A.G. Sciarone & M.R. Ahmadi, Delft University of Technology

Towards a Universal Dictionary

Abstract

In this paper we demonstrate that the process of writing bilingual dictionaries can be organized in such a way that French-English or French-Dutch dictionaries, for example, may be generated automatically on the basis of English-Dutch and English-French dictionaries. To make this possible, we have built a set of word-meanings that can be expanded and modified, and a filtering mechanism to extract appropriate meanings for individual language pairs. In addition, we have implemented a computer program using a relational database management system to demonstrate a prototype of the "Universal Dictionary".

1. Introduction

Making a dictionary is an enormous and hence time-consuming and expensive task. Unfortunately, when this enormous task has been completed for a dictionary translating, for example, words of English into words of French, the result of those efforts cannot be used for translating the same words in the opposite direction, because languages differ in the ways they map meanings onto word forms. From this it follows that, depending on the language and on the direction into which words are to be translated, words may have to be divided differently into meanings (cf. Sciarone 1984; Marello 1989:19).

The semantic distinctions that need to be made for the words of a given language, depending on the language into which they are to be translated, are not really a matter of content – meanings are indeed more or less fixed entities – but of the number of distinctions that need to be made: pairs of words in different languages whose range of meanings are the same, require no more than a single unitary translation. Accordingly, when different bilingual dictionaries are made for the same source language, part of the necessary work of subdividing source-language words may have to be done more than once.

Similar inefficiencies occur with bilingual dictionaries that differ only in the direction of translation: from a particular foreign language into one's mother tongue, or vice versa. The problem here is that in translating from their mother tongue into a foreign language, users need more information to be able to choose the correct translation than they do the

other way round. With their mother tongue the target language, users can choose among the various translations on the basis of their native knowledge of the language, so that this knowledge need not be included in the dictionary.

In two earlier publications (Sciarone 1983, Sciarone 1984) it was argued that on the basis of a dictionary translating from the mother tongue into a foreign language, it should be possible to almost automatically produce a dictionary for the opposite direction, which would substantially reduce the work involved in dictionary making.

As far as we can see, the ideas then proposed have not been worked out by lexicographers or put into practice by publishers of dictionaries. A possible reason might be that in those days, lexicographers used computers almost exclusively as word processors – not as devices allowing them to rearrange large amounts of data.

In those days, lexicographers – rightly – emphasized differences between bilingual dictionaries which depended on the direction of translation, but were not aware that these differences need not conflict with the idea of a single computerised dictionary that can be organised in different ways (cf. Hartmann 1984). To illustrate this point, Sciarone (1984:415) presented the English entry to *appear*, with its Italian translations, as derived from an Italian-English dictionary.

In Sciarone (1984) it was argued, furthermore, that bilingual dictionaries could not only be inverted rather easily, but that new bilingual dictionaries (English-French, for example) could be derived from sets of bilingual dictionaries with common languages (English-Italian and Italian-French, for example). The point was illustrated with translations taken from the Sansoni Italian-German dictionary:

apparire	= (mostrarsi visibilmente) 'erscheinen'
apparire	= (risultare) 'sich ergeben'
comparire	= (apparire) 'erscheinen'
comparire	= (di pubblicazione) 'erscheinen'
comparire	= (in giudizio) 'erscheinen'
parere	= (sembrare) 'scheinen'
presentarsi	= (comparire) 'erscheinen'
sembrare	= (parere) 'den Eindruck haben, scheinen'
sembrare	= (ritenere) 'scheinen'

from these translations and those derived before, the following English-German list of translation equivalents can be derived automatically:

appear = 'erscheinen, sich ergeben, hervorgehen, etc.'

In the remainder of this paper, we investigate how a "universal" dictionary might best be organised so that new languages may be added freely and new bilingual dictionaries may be derived. First, we discuss what information the dictionary should contain and show that this comes close to what is found in a traditional dictionary. Second, we elaborate on how meanings may be subdivided, modified, or added. Third, we discuss a mechanism that chooses, among the various meanings which have been distinguished for the words of the various languages, those meanings that are relevant for a particular language pair. Finally, we briefly present some information on how the computer program has been organised.

2. Basic information

A bilingual dictionary of, say, English to French, typically contains the following basic information:

(1) source language word	leg
(2) target language word	jambe
(3) meaning description	lower part of living object
(4) example(s)	humans have two legs, dogs have four
(5) semantic field	part of body
(6) [language specific information]	[...]

Most of these items are unproblematic in the present context, except for (2) and (5), which concern the problems raised by the fact that source language words may be ambiguous relative to a particular target language, and the related problem of selecting, from among the various meanings in the universal dictionary, those meanings that are relevant for a particular language pair, using semantic field specifications similar to those found in a Roget-type thesaurus. The following three sections deal with these problems in more detail.

2.1 Subdivision into meanings

A central problem in dictionary making involves the division of words into distinct meanings, as can be seen from the fact that different dictionaries for a given language typically differ to some degree in the distinctions that are made in this regard.

Clearly, the making of such distinctions is not a simple technical problem that can be dealt with according to a fixed procedure (Van der Eijk et al. 1995), but also seems to depend on the ideas of individual

lexicographers about meanings. However, in order for dictionaries for specific language pairs to be derivable from other dictionaries with a common language, the meanings of this common language will have to be subdivided in the same way.

One way of securing such uniform divisions would be to establish a fixed set of meanings to be used by dictionary makers (cf. Eaton 1940). But since different languages subdivide words into meanings differently, and since we cannot foresee all possible meanings, it is impossible to set up such a set of meanings from the outset and to assign the corresponding words for each language. What we need is a system in which meanings can be distinguished freely. As a consequence, all meanings distinguished for the various languages included in the universal dictionary will be recorded in a list that serves as input for languages that may be added later.

Once meanings have been distinguished, they need to be described, so that lexicographers (and dictionary users) may know what meaning is meant. To this end, entries will include meaning descriptions and examples, both firmly rooted in lexicographic tradition. Both meaning description and examples may be changed, and they may even be different for different languages, as long as the relevant meaning doesn't change.

In order that a bilingual dictionary for a particular language pair may be used in either direction, meaning descriptions and examples should be given for both languages. This is also helpful when other languages are to be dealt with: once an English-French dictionary has been written in accordance with these principles, both English-speaking and French-speaking lexicographers may profit from the work done by the makers of that dictionary when they attempt to write new dictionaries for English-Italian, FrenchDutch, etc.

2.2 Adding or modifying meanings

Because different languages semantically partition reality in different ways, and because linguistic views on meaning may differ too, lexicographers should be allowed to freely add and modify meanings. When making a French-Dutch dictionary, for example, a lexicographer would need to assign two Dutch translation to the French word jambe 'leg': (1) *been*, and (2) *poot* – depending on whether 'jambe' refers to the lower limbs of a human being or to those of an animal. This means that in the existing version of a universal dictionary, the meaning description 'lower part of a living object' will be replaced by two new descriptions: 'lower

part of human being', and 'lower part of animal'. This new semantic distinction should then also be imported consistently – and automatically – into the other languages included in the universal dictionary. As a rule, however, distinctions thus imported into a given language will not affect translations into that language:

leg ('lower part of human being') => jambe
leg ('lower part of animal') => jambe.

For French-to-Dutch translations, a lexicographer should add a third Dutch translation for *jambe*: (3) *pijp* ('lower part of trousers') – a meaning which belongs to the semantic field of 'clothes'. Unlike modified meanings, of course, new meanings of this kind will have to be translated, manually, for all other languages already included in the dictionary.

2.3 Filtering meanings

Adding meanings on the basis of a particular language leads to distinguishing meanings for languages which do not distinguish them, or which have meanings with a wider or a different domain:

jambe ('lower part of animal') leg

Although specifying distinct meanings for languages which do not distinguish them does not lead to incorrect information, and although the number of different meanings will not grow excessively because many languages partition reality along similar lines (human beings resemble each other) specifying meanings for languages that do not distinguish them is not efficient. To avoid this, we can make use of the information included in semantic fields. When different meanings are translated by the same word, and when these meanings belong to the same semantic field, we are obviously dealing with a language that does not distinguish those meanings, so that we can filter them out. In the case of English to French, both meanings of *jambe* belong to the semantic field 'part of the body', which triggers the filter. However, when we add a meaning that belongs to a different semantic field (e.g. the field of 'clothes' for *jambe* meaning 'lower part of trousers') the filter will not be triggered even though this meaning is translated into English as 'leg', too.

3. The Universal Dictionary

With uniform semantic distinctions for all languages as a result of adding and modifying meanings along the lines outlined above, new bilingual dictionaries can be derived quite easily from existing ones and formatted to the needs of users by filtering out meanings that are not used in a particular language. As more and more languages and pairs of languages are dealt with in this way, the result will be a universal dictionary. From this universal dictionary we can not only derive bilingual dictionaries for arbitrary language pairs but also monolingual dictionaries. Indeed, a monolingual dictionary is not really different from a bilingual one: instead of meanings being translated with foreign words, they are described ('translated') with words and phrases of the same language, similar to those we find in the informal descriptions of meanings in bilingual dictionaries.

Making dictionaries along those lines has three major advantages:

1. it saves much time by avoiding work that has been done already;
2. it increases the quality of dictionaries, because meanings are identified in a systematic way and because dictionaries are the product of co-operative efforts;
3. it shortens the production time of dictionaries considerably.

4. The Program

The above ideas have been implemented in a computer program with the overall structure of a relational database management system.

The heart of the database is an extendible set of meanings based on all the languages included in the dictionary. For all languages, even for those that not distinguish certain meanings, the same set of meanings is used to point to the word(s) that translate those meanings. Unlike a traditional dictionary, the universal dictionary, by starting from the meanings, gives one-to-one translations. All languages are, in a sense, made equal, which results in a rather simple management system. The program, in rearranging the data, takes care of differences between languages, using the specially created filtering mechanism. The filtering power of the program may be changed by the lexicographer by changes to the semantic fields – by changes to the number and variety of fields that can be used, and/or by changes to the fields that are assigned to individual meanings.

The program also takes care of updating all the languages included in the dictionary when changes are made in one specific language. Depending on the authority granted to the lexicographer, changes will be temporary or permanent.

Figure 1 shows the introductory screen of the program, which prompts the lexicographer (or the dictionary user) to choose a pair of source and target languages. Figure 2 shows the screen on which the lexicographer would be working: source and target words are to be typed in the appropriate boxes, along with the other information required (see 2.0). The screen for the user of the dictionary is basically the same, except in that it does not allow for data to be changed or altered.

The program is a well-working prototype but will have to be extended with whatever further features may be required by lexicographers or by their publishers.

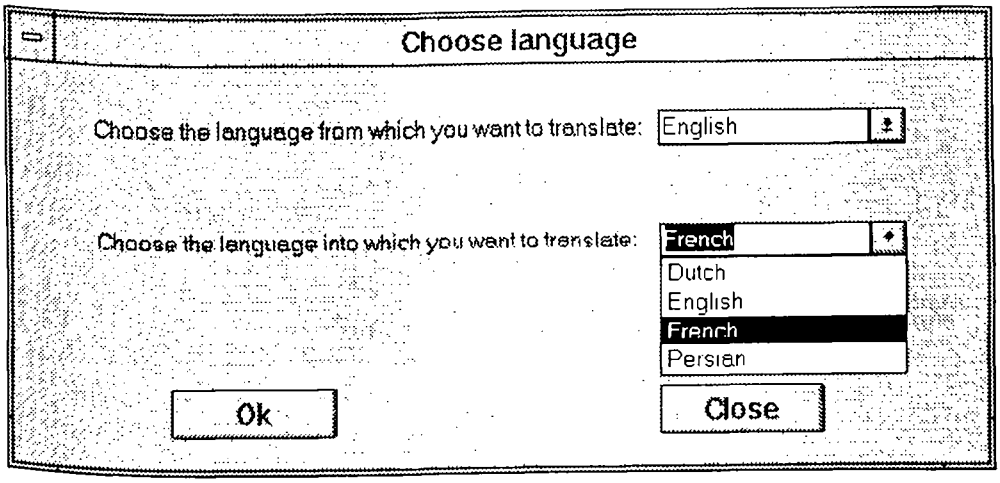

Figure 1: opening screen for the lexicographer.

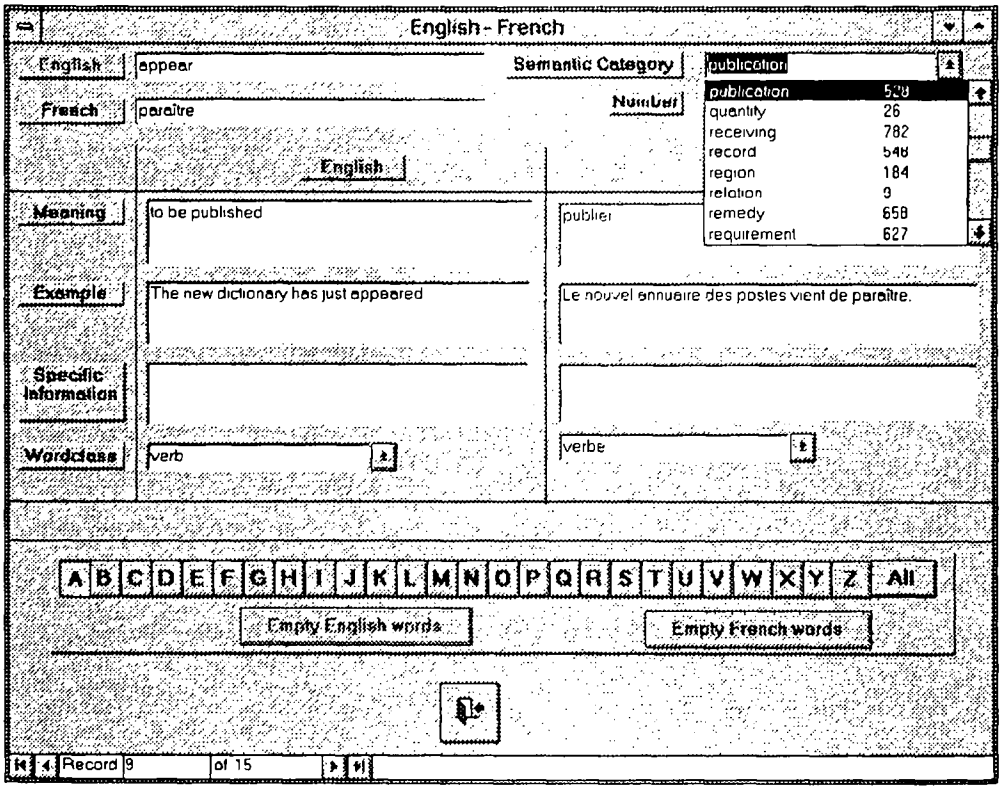

Figure 2: working screen for the lexicographer.

References

Eaton, Helen S. 1940. *Semantic Frequency List for English, French, German, and Spanish*, Chicago, Chicago University Presss.

Hartmann, R.R.K. 1985. "Contrastive text analysis and the search for equivalence in the bilingual dictionary" in: K. Hyldgaard-Jensen & A. Zettersten (eds), *Lexicographica Series Maior*, Volume 5, Tübingen, Max Niemeyer Verlag, pp. 121–132.

Marello, Carla 1989. *Dizionari bilingui*, Bologna, Zanichelli.

Sciarone, A.G. 1983. "Analyse sémantique: l'homme et la machine" in: *Lexique*, Lille, Presses universitaires de Lille, volume 2, pp. 133–146.

Sciarone, A.G. 1984. "The organization of the bilingual dictionary" in: R.R.K. Hartmann (ed) *Lexeter'83 Proceedings*, Tübingen, Max Niemeyer Verlag, pp. 413–419.

Van der Eijk, Pim; Alejandro, Olga; Florenza, Maria 1995. "Lexical semantics and lexicographic sense distinction" in: *International Journal of Lexicography*, Oxford, Oxford University Press, number 1, pp. 1–27.