Willy Martin, Free University of Amsterdam
Anne Tamm, Karolí Gaspar University, Budapest

OMBI: An editor for constructing reversible lexical databases

Abstract

In this paper OMBI (Dutch acronym for **Om**keerbare **B**ilinguale Lexicale Databases (= Reversible Bilingual Lexical Databases)) is presented and its approach, in particular with regard to the reversal function, is dealt with.

First a description of *OMBI's basic architecture*, aimed at genericity and flexibility is given. In this respect a distinction is made between three organizational levels: the UDS (Universal Deep Structure) of bilingual lexicons, the PDS (Product Specific Deep Structure) and the SUS (Surface Structure).

Thereafter *OMBI's main characteristics* are briefly mentioned (editing, genericity, import/export), to finally discuss the *reversal function and approach*. Here several parameters such as conceptual equivalence, pragmatic contrast, variant status and lexicalization status are used in order to link lexical units (not form units) from different languages to each other.

The end result should be a non-directional but *linkable* bilingual database from which databases and/or dictionaries in both directions can be automatically derived at a subsequent stage, and which can be used to be linked with languages outside of the original language pair.

1. Background

OMBI, an acronym based on the Dutch wordgroup '**Om**keerbare **Bi**linguale Lexicale Databanken' (= Reversible Bilingual Lexical Databases), is an editor which has been developed during the academic year 1994 – 1995 by the Dutch software house SERC (= Software Engineering Research Centre, Utrecht, The Netherlands) under the auspices of the CLVV – Committee (**C**ommissie voor **L**exicografische **V**ertaal**v**oorzieningen = Committee for Lexicographical Translation Resources). This Committee is an intergovernmental body of lexical experts set up in 1993 by the Ministry of Education and Science of both Flanders and the Netherlands in order to improve and stimulate the production of bilingual dictionaries and lexical databases with Dutch as a source or target language. The Committee has been given an initial budget of 2 Mi ECU (for the period March 1993 – February 1996) and has launched up till now several lexicographical projects which are, commercially speaking, non-viable, yet of great social relevance.

In this respect projects such as Turkish-Dutch v.v., Arabic-Dutch v.v., Hungarian-Dutch v.v., Polish-Dutch v.v., Italian-Dutch v.v. and Swedish-Dutch v.v. need to be mentioned.

However not only is it the Committee's task to have concrete products realized, but also to see to it that, if needed, adequate lexicographical tools and infrastructure are provided for.

The construction of OMBI is to be situated within this second domain, aiming at providing lexicographical teams with a generic and powerful editing tool.

In what follows we will point at the main characteristics of OMBI paying special attention to its reversal function. However in order to understand fully OMBI's characteristics it is necessary to deal with some of its architectural aspects as well.

In the section to follow therefore we will briefly mention some of these.

2. Some aspects of the OMBI-architecture

OMBI contains three levels:

- the UDS or universal deep structure
- the PDS or product specific deep structure
- the SUS or surface structure

With the UDS is meant those elements that all bilingual dictionaries (should) share and the relations that hold between them. So e.g. all bilingual dictionaries (should) have form units (FUs), lexical or meaning units (LUs) and example units (EUs)/combinations in two languages. These units are connected to each other, see fig. 1

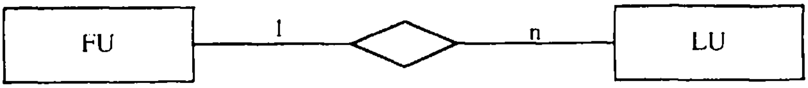

Fig. 1

meaning that

- corresponding to an LU there is exactly one FU
- one FU has at least one LU

Other relations state that an EU can be shared by more than one LU, see fig. 2.

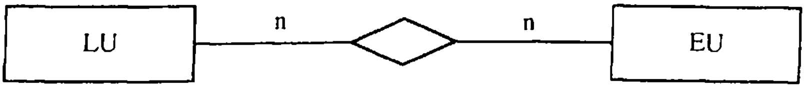

Fig. 2

So too it is stated that translation is a relation (a link) between two units from two different languages, of a specific nature, either an LU or an EU, see fig. 3.

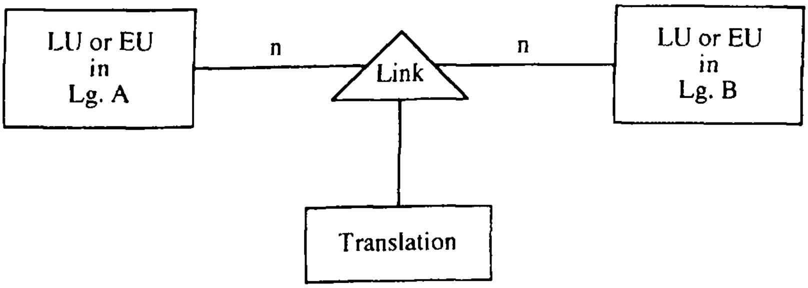

Fig. 3

In other words, the UDS describes a basic, yet fundamental, data model which is used for all OMBI-dictionaries, giving OMBI its generic character. Fig. 4 and 5, respectively, summarize the 'empty' data model and 'fill' or illustrate it.

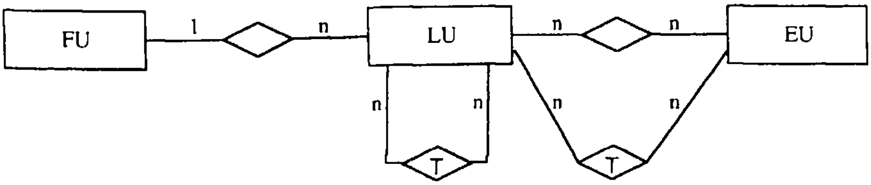

Fig. 4

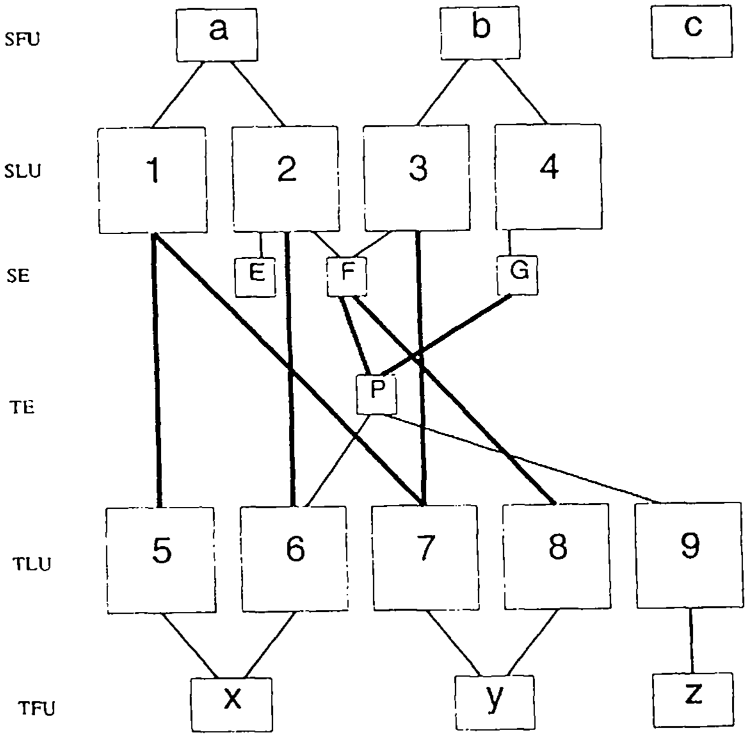

Fig. 5

In fig. 5 e.g. one finds E, F and G as EUs of the Source Language LUs 2, 3 and 4. P is an EU of the Target Language LUs 6 and 9, serving as a translation of the Source Language EUs F and G. F (an EU) is translated both by P (an EU) and by 8 (an LU) etc.

As to the PDS, it reflects the language and product specific features of the database, e.g. the categories and subcategories used to describe the language(s) in question. Whereas changing the basic UDS structure would require changing the basic structure of the program, the lexicographer can easily customize OMBI if he wants to add, change, delete, reorder etc. data categories. In other words, it is entirely up to the lexicographer which data categories and which values he wants to include in his database.

The SUS or surface structure, finally, reflects the organization of the final form of the output, e.g. a real dictionary. Through the main interface, the lexicographer can choose the fields he wants to print in the dictionary, and define the order of those fields. For example, in a productive/active dictionary one can choose to leave out all the flections and variants in the source language, and provide all the morphological and syntactic information and illustrations about the target language. In a decoding/passive dictionary, on the contrary, one can include the irregular flections and form/spelling variants in the source language with a reference to the base forms.

3. OMBI's main characteristics

OMBI's main characteristics basically come down to the following:

- it functions as an *editor*
- which is *generic*
- having *importing and exporting* facilities
- and the power to *reverse* lexical databases, trying to do so in *as accurate as possible* a way

As the last aspect is the most innovative one we will deal with it in a separate section. Before doing so we briefly mention the other characteristics as well.

3.1 Editing

First of all, OMBI has all the classical editing devices which structure, guide and correct the input of data, according to a pre-defined grammar for the lexical database. The benefits of these devices are clear: the input is efficient, and the data are consistent and structurally correct. As was stated before, the pre-defined grammar can be customized. As is usually the case the editing process is guided by making use of menu dialogues presented as screens with a set of possible actions to perform. To edit the maximal graphemics of a form unit (FU) e.g. the following menu/screen containing the fields will be opened.

- spelling
- spelling pragmatics (e.g. BE, AE etc.)
- spelling status (e.g. official)
- spelling type (e.g. full form, abbreviation etc.)
- hyphenation
- hyphenation pragmatics
- spelling variants
- form variants
- comments and
- illustrations

3.2 Genericity

A second, non-trivial function of OMBI resides in its genericity. As stated, OMBI is able to create databases on the basis of different database grammars, as long as they respect the fundamental architectural OMBI-principles (the UDS), such as the distinction between FU and LU. This makes OMBI a highly flexible and multifunctional tool, capable of being used in many different environments.

The implementation of this function is such that OMBI is delivered with a standard database grammar. This standard grammar can be varied in a simple way by the user, by adding or deleting data categories, changing finite lists, changing feature co-occurrences etc. More fundamental changes to the grammar have to be programmed on a lower level, in the underlying system Paradox. This is still a relatively simple operation, and does not require highly sophisticated programming skills.

The OMBI User Manual (see Wijne & van Elswijk 1995) contains a complete description of the OMBI-SGML Grammar. The notation used is BNF (Backus-Naur Form) a well-known standard for describing grammars. The top-most rules for e.g. form units, respectively lexical units, read:

```
form-unit       ::=  <FORM> spelling [ # wordcat]
                     form-field*
                     lexical-unit*
                     </FORM>

lexical-unit    ::=  <LU> resume
                     [ # syntactic-subcategory
                     [ # semantic-type]]
                     lu-field*
                     {translation/description}*
                     example-unit*
                     </LU>
```

3.3 Importing and exporting

OMBI's third function is that it can import into its database structure existing MRD's and databases, as long as they have certain minimal structural indications for the recognition of the basic units that OMBI works with. In addition, OMBI can export from its database different SGML-databases and/or dictionaries for specific purposes, by using transformation components. Along the lines mentioned above, a standard transformation component can be supplied to a number of standard export formats (dictionary models); this component can then be varied easily or changed more fundamentally, according to the needs of the user. Although we have not dealt yet with OMBI's reversal function, it should be clear that OMBI, when importing an A-B database, is able to export it as a B-A database. The result is of course strongly dependent on the degree of well-structuredness, formalization and OMBI-SGML-compatibility of the data to start from. Furthermore, it should go without saying that post-editing is necessary. The following, however, can give an idea of what one can expect. The starting-point is part of a new, not yet published, English-Portuguese dictionary (part from the letter A, publisher: Verbo Editorial)

abandon **1** <*n*> ***abandono*** *m* **with gay abandon** ***com desenvoltura*** **2**<*vtr*> ***abandonar*** [person, town, car, hope, etc] (**to** a); ***renunciar a*** [activity, claim, idea]; **3** <*v refl*> **to abandon oneself** ***abandonar-se***

abort **1** <*vtr*> a) [terminate] ***fazer abortar*** [foetus, embryo, pregnancy]; b) [interrupte] ***interromper*** [mission, plan]; c) *Comput* [abandon] ***abandonar*** [program, operation]; **2** <*vi*> a) [mother, embryo] ***abortar***; b) [plan, mission, attack] ***fracassar*** ; c) *Comput* [program] ***abortar***

This input/import yields a.o. the following Portuguese-English entries: **abandono, abandonar, renunciar, abandonar-se, fazer abortar** (!), **interromper, abortar, fracassar**.

abandonar <*vtr*> **1** [person, town, car, hope, etc] ***abandon*** [x]; **2** [program, operation] ***abort*** [abandon]

abandonar-se <*v refl*> **1** ***abandon*** [x]

abandono <*n*> **1** ***abandon*** [x]; ***abandonment*** [x] (**a.a**) *com desenvoltura* ***with gay abandon***

abortar <*vi*> **1** [mother, embryo] ***abort*** [x]; **2** [program] ***abort*** [x]

fazer abortar <*vtr*> **1** [foetus, embryo, pregnancy] ***abort*** [terminate]

fracassar <*vi*> **1** [plan, mission, attack] ***abort*** [x]

interromper <*vtr*> **1** [mission, plan] ***abort*** [interrupte]

renunciar a <*vtr*> **1** [activity, claim, idea] ***abandon*** [x]; **2** [throne] ***abdicate*** [give up]; **3** [rights, claims] ***abjure*** [x]

Because of the fact that OMBI expects a meaning description for each LU (in both source and target language), a dummy (x) is inserted when it was lacking. It was put immediately behind the translation equivalent. Selection restrictions were put between square brackets preceding the translation equivalent. As one can notice 'integration' of fragments has already taken place (see *abandonar* e.g.), yet not always in a correct way (see *com desenvoltura*)

4. The reversal function

Although with the advent of the reusability-concept (see e.g. Calzolari 1991) the notion of reversibility has also come to the fore (see e.g. P. van der Eijk e.a. 1992), editors which pay attention to this aspect in a non-trivial way are rather scarce or even non-existent. In what follows therefore we will present this function somewhat more in detail.

A fourth function, which makes OMBI more innovative than other editors, is that it can reverse translational relations and directional databases in general. While the editing function is busy creating a bilingual database X → Y, and as such taking in translations from X to Y, OMBI simultaneously stores the reversed counterparts, thereby building a reverse database Y → X. The end result is a non-directional bilingual database, from which databases and/or dictionaries in both directions can be automatically derived at a subsequent stage.

In order for the process and outcome of reversal to be non-trivial, the tool should not merely state that if word form x is a translation of word form y, then word form y is a translation of word form x. This is in many cases not only too limited a conclusion, but also a wrong one: only rarely is translation a straightforward symmetrical relation between word forms.

The first, highly important observation about translation relations is that it is not *words* that are translated into other words, but rather *words in a specific meaning*. The English word *horse* is a translation of the Dutch word *paard*, but only in the meaning of the latter as 'certain animal', not in its meaning 'certain chesspiece'. This insight has had a fundamental influence on the architecture of the databases that OMBI builds (see section 2). The database distinguishes between Form Units or FUs (word forms) and Lexical Units or LUs (meanings): every Form Unit (e.g. *horse*) can have one or more meanings (e.g. 1 'certain animal', 2 'certain chesspiece', etc.); only a LU (which always belongs to an accompanying FU) can be translated by a LU into another language.

The second important observation is that translation, and reversal of translation in particular, only holds if certain conditions are met. In OMBI the translation relation is analysed into four relevant parameters that influence reversibility, and which therefore have to be specified and taken into account in 'calculating' whether the reversal of the relation is valid or not. The four parameters are the following:

- conceptual equivalence
- pragmatic contrast
- variant status
- lexicalization status

Before entering into details for each of these parameters separately, a general remark may be in order: a useful view on the layout of the LU and the presentation of the components to be filled in by the lexicographer is that the LU has three different aspects:

- the monolingual information it contains
- the EUs (or combinations) it contains, with all there is to say about them
- the link it has to another unit in another language (LU or EU)

The different parameters used in this linking process can be characterized and illustrated as follows:

Conceptual equivalence

If x is translated as y, it is also necessary to specify whether or not the conceptual equivalence between x and y is complete. If it is incomplete, there are several sub-categories of 'non' or 'partial equivalence' which may be relevant; for instance that y is a hyperonym of x, or that y is a hyponym of x. In such cases, reversal of the translation relation x → y is permitted, but additional information is needed about the semantic restrictions/specifications of x or y. For example, the French words *fleuve* and *rivière* can both be translated into English by the hyperonym *river*; in the reversal, however, additional information is needed, so that *river* is translated as *fleuve* if the water flows into sea, and as *rivière* if it does not. At the moment OMBI distinguishes between the following values for this feature (conceptual equivalence):

- complete equivalence (i.e. there is complete conceptual equivalence between SLU and TLU)
- hyperonym (i.e. the TLU is a hyperonym of the SLU)
- hyponym (i.e. the TLU(s) is/are (a) hyponym(s) of the SLU)
- substitution by near equivalent
- related (e.g. the English series *shine*, *glimmer*, *glister*, *glow*, *glitter* etc. shows different degrees of overlap and partial equivalence with the Dutch 'counterparts' *schijnen*, *schitteren*, *flikkeren*, *glimmen*, *glinsteren*, *flonkeren*, etc.) (see Martin e.a. 1992)

These values group the complete and partial (conceptual) equivalents. Next to that there are also non-equivalents for which the values: 'description' and 'borrowing' have been reserved. Of course the several values can be extended, deleted, refined or otherwise modified. They entail with them a conceptual calculus. So e.g. there will be no reversal from A → B to B → A if the equivalent in B is a description or a borrowing. In other words, the reversal of the English LU *haggis* e.g., whether it is rendered as *haggis* in Dutch (borrowing), or by 'Scottish dish made from the heart and other organs of a sheep, cut up and boiled in a skin made from the sheep's stomach' (see LDOCE) (description), will be blocked in the Dutch-English lexical database.

So too the hyponym-link between the English LU *inflection* and the Dutch *vervoeging*, *verbuiging*, will be inversed into a hyperonym-link when reversal takes place. Moreover the semantic constraints in the hyponym-linking (viz. < w.r.t. verbs > and < w.r.t. nouns and adjectives>) will now be transformed into semantic specifications. Thus D. *vervoeging* = E. *inflection* < of verbs >.

Pragmatic contrast

Each LU must be accompanied by a specification of its pragmatic value. This information is part of the monolingual description. Via the database grammar, OMBI can be programmed in such a way that if the user tries to link two LUs x and y that have highly different pragmatic values in a translation relation x → y, the interface gives the user a warning signal, or even simply prevents such a relation from being stated. Along the same lines, the grammar could allow a translation relation x → y, but warn about or prevent the reversal y → x. For example, the obsolete Dutch expression *sponde* would be translated in English (with a warning concerning the pragmatic contrast) by the neutral, contemporary expression *bed*. The reversal of this translation must, however, be blocked.

At the moment the pragmatic component consists of the following subcomponents (and of course here too changes can apply): style, connotation, chronology, frequency, geography, subject field. Some of the values of these features are linearly ordered (such as style, chronology and frequency e.g.), others are value bound (connotation e.g. which is either (more/less) positive or (more/less) negative), still others are non-ordered (geography, subject-field).

The ordered and value-bound values can be arranged in groups, to which rules such as those below apply.

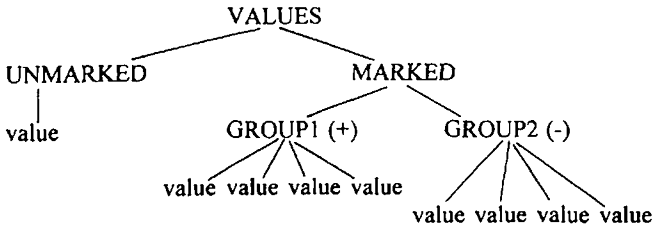

Fig. 6

Rule-conditions:

1. If the values are identical, rule (a) holds
2. If the values differ but are from the same group, rule (b) holds
3. If the values differ, and both are of the type 'marked' and both are from different groups, rule (d) holds
4. If the values differ in type, either rule (b), (c), or (d) holds: this is stated separately under the subcomponent.

Rules:

a. symmetrical link, no warning
b. symmetrical link, with warning
c. a-symmetrical link with warning: translation is possible, but reversion is blocked
d. no link allowed: warning

Example: Chronology

values: obsolete 1
obsolescent 2
contemporary 3
neologism 4

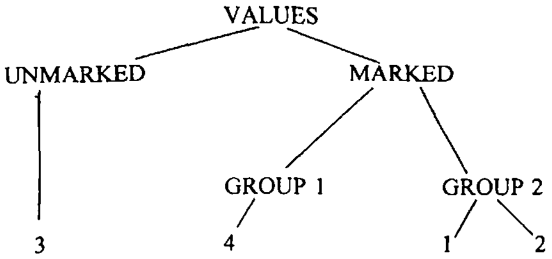

Fig.7

Extra rule: If the values differ in type, rule b holds.

Variant status

The fact that in a given situation y is the main or prototypical equivalent (see B. Defrancq, 1995) for x does not necessarily mean that the reverse is also the case. Therefore, this aspect of the translation relation has to be stated explicitly for the two directions x → y and y → x. So e.g. although E *car* gets two Dutch translations namely *auto* and *wagen*, both with a different variant status namely 'main' for *auto* and 'variant' for *wagen*, in reversing, the status of *car* for *wagen* will be 'main'. Furthermore, although the 'status' of a translation equivalent is for the greater part dependent on a.o. its pragmatics, its value cannot (always) completely be inferred from it, other factors, such as the mere existence of other alternatives, also playing a role.

Lexicalization status

Sometimes the degree of lexicalization of an expression and its translation differ. This is of course relevant information for the reversal. Items which are non-lexicalized in the target language (see under conceptual equivalence: 'description') will be blocked in the reversal process. On the other hand typical culture-bound items from the target language will never appear in the source language and therefore cannot be 'generated' by the reversal. As a rule therefore, the B → A database resulting from the A → B reversal, will always show gaps which will need to be filled.

5. Conclusions and Further Prospects

By making use of tools such as OMBI the idea of creating *linkable* lexical databases has been given shape. The advantages are obvious: while making an A → B dictionary/lexical database, a greater part of the B → A database is already created, reducing the amount of work drastically. After a testing period in which language pairs such as Dutch-Estonian and English-Portuguese have been used we can estimate the reduction of labour to be at least one third of the total workload (for the two databases).

On the other hand, it should be clear that, in order to work with OMBI properly, one needs to have a fairly thorough bilingual competence. Although OMBI has been developed with Dutch and Flemish government money this does not imply that it can only be used in projects with Dutch as a source and/or target language. On the contrary, at the moment we are investigating the possibility to use OMBI in establishing an efficient, high-quality, yet economically justified infrastructure for the African languages in South-Africa, where recently instead of two official languages (English and Afrikaans), eleven languages have been given the status of 'official language'. The fact that OMBI links at meaning level and calculates the possibility or impossibility of equivalents, makes it very suitable to function within what we have called elsewhere the 'hub-and-spoke' model, connecting (spoke) languages not directly but via a hub (see Martin, 1995a and Martin & Mashamaite, forthcoming).

References

Calzolari, N. 1991. "Representation of semantic information in Acquilex." In: *Feasability of standards for semantic description of lexical items*. Eurotra-7 Report, 31–42.

Eijk, P. van der, e.a. 1992. "Towards developing reusable NLP dictionaries." In: *Coling 1992 - Proceedings*, 53–54.

Defrancq, B. 1995. "The proto-equivalent." In: *Contragram* 3, 1995, 4–7.

Honselaar, W. and M. Elstrodt 1992. "The electronic conversion of a dictionary: from Dutch-Russian to Russian-Dutch." In: *Euralex-92-Proceedings*, 229–238.

Martin, W. 1995a. "Government Policy and Bilingual Lexical Databases: The Action Plan for Dutch." In: *Second Language Engineering Convention*, London, 133–144.

Martin, W. 1995b. "Maschinelle Lexikographie: Ein Blick in die Zukunft." In: Hitzenberger, L. (Hg.), *Sprache und Computer*, 1–21. Olms, Hildesheim.

Martin, W. e.a. 1992. *Multilex: Standards for the terminological description of lexical items*, Amsterdam.

Martin, W. and K. Mashamaite forthcoming. The hub-and-spoke model: a recipe for making dictionaries between African languages in South-Africa.

Wijne, C. and M. van Elswijk 1995. *OMBI: User Manual*, ms., The Hague.