Lexical and Translation Equivalence in Parallel Corpora

Tamás Váradi

Linguistics Institute, Hungarian Academy of Sciences
H-1014 Budapest Színház u 5-9
varadi@nytud.hu

Abstract

In the present paper we intend to investigate to what extent use of parallel corpora can help to eliminate some of the difficulties noted with bilingual dictionaries. The particular issues addressed are the bidirectionality of translation equivalence, the coverage of multiword units, and the amount of implicit knowledge presupposed on the part of the user in interpreting the data. Three lexical items belonging to different word classes were chosen for analysis: the noun *head*, the verb *give* and the preposition *with*. George Orwell's novel 1984 was used as source material, which is available in English-Hungarian sentence aligned form. It is argued that the analysis of translation equivalents displayed in sets of concordances with aligned sentences in the target language holds important implications for bilingual lexicography and automatic word alignment methodology.

1. Introduction

It is a truism to state that languages carve up reality into different sets of lexical items. Hence the chunks of experience embodied in lexemes will inevitably differ from each other, which rules out any neat correspondences at the lexical level between languages. Bilingual dictionaries traditionally attempt to map cross-linguistic equivalence by defining a headword in terms of a list of foreign language equivalents. Typically, they provide very little information as to which alternative would be suitable in the given context. In fact, as has been recently pointed out by Wolfgang Teubert (1999) bilingual equivalence between dictionary entries is very often not bi-directional. As a useful complement to bilingual dictionaries and conceptual ontologies, Teubert recommends the corpus linguistic approach, which he illustrates through data from monolingual corpora. In this paper we explore this line of research by adducing evidence from a bilingual parallel corpus. For a related effort involving the same parallel corpus see (Ide 2000), which shows how selected words in English are lexicalized differently across a variety of languages.

2. The problem with bilingual dictionaries

Even if available in electronic format, ordinary dictionaries created for human users present certain problems for natural language processing (Boguraev and Briscoe 1989). Some limitations derive from the fact that lexicographers inevitably rely on the co-operation of the readers to exploit the information compiled in the body of dictionary entries. One source of relatively low level difficulties is that all dictionaries make shortcuts in presenting the data in an effort to compress maximal amount of information in the available space. Users are expected to decode and apply the formatting conventions that are usually set out in an introductory section. This is a task that is not always trivial to automate as was reported by the CONCEDE project for several dictionaries (Erjavec et el. 1999). More serious than this procedural difficulty are the general deficiencies in content. As was demonstrated by Teubert (op.cit.) bilingual dictionaries tend to give a list of equivalents with very little help (apart from usage notes) as to which one it to be used in the particular context the dictionary user needs. There is, furthermore, an undue preponderance of single-word equivalents at the expense of multi-word units. When one tries to look up the equivalents in the other side of a bilingual dictionary, one finds surprisingly few bidirectional equivalents. Teubert presents the situation graphically in Figure 1 through the analysis of the semantic field associated with the German word *Trauer*. The figure was arrived at by first looking up the equivalents of Trauer in Langenscheidt Enzyklopädisches Wörterbuch and successively looking up the equivalents found in one language in the other side of the dictionary until the senses started to become remote from the semantic field of the original word.

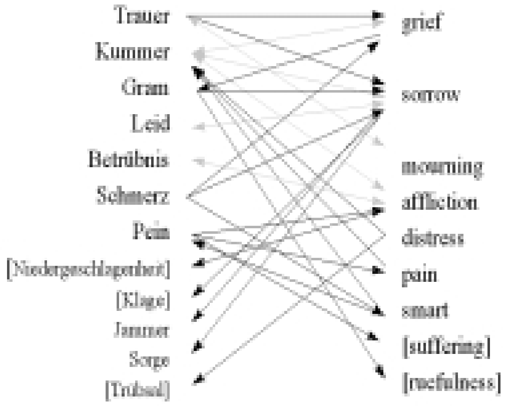


Figure 1: Successive trace of German-English translation equivalents related to *Trauer* (based on Teubert 1999)

3. The rationale for the present work

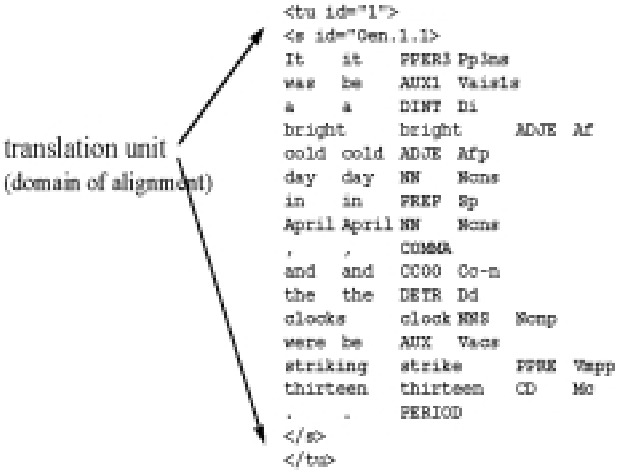
In the present paper we intend to investigate to what extent use of parallel corpora can help to eliminate some of the difficulties noted with bilingual dictionaries. It was assumed that parallel corpora are amenable to the same procedure of traversal of translation equivalents. At the same time, the data are sufficiently different in key aspects to warrant the replication of the methodology.

In particular, we set out to investigate what picture emerges from parallel corpora with regard to a) consistency of coverage (bidirectional vs. unidirectional equivalences) b) coverage of multiword units, collocations c) the amount of user knowledge presupposed and d) what implications are there for bilingual lexicography and NLP, e.g. automatic word alignment.

4. Methodological issues

4.1 Data and encoding scheme

As source data, we used the sentence aligned Hungarian-English parallel corpus of Orwell's 1984 developed in the MULTEXT-EAST project (Erjavec and Ide 1998). The corpus was processed with the IMS Corpus Workbench system (Christ 1994) developed by the Institute für maschinelle Sprachverarbeitung of Stuttgart University. We used a slightly simplified version of the corpus encoding scheme developed by Erjavec (1999) for the ELAN Slovene/English. Figure 2 shows a sample of the corpus annotation. There were only two tags used to mark up the structure of the texts: <tu> (translation unit) and <s> (sentence) with their respective id attributes. As against the relatively simple structural mark-up, the annotation attached to each token was substantially richer in content. Each line of text included the word form, lemma, corpus tag and the morphosyntatctic description, arranged in a tabular format. While the IMS Workbench Tool is somewhat limited in handling SGML tagged corpora, it offers remarkable facility in handling linguistic mark-up associated to each token. Technically, the two languages of the parallel corpus are stored in separate files and the alignment is made with reference to the respective offset figures of the corresponding translation units.

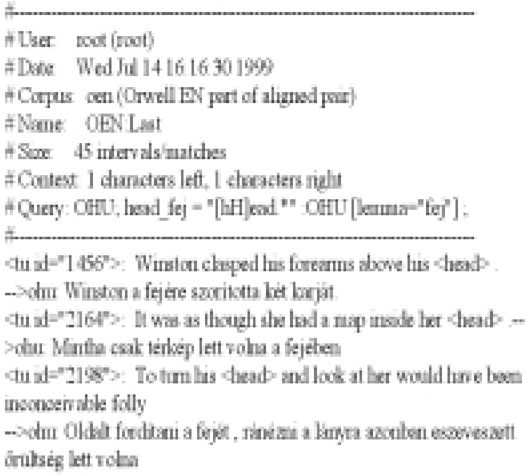


```
<tu id="1">
<s id="Oen.1.1>
It      it      PPER3 Pp3ns
was     be      AUX1  Vais1s
a       a       DINT  Di
bright          bright      ADJE  Af
cold    cold    ADJE  Afp
day     day     NN    Ncns
in      in      PREP  Sp
April   April   NN    Ncns
,       ,       COMMA
and     and     CCOO  Cc-n
the     the     DETR  Dd
clocks          clock NNS   Ncnp
were    be      AUX   Vacs
striking        strike      PPRE  Vmpp
thirteen        thirteen    CD    Mc
.       .       PERIOD
</s>
</tu>
```

Figure 2: The encoding of the data

Figure 3 shows a sample output of a query. The query string `[hH]ead.* :OHU [lemma="fej"]` is a regular expression designed to retrieve all occurrences of any inflected forms of the word *head* (whether it begins in lowercase or uppercase) where the corresponding alignment unit in the Hungarian corpus OHU includes the lemma *fej*.

It should be noted that because the two parts of the parallel corpus were aligned at the sentence level, it was not possible to establish automatically whether the two words in the search expressions were actually translation equivalents. All that can be stated with certainty is that the aligned sentences contained the two words in question. The corresponding parts of the aligned sentence pair had to be related manually.



```
#------------------------------------------------------------------------
# User:    root (root)
# Date:    Wed Jul 14 16 16 30 1999
# Corpus:  oen (Orwell EN part of aligned pair)
# Name:    OEN Last
# Size:    45 intervals/matches
# Context: 1 characters left, 1 characters right
# Query: OHU; head_fej = "[hH]ead.*" :OHU [lemma="fej"] ;
#------------------------------------------------------------------------
<tu id="1456">:  Winston clasped his forearms above his <head> .
-->ohu: Winston a fejére szorította két karját.
<tu id="2164">:  It was as though she had a map inside her <head> .--
>ohu: Mintha csak térkép lett volna a fejében
<tu id="2198">:  To turn his <head> and look at her would have been
inconceivable folly
-->ohu: Oldalt fordítani a fejét , ránézni a lányra azonban eszeveszett
őrültség lett volna
```

Figure 3: A sample query output

4.1. The analysis

We have selected three English lexemes for analysis ie. *head*, *give* and *with*. By focussing on three words of so radically different parts of speech we intended to examine whether our findings were sensitive to word class membership. In schematic form, we carried out our analysis through the following steps.

Find prototypical equivalents for the English words. For each of the three words it was easy to find single uncontested candidates for this status: *head*=fej, *give=ad with*=*-val/-vel.* (*-val/-vel* are variant forms of the instrumental suffix governed by vowel harmony.) Below we will refer to members of prototypical equivalent pairs simply as L_1 word and L_2 word

Generate three concordance sets where

L_1 word is translated with L_2 word

L_1 word is translated with non-L_2 word

L_2 word is translated with non-L_1 word

To automate this step a perl script was developed that produced the three sets from two words specified on the command line.

Repeat step 2) with other L_1 and L_2 words from 2b) and 2c) until the semantic field of the original English lexeme seems to be saturated.

4.2. Limitations of the approach

Our analysis inevitably faced certain limitations owing to the scope and nature of the source data used. One immediately obvious constraint was the size of the data which was about a hundred thousand words. While monolingual corpora do exist for Hungarian as well – the Hungarian National corpus currently number more than 80 million words (Váradi 1999), parallel corpora are much harder to come by even for other language pairs let alone for Hungarian-English. Another such corpus is Plato's Republic developed as a TELRI joint research effort but the Hungarian English alignment was not available to us at the time the work reported here was undertaken.

Apart from size, another practical limitation was the rather limited language variety used for source data. Again, this problem could be remedied with the extension of the data not just in terms of size and register as well.

A third peculiarity of the data that one must bear in mind is its inherently unidirectional nature. Although it is tempting to look at the aligned sentences from either direction, it still remains true that for any pair of sentences one is the source and the other is the target of the translation. Hence, we simply cannot speak of ‘the translation equivalents’ of any Hungarian word in our data. This deficiency could be compensated by involving data that are translations of Hungarian source texts though this could raise all sort of issues about how close a match there is between the two source texts. In this sense, there is hardly any genuine bidirectional parallel corpus.

5. Findings

5.1. Prototypical vs. other equivalents

Table 1 presents the statistical summary of our findings. The figures afford several interesting conclusions. It appears that the ‘fit’ between the actual translation equivalence and the presumed prototypical equivalents, as measured in the ratio of the prototypical cases within the total, varies with the language as well as the word class if not the individual lexeme. For example, while close to 70 % of the instances of *head* were rendered with the expected prototypical equivalent *fej* in Hungarian, the same ratio for *with* is 54 % and *give* is translated with *ad* in less than 25% of the cases. If we look at all occurrences of the Hungarian equivalents, we find the same ranking of the items in terms of the ratio of the prototypical equivalent to all other translation equivalents (*fej, -val/-vel, ad*) but at a higher level (*80%, 44%, 38%*). Recall that given the unidirectional nature of our text data, one should interpret the Hungarian figures for *fej*, for example, as the number of times *fej* was used as the translation of *head* vs. of other English words.

	head	other	total
fej	45	11	56
other	21		
total	66		

	give	other	total
ad	26	43	68
other	79		
total	105		

	with	other	total
-val/-vel	337	412	759
oher	285		
toal	622		

Table 1: The actual distribution of the assumed prototypical translation equivalents

5.2. A close up profile

Figure 4 shows the distribution of all the translation equivalents of the word *head* found in our data. The figure on the left shows a complete listing of the equivalents summarized in Table 1. It traces the corresponding items in both directions at one level of depth. The graphic display of the ‘other’ variants make it immediately clear that the spread of the Hungarian translation equivalents of *head* is much wider than the range of words rendered as *fej*. Except for the last two cases all uses of the word *head* made reference to the body part in a non-metaphorical sense. It is interesting to note that some uses were rendered with a verb or verb phrase (*eszébe jut - come to one’s head, felfigyel ‘listen up’ - raised his head).* Also note the number of cases (4) where there was no English source for *fej* at all.

The diagram on the right in Figure 4 traverses the links between translation equivalents one step further, eliminating for clarity all the cases with a single occurrence.

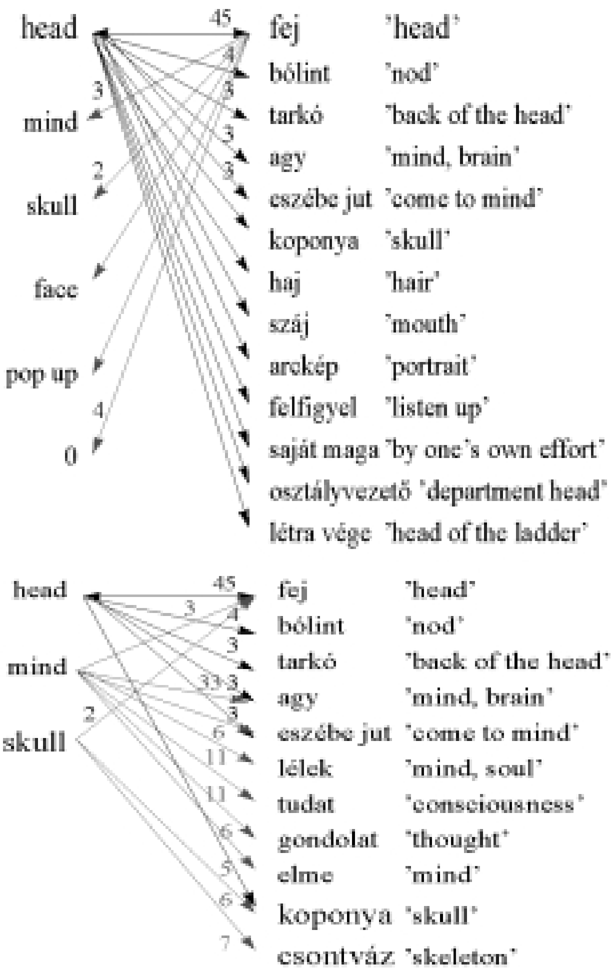


Figure 4: Tracing the translation equivalents of ‘head’

head – haj ‘hair’
<tu id=”5920”>: He plucket at Winston’s <head> and brought away a tuft of heair.
ohu: Belemarkolt Winston **hajába**, és kihúzott egy csomót.
head – száj ‘mouth’
<tu id=”5294”>: And the few you have left are dropping out of your <head>.
->ohu: S az a néhány is, ami még megvan, kiesik a **szádból**.
head – saját maga ‘by one’s own effort’
<tu id=”693”>: A great deal of the time you were expected to make it up out of your <head>.
->ohu: Az ember gyakran **saját maga** volt kénytelen kitalálni.

Figure 5: A sample of the translation equivalents in context

Figure 5 displays some instances where the Hungarian equivalents of *head* (i.e. *száj* ‘mouth’, *saját maga ‘by one’s own* effort’) may seem odd when viewed purely at the lexical level. On the strength of the English examples alone it is easy to see that there is nothing contrived about the uses involved here, yet any lexicographer would probably feel reluctant to introduce such equivalence as *head - száj ‘mouth’* in a dictionary.

The richness of variety of data brought to the fore with this method is further illustrated by Figure 6 displaying the different expressions rendered as *eszébe jut* ‘come to mind’. It is important to note that none of the corresponding items is a single word unit.

When we turn to the verb *give* we find that it is practically impossible to even describe the bare items standing in correspondence without making recourse to multi-word expressions. The data in Figure 6 clearly argues for the importance of the context in even defining the units that enter into a bilingual equivalence relationship. Note that without considering the context in which the translation equivalents occur one may get paradoxical correspondences like *give* = *kap ‘get’*. Such cases can only be interpreted if the different organisation of the sentences that they occur in are also considered. In other cases, it is enough to make reference to the typical objects that the item co-occurs with. One feature in Hungarian that provides for a proliferation of equivalents in English is the presence of the coverb particle that creates new meanings of the verb stem that are often rendered in English with a separate lexeme. Examples include *kiad – publish/issue, átad - hand over, elad - sell* etc.

give	→	kap ‘get, receive’
was given as	→	jelölték meg ‘was marked’
give sb. the impression	→	az volt az érzése ‘had the feeling’
give a glance	→	körülnézett ‘looked around’
give his name	→	megmond ‘say’
give way to	→	következett ‘followed’
give off (smell)	→	áraszt ‘ooze’
give one away	→	elárulhat ‘could betray’
give up trying	→	felhagynak ‘abandon/cease to do’
at any given moment	→	mindig ‘always’
by a given date	→	záros határidőn belül
with no reason given	→	minden indoklás nélkül
don’t give a damn	→	senkinek sem akarunk ártani

Figure 6: Translation equivalents of *give*

6. Conclusions

We do not have the space here to discuss the data uncovered by the analysis in the detail that it clearly merits. However, we are positive that the evidence presented above is sufficient to draw the following conclusions.

The corpus linguistic approach advocated by Teubert has received ample corroboration from the bilingual corpus evidence presented. On the issue of bi-directionality of equivalents, we did not find that the set of bilingual equivalents in the corpus formed a closed set either. However, this may well be due to the sparseness of data used in this pilot experiment.

Our findings have important implications for bilingual lexicography. The most important point to note is the vital need to integrate corpus evidence. Enriching the dictionary with contextual evidence serves to eliminate several shortcomings noted earlier. As the data is embedded in context, it will almost inevitably brings with it guidance as to how the particular item is to be used. Showing usage through examples will also obviate the need for terse and highly abstract formulations. True, the intuitions of the dictionary users are required here too but developing intuitions through actual language data is a task that the average human user is better equipped to handle than dealing with an arid list of bilingual equivalents.

More extensive direct integration of the context should also narrow the current gap between lexical and textual equivalence. We have presented numerous examples for translation equivalents that make perfect sense in the particular context they are used which, however, may seem puzzling if not downright false when viewed out of context. Any attempt to base definitions on real translation equivalence will result in more numerous use of multi-word expressions simply because most of the time it is just not feasible and certainly not practicable to tease out some single word and equate it with another one in the target language at a relatively abstract level. The data presented in the paper clearly suggests that one can only do justice to the intricate and rich texture of context if the two languages are related not at the lexical level of the word but rather at the contextual level embodied in phrases.

The difficulties of pinning down bilingual equivalence on the individual words also has implications for automatic word alignment methodology. No matter how wide a window one establishes within which to scan for equivalents, as long as the search is centred on individual words, the procedure is faced with a serious limitation.

7. References

Boguraev, B. and E. J. Briscoe, 1989. “Computational Lexicography for Natural Language Processing.” London and New York: Longman.

Oliver Christ 1994. “A Modular and Flexible Architecture for an Integrated Corpus Query System.” Papers in Computational Lexicography COMPLEX’94, ed. by Kiefer et al., 23-32. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Erjavec, Tomaž and Nancy Ide, 1998. “The MULTEXT-East corpus.” First International Conference on Language Resources and Evaluation, LREC’98 ed. by Rubio et al., 971-974. Granada: ELRA.

Erjavec, Tomaž, 1999. “Making the ELAN Slovene/English corpus.” Proceedings of the Workshop Language Technologies – Multilingual Aspects, ed. by Špela Vintar., 23-30. Ljubljana: Department of Translation and Interpreting.

Erjavec, Tomaž, Dan Tufiş and Tamás Váradi, 1999. “Developing TEI–Conformant Lexical Databases for CEE Languages.” Papers in Computational Lexicography COMPLEX’99, ed. by Kiefer et al., 205-209. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Ide, Nancy, 2000 “Parallel Translations as Sense Discriminators” International Conference on Language

Resources and Evaluation, LREC'2000, (this volume), Athens:ELRA.

Teubert, Wolfgang, 1999. "Starting with *Tauer*. Approaches to Multilingual Lexical Semantics." Papers in Computational Lexicography COMPLEX'99, ed. by Kiefer et al., 153- 169. Budapest: Linguistics Institute, Hungarian Academy of Sciences

Acknowledgement

The research reported in the paper was supported by Országos Tudományos Kutatási Alapprogramok (grant number T026091).