring translation equivalents.

In Fig. 2 I present a small portion of Laffling's results in relation to dictionary coverage. The information is arranged in four columns. In the first, on the left, there are four phrases (you might call them political clichés) from the German corpus.

The second column contains English equivalents of these phrases as found in the official and unofficial translations of the texts, which on the whole are fairly literal renderings.

Column 3 is the most interesting: here we have 'real' textual equivalents not found in translations, but in separately formulated parallel texts, i.e. the party manifestoes of the corresponding political party in the United Kingdom.

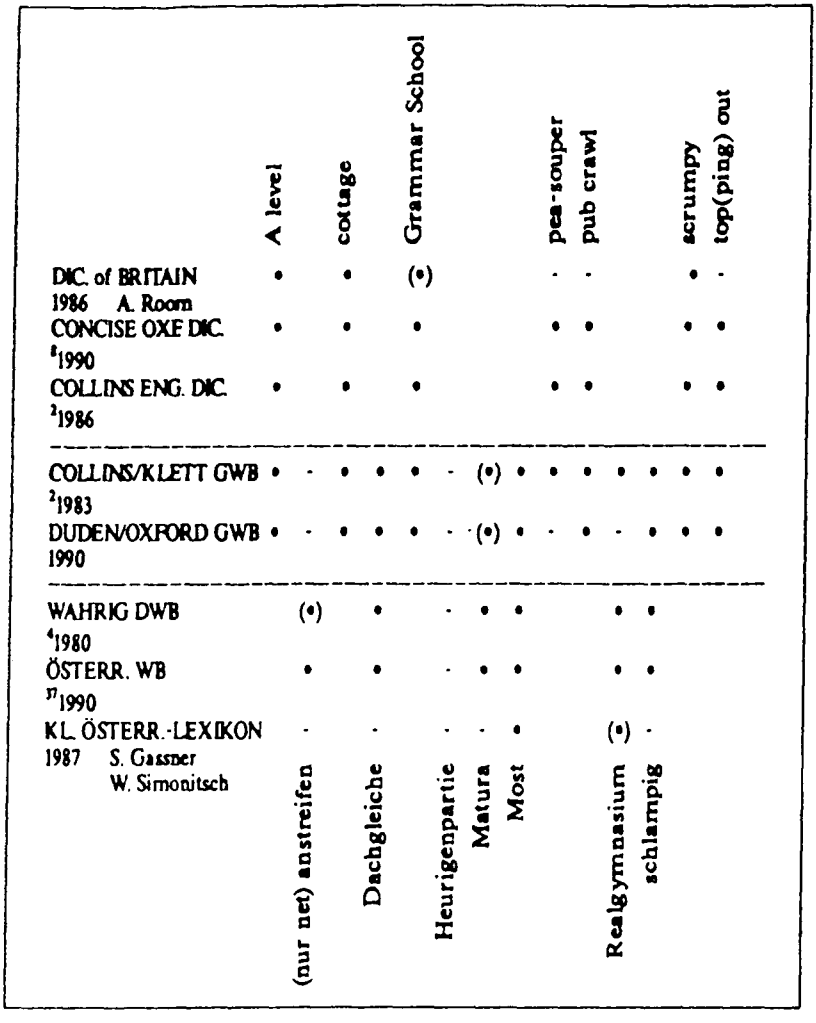

Figure 1: Dictionary coverage.

These are phrases that occur independently in matching discourse, with identical meanings, although formally less close than the translation equivalents in Column 2.

If we now go a step further and ask how these lexical equivalents are codified in the bilingual dictionaries of the two languages, we may be in for a surprise. Many cannot be found at all in the current dictionaries. In Column 4 I have exemplified the coverage in only one of the better German–English dictionaries, the DUDEN–OXFORD (1990). None of the English phrases supplied in that bilingual dictionary offers a translation equivalent that fully

matches the naturally occurring phrases from the English parallel texts, although they come reasonably close.

	translation	parallel texts	DUDEN-OXFORD
politische Auseinandersetzung	political argument	political debate	political [+] debate
Bildungsangebot	educational opportunities	educational provision	educational [+] offer
Not und Elend	want and poverty	misery and hardship	poverty (and hardship) [+] misery
breite Schichten der Bevölkerung	broad layers of the population	large sections of the population	broad sections of the population

Figure 2: Parallel text analysis.

[The full version of this paper will elaborate on the problems and possibilities of the methodology of computer-assisted generation of translation equivalents from other parallel text corpora, based on the results of research to be undertaken at Macquarie University in August 1994.]

6. Implications for bilingual lexicography

The existing literature on bilingual dictionary-making (cf. Bartholomew & Schoenhals 1983, Marello 1989, Svensn 1993) is strangely silent on these issues. However, recent impulses have come from machine translation (see Laffling 1991 as discussed in Section 5 above), artificial intelligence and computer technology.

Kenneth Church and William Gale (1991), for example, have explored the use of parallel text concordances, such as those based on the French-English Canadian Hansard, if in remarkable ignorance of the pioneering theoretical work mentioned above. Church & Gale claim that translation equivalents can be extracted from such bilingual corpora by aligning the parallel texts at the sentence level.

Eugenio Picchi and his Pisa colleagues (1992) have proposed a 'workstation' for lexicographers intent on monitoring and processing lexical

equivalents derived from English–Italian text corpora. Each of these sets of bilingual texts from different language varieties can be first 'synchronized', using morphological procedures and information from an electronic bilingual dictionary, and then searched for 'direct links' between the texts, which produces a choice of potential translation equivalents.

I would venture to suggest that we are not far off the time when these techniques not only become more widely available, but also could help us design bilingual thesauruses (cf. Hartmann 1994) from text corpora taken from corresponding genres in selected pairs of languages and thus benefit dictionary compilers and dictionary users, especially translators.

References

(a) Cited dictionaries

Collins Dictionary Of The English Language. Ed. L. Urdang. London & Glasgow: Collins 2nd ed. 1986

Collins–Klett [Pons] Grosswörterbuch Deutsch–Englisch Englisch–Deutsch. Comp. P. Terrell et al. Glasgow: Collins & Stuttgart: E. Klett 1980/1983

Concise Oxford Dictionary Of Current English. Ed. R.E. Allen. Oxford U.P. 8th ed. 1990

Dictionary Of Britain. An A To Z Of British Life. Comp. A. Room. Oxford U.P. 1986/1990

Duden–Oxford Grosswörterbuch Englisch [–Deutsch]. Eds. W. Scholze–Stubenrecht & J. Sykes. Mannheim: Dudenverlag & Oxford U.P. 1990

Kleines Österreich–Lexikon. Wissenswertes Ber Land Und Leute. Comp. S. Gassner & W. Simonitsch. München: C.H. Beck

Österreichisches Wörterbuch. Wörterbuchstelle/Bundesministerium für Unterricht [et al.]. Vienna: Österreichischer Bundesverlag 37th ed. 1990

Wahrig Deutsches Wörterbuch. Comp. G. Wahrig [et al.]. Gütersloh: Bertelsmann & Mosaik 4th ed. 1980

(b) Other literature

Bartholomew, D.A. & Schoenhals, L.C. 1983. *Bilingual Dictionaries for Indigenous Languages.* Mexico: Summer Institute of Linguistics.

Bujas, Z. 1975. "Testing the performance of a bilingual dictionary on topical current texts" *Studia Romanica et Anglica Zagrabiensia* 39: 193–204.

Church, K. & Gale, W. 1991. "Concordances for parallel text", in *Using Corpora.* Proceedings of the 7th Annual Conference of the Centre for the New Oxford English Dictionary and Text Research [Oxford] ed. L.M. Jones. Oxford U.P., 40–62.

Flowerdew, L. & Tong, K.K. Eds. 1994. *Entering Text.* Papers from the HKUST/GIFL Joint Seminar on Corpus Linguistics and Lexicology. Hong Kong: HKUST Language Centre.

Hartmann, R.R.K. 1980. *Contrastive Textology.* Comparative Discourse Analysis in Applied Linguistics (Studies in Descriptive Linguistics 5). Heidelberg: J. Groos.

Hartmann, R.R.K. 1985. "Contrastive text analysis and the search for equivalence in the bilingual dictionary" in *Symposium on Lexicography II* ed. K. Hyldgaard–Jensen & A. Zettersten (Lexicographica Series Maior 5). Tübingen: M. Niemeyer, 121–132.

Hartmann, R.R.K. 1992a. "300 years of English–German language contact and contrast: The translation of culture–specific information in the general bilingual dictionary" in *Language and Civilization. A Concerted Profusion of Essays and Studies in Honour of Otto Hietsch* ed. C. Blank. Frankfurt: P. Lang, 300–327.

Hartmann, R.R.K. 1992b. "Contrastive linguistics: (How) is it relevant to bilingual lexicography?" in *New Departures in Contrastive Linguistics* Med. C. Mair & M. Markus (Innsbrucker Beiträge zur Kulturwissenschaft, Anglistische Reihe 4 & 5). Innsbruck: Universität, Institut für Anglistik, I: 293–299.

Hartmann, R.R.K. 1994. "The onomasiological dictionary in English and German. A contrastive textological perspective" in *Die Welt in einer Liste von Wörtern/The World in a List of Words*. ed. W. Hüllen (Lexicographica Series Maior). Tübingen: M. Niemeyer, 172–185.

Hartmann, R.R.K. forthcoming. "Contrastive textology, bilingual lexicography and translation" in *Encyclopedic Dictionary of Chinese–English/English–Chinese Translation* ed. Chan Sin–wai. Hong Kong: Chinese University Press.

Hatim, B. & Mason, I. 1990. *Discourse and the Translator* (Language in Social Life Series). London: Longman.

Laffling, J. 1991. *Towards High–Precision Machine Translation, Based on Contrastive Textology* (Distributed Language Translation 7). Berlin: Foris Publications.

Marello, C. 1989. *Dizionari bilingui con schede sui dizionari italiani per francese, inglese, spagnolo, tedesco* (Fenomeni Linguistici 6). Bologna: Zanichelli.

Péry–Woodley, M.–P. 1990. "Contrasting discourses: Contrastive analysis and a discourse approach to writing" *Language Teaching* 23:143–151.

Picchi, E. et al. 1992. "The Pisa Lexicographic Workstation: The bilingual component" in *EURALEX '92 Proceedings ed.* H. Tommola et al. (Studia Translatologica A.2). Tampere: Yliopisto, I: 277–285.

Svensn, B. 1993. *Practical Lexicography. Principles and Methods of Dictionary–making*. Oxford: Oxford U.P.

Zgusta, L. 1984. "Translational equivalence in the bilingual dictionary" in *LEXeter '83 Proceedings* ed. R.R.K. Hartmann (Lexicographica Series Maior 1). Tbingen: M. Niemeyer, 147–154.