PILOT IMPLEMENTATION OF A BILINGUAL KNOWLEDGE BANK

Victor Sadler & Ronald Vendelmans
BSO/Research
P.O. Box 8348
NL-3503 RH Utrecht
The Netherlands

email: sadler@dlt1.uucp

Abstract:

A Bilingual Knowledge Bank is a syntactically and referentially structured pair of corpora, one being a translation of the other, in which translation units are cross-coded between the corpora. A pilot implementation is described for a corpus of some 20,000 words each in English, French and Esperanto which has been cross-coded between English and Esperanto and between Esperanto and French. The aim is to develop a corpus-based general-purpose knowledge source for applications in machine translation and computer-aided translation.

1. Introduction

Harris (1988) has called for a "hyper-bitext" tool for professional translators, a tool which would permit them easy on-line retrieval of bilingual equivalences, or "translation units", they have used in the past. The translator's previous output would be stored as hyper-text, with the parallel texts as far as possible aligned. A search for a given expression or term would thus display, for each occurrence in the corpus, a chunk of source language context together with the corresponding fragment in the target language.

At the same time, but independently, the authors and their colleagues at BSO/Research have been experimenting with bilingual corpora as a potential knowledge source for the Distributed Language Translation system (for an overview of this machine translation project, see Witkam 1988). They have argued that a bilingual corpus, appropriately structured, can largely replace conventional dictionaries (Sadler 1989: 133) and grammar rules (van Zuijlen 1989) in machine translation. The aim is to automate as far as possible the acquisition of the various types of knowledge required for machine translation – from monolingual knowledge of morphology, word classes, syntactic structures etc., through bilingual knowledge of lexical equivalences and translation syntax, to purely extra-linguistic knowledge-of-the-world – by structuring the evidence explicitly and implicitly available in human translations. The structured bilingual corpus is termed a "Bilingual Knowledge Bank", or BKB. It appears that the tools now under development for constructing a BKB may also provide the professional translator with a more sophisticated form of "hyper-bitext" than that envisaged by Harris.

2. Building a Bilingual Knowledge Bank

There are basically three steps involved in building a BKB structure. First, each language version must be structured syntactically if it is to serve as a source of (monolingual and contrastive) grammatical knowledge. Second, semantically equivalent units (translation units) must be identified and cross-linked between the two versions. Third, referential or conceptual links must be added to identify various types of deixis and co-reference. The process can be illustrated with the following English-French example from Harris (1988).

[1] *The board of PAC unanimously confirms the mandate.*
= Le conseil du PAC est unanime dans sa confirmation du mandat.

The Distributed Language Translation project has adopted dependency, rather than constituency, syntax (Schubert 1987; Maxwell & Schubert 1989), and figure 1 shows the dependency trees for this example, cross-coded for translation units (TUs). Each ellipse corresponds to a subtree. The basic TUs are dependency (sub)trees. Each of the seven subtrees which are directly identifiable as translation units has been assigned an identification number.

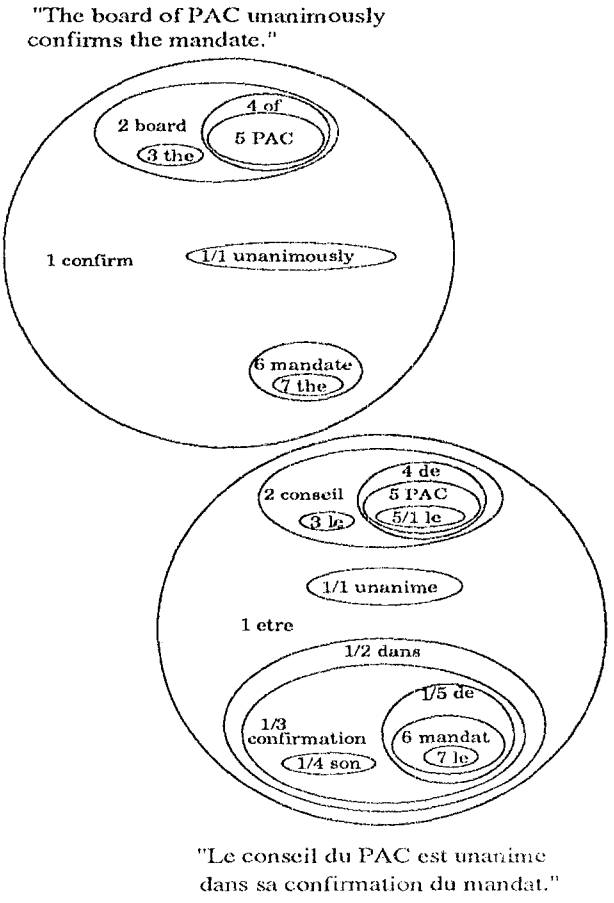

Figure 1: Dependency structures and translation units for example [1]

Table 1 lists the TU numbers with the corresponding equivalences. For example, TU 1 identifies the complete sentence, TU 2 is the subject noun phrase, 3 the determiner, 4 the prepositional phrase, etc. While each of the basic translation units corresponds to a (sub)tree, not every subtree corresponds to a translation unit. The French subtree governed by *dans*, for instance, does not constitute a translation unit. In the TU coding, this is shown by the identification "1/2" attached to *dans*, which indicates that this subtree is the second bound dependent in TU 1.

Table 1: Translation units identified in figure 1.

TUs	English phrase	French phrase
1	The ... mandate.	Le ... du mandat.
2	the board of PAC	le conseil du PAC
3	the	le
4	of PAC	du PAC
5	PAC	le PAC
6	the mandate	le mandat
7	the	le

The subtree approach to translation units allows for a process of tree subtraction which amounts to a kind of generalization. This allows the productive use of all the equivalences in the text, even if they do not constitute independent subtrees. For example, subtracting TUs 2 and 6 from TU 1 in figure 1 yields the equivalence of *to unanimously confirm* with *être unanime dans sa confirmation de*. In a machine translation application, TUs 2 and 6 can be thought of as variables in a productive translation rule. Table 2 lists the remaining possibilities and the corresponding subtractions. Once the basic TUs have been identified, these other equivalences can be automatically deduced by tree subtraction.

Table 2: Translation units derived from table 1.

TUs	English phrase	French phrase
1-2-6	unanimously confirm	être unanime dans sa confirmation de
2-3	board of PAC	conseil du PAC
2-4	the board	le conseil
2-3-4	board	conseil
4-5	of	de
6-7	mandate	mandat

The remaining step in BKB construction is the coding of references. In figure 1, TU 6 (*the mandate = le mandat*) will be linked by a pointer to its antecedent in a previous sentence. This link is bilingual, but other references may be language-specific. For example, the possessive pronoun in the French sentence has no correspondent in the English version, as shown by the coding "1/4" in figure 1. Nevertheless, a monolingual link must be established between *sa* (or its normalized form *son*) and the antecedent, which can be identified as unit 2 (*le conseil du PAC*).

Interconnecting the various surface forms used to refer to a given concept multiplies, for any given surface form, the contextual constraints which can be derived from the BKB, e.g. for the purposes of automatic disambiguation. It also allows the BKB structure to be regarded as a type of knowledge representation to which inference rules can be applied (Sadler 1989: 149-233).

The building of a Bilingual Knowledge Bank entails a great deal of interactive text processing. Even after the text in each language has been correctly parsed, the conversion of the parallel dependency trees to the BKB structure cannot be performed automatically. However, it does appear that a great deal of the work can become automatic. There are two reasons for this. First, the BKB itself can provide more and more support, in a kind of boot-strapping process, the larger it becomes. Second, the information contained in one language version can support the disambiguation of the other version.

3. The pilot implementation

In order to serve as a general provider of linguistic and world knowledge, a BKB should contain large amounts of data. When considering time-critical BKB applications, such as the BKB within a machine translation system, it is clear that efficient data storage techniques are needed. Of course, it is not possible to investigate BKB techniques on a very large scale at present, because it takes a relatively long time to process the corpus. For this reason a small-scale implementation has been designed which gives a good impression of a future large-scale BKB system. The basis for this pilot BKB is formed by three parallel 20,000-word text corpora in the field of computer manuals. From these corpora, two BKBs have been built: one for English/Esperanto, the other for French/Esperanto. The pilot implementation consists of three main parts: the parser, the "synsemizer" and the retrieval system.

The **parser** is used to parse each input text. Since each sentence which is stored in the BKB should have only one meaning (i.e., should contain no syntactic ambiguities), the parser yields only one analysis per sentence. This deterministic behaviour is produced by a simple category-based grammar on the one hand, and built-in mechanisms which take care of coordination, ellipsis and uncertain syntagma attachments on the other hand. The analysis found is presented graphically to the user, and can be edited as required before it is stored in the BKB. Words are stored in their normalized forms with categories and some basic syntactic features. The parsing process is BKB-supported: with each new sentence, the information that was stored earlier is used to give clues to categories, features and normalized forms. Besides this learning capability, a future BKB system will also use the structure of sentences already parsed to resolve attachment problems that the parser was unable to resolve.

The **synsemizer** is used both to define translation units by establishing bilingual relations between corresponding monolingual subtrees, and to establish monolingual referential relationships. The first part of the work is presented to the user graphically: the computer searches for probable TU constituents and displays them for the user's confirmation or correction. Subsequent proposals are influenced by the user's response. The system is self-improving, since the computer's guesses are based on the whole of the text processed so far. Referential relations must be

identified manually in this pilot implementation. However, since bilingual relations (TUs) have already been established before this process begins, there is additional information available to aid the operator.

The **retrieval system** is a tool which extracts information from a BKB that has been built using the parser and the synsemizer. On the basis of input phrases, which can be augmented with syntactic information, the BKB is queried. The resulting answers are presented to the user, either graphically or textually. Possible queries include concordance queries, translation and back-translation queries, and – to some extent – bridge translation (e.g. simulated English-to-French translation via Esperanto by "chaining" the two available BKBs).

An interesting aspect of this pilot implementation is that it is not just a simplified prototype system in which decisions about various difficult issues are postponed. On the contrary, it contains the required functionality for building a real large-scale BKB. Any weaknesses of the pilot system derive from its limited size and from inefficiencies in implementation, rather than from its functionality. The system can therefore be used for examining various extrapolation-directed aspects such as linguistic and technical applicability, consistency mechanisms and also user interface presentation at the BKB building stage.

4. Comparison with other research

The corpus-based approach to dictionary acquisition, which is part of the motivation behind the Bilingual Knowledge Bank, should not be confused with attempts made elsewhere to derive lexical equivalences from a bilingual corpus by purely probabilistic means (e.g. Brown *et al.* 1988). Syntactic structure is an essential BKB ingredient. Sumita & Tsutsumi (1988) have implemented a database of equivalent sentences in Japanese and English, but no full syntactic parsing is done, and retrieval is based on patterns of function words in the Japanese text. In their tool, sentences retrieved in bilingual form serve merely as models for the human translator. Another translation aid has been described and implemented by Kjærsgaard (1987). This system allows the translator to retrieve a key word from one half of a bilingual corpus, together with its context in the source language and the corresponding chunk of text in the target language. It is up to the user, however, to decide which, if any, is the equivalent expression in the target language chunk.

The closest comparable research appears to be that of Ogura *et al.* (1989), who have structured some 40,000 words of running text in Japanese and English in what they term a "linguistic database". This does comprise a hierarchical syntactic and text-level structure, as well as cross-references between equivalent expressions in the two languages, although it is not clear whether all translation units have been coded. Their primary aim is to provide a friendly interface for the linguist, answering queries on word-class statistics, displaying the context and translations of key expressions, etc. In contrast, the present research is directly primarily towards applications in machine translation.

5. Conclusions

As compared with traditional methods of lexicography and the writing of conventional grammar rules, this corpus-based approach takes advantage of the fact that vast amounts of human translation expertise are readily accessible in readable form. Instead of extracting vocabulary and grammar rules from text, the method described structures the text in such a way that the knowledge is directly accessible in the text itself. The BKB is a completely symmetrical construction, in which no distinction is made between source and target languages. The (virtual) dictionary and rule system it comprises are thus 100% reversible.

References

Brown, P. / J. Cocke / S. Della Pietra / V. Della Pietra / F. Jelinek / R. Mercer / P. Roossin (1988): A statistical approach to language translation. In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, pp. 71-76.

Harris, Brian (1988): Interlinear bitext. *Language Technology* Nov/Dec 1988, 10, p.12.

Kjærsgaard, Poul Søren (1987): REFTEX – A context-based translation aid. In: *Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, 1-3 Apr. 1987, pp. 109-112.

Maxwell, Dan / Klaus Schubert (eds.) (1989): *Metataxis in practice: Dependency syntax for multilingual machine translation.* Dordrecht/Providence: Foris. DLT 6.

Ogura, Kentaro / Kazuo Hashimoto / Tsuyoshi Morimoto (1989): Object-Oriented User Interface for a Linguistic Database. In: *Proceedings of the Working Conference on Data and Knowledge Base Integration*, University of Keele, 5-6 Oct. 1989.

Sadler, V. (1989): *Working with Analogical Semantics: Disambiguation Techniques in DLT.* Dordrecht/Providence: Foris. DLT 5.

Schubert, K. (1987): *Metataxis. Contrastive dependency syntax for machine translation.* Dordrecht/Providence: Foris. DLT 2.

Sumita, E. / Y. Tsutsumi (1988): A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching. In: *Proceedings Supplement, Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Carnegie Mellon University, Pittsburgh, 12-14 June 1988.

Witkam, Toon (1988): DLT – an industrial R & D project for multilingual MT. In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, pp. 756-759.

Zuijlen, J. van (1989): Probabilistic methods in dependency grammar parsing. In: *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, Pittsburgh, August 1989, pp. 142-151.