Standardising bilingual lexical resources according to the Lexicon Markup Framework

Isa Maks, Carole Tiberius, Remco van Veenendaal

Dutch HLT agency (TST-centrale)

Institute for Dutch Lexicology (INL)

Matthias de Vrieshof 2-3

2311 BZ Leiden

E-mail: {maks,tiberius,veenendaal}@inl.nl

Abstract

The Dutch HLT agency for language and speech technology (known as TST-centrale) at the Institute for Dutch Lexicology is responsible for the maintenance, distribution and accessibility of (Dutch) digital language resources. In this paper we present a project which aims to standardise the format of a set of bilingual lexicons in order to make them available to potential users, to facilitate the exchange of data (among the resources and with other (monolingual) resources) and to enable reuse of these lexicons for NLP applications like machine translation and multilingual information retrieval. We pay special attention to the methods and tools we used and to some of the problematic issues we encountered during the conversion process. As these problems are mainly caused by the fact that the standard LMF model fails in representing the detailed semantic and pragmatic distinctions made in our bilingual data, we propose some modifications to the standard. In general, we think that a standard for lexicons should provide a model for bilingual lexicons that is able to represent all detailed and fine-grained translation information which is generally found in these types of lexicons.

1. Introduction

The set of bilingual lexical resources currently available at the Dutch HLT agency comprises Dutch-Arabic, Dutch-Portuguese, Dutch-Indonesian and Dutch-Danish. In the near future this set will be extended with lexicons for Dutch-Finnish, Dutch-Greek and Dutch-Estonian. These lexical resources were developed under the auspices of the CLVV (Committee for Interlingual Lexicographical Resources), an intergovernmental body of lexical experts installed in 1993 by the governments of the Netherlands and Flanders in order to improve and stimulate the production of bilingual dictionaries and lexical databases with Dutch as source or target language (Martin 2007:222).

The goal was to develop multifunctional and reusable electronic lexical databases. As such they contain information useful for different types of lexicons. For instance, information such as free text definitions, lexicographic comments and descriptions are mainly useful for human use, whereas semantic type, example types and complementation patterns are more useful for computational applications and information like lemma, word form, part of speech, pragmatic labels, collocations, idioms, translation equivalents can be used by both humans and computers.

The data were compiled using the dictionary tool OMBI (Omkeerbare Bilinguale Bestanden = Reversible Bilingual Lexical Databases) specifically designed for creating and editing rich multi-purpose bilingual resources (Maks 2007, Martin and Tamm 1996). One of the most distinctive features of the tool is the reversal of source and target language at sense level. All bilingual resources in the set available at the Dutch HLT Agency use the same Dutch component as a base, i.e. Referentiebestand Nederlands (Reference Database for Dutch, henceforth RBN (van der Vliet 2007)). The RBN is a lexical database based upon modern Dutch written corpora; it has a macrostructure of about 45.000 entries and a rich and explicit microstructure describing orthographical, morphological, syntactic, collocational, semantic and pragmatic features of Dutch lexemes (Martin 2007:230).

The compilation of these lexical databases started in 1998 and the first was completed in 2002. Several bilingual printed paper dictionaries have been derived from these resources and have been published since (for details see References and Table 1).

Project	Start	End	Publisher	Distributor
Dutch-Arabic vice versa	1997	2002	Bulaaq Amsterdam 2003	TST-centrale 2007
Dutch-Danish	1997	2001	Het Spectrium Utrecht 2004 Gyldendal Copenhagen 2004	TST-centrale 2008
Dutch-Estonian	1998			TST-centrale
Dutch-Finnish vice versa	2002			TST-centrale
Dutch-Indonesian	1997	2002	KILTV Leiden 2004	TST-centrale 2007
Dutch-New Greek vice versa	1998	2008	Het Spectrum Utrecht *to appear*	TST-centrale
Dutch-Portuguese vice versa	1998	2002	Het Spectrum Utrecht 2004 Verbo Lisbon 2004	TST-centrale 2008

Table 1: OMBI bilingual dictionary projects

Although the data themselves are particularly appropriate for use in computational applications such as machine translation, computer assisted translation, and cross-lingual retrieval, they have never been used as such until now. The main reason for this is that both lexicon model and technical format are heavily dependent on the OMBI-tool they were created with and therefore difficult to understand and reuse. Moreover, as the databases were developed in the late 1990s, they do not comply with current standards like Unicode or XML. The task of the Dutch HLT agency is to solve this by standardising the data and converting it to a more accessible format. Of course, the original data format will remain available too
The paper is structured as follows. First we discuss some of the characteristics of our bilingual data. Then we give an overview of the process we adopted to convert the bilingual resources into LMF (Lexical Markup Framework) conformant lexicons. Finally, we describe the LMF extension for bilingual lexicons that we propose.

2. The Bilingual Lexicons

One of the most distinctive features of the OMBI-bilingual LRs is that they contain information on translation equivalency with regard to both conceptual differences and usage differences between languages and differences in usage. They specify:

a) Difference in degree of conceptual equivalence between the two languages ranging from complete conceptual equivalent to partial conceptual equivalent (hypernym or hyponym equivalent) and near equivalent. See example (1) : *padi, beras* and *nasi* are partial translation equivalents of *rijst*. The target words refer to narrower concepts than the source word.
b) Lack of translation equivalent in the target language: if the source language concept does not exist in the target language, a description is given. e.g. the Dutch concept *chocoladeletter* (lit. chocolate letter) will be described as 'letter made of chocolate for the celebration of Saint Nicholas'.
c) Contrast in degree of lexicalisation status ranging from fully lexicalised to non-lexicalised. See example (1): the lexical collocation *kleffe rijst* (lit. sticky rice) is less standardised, i.e. semi-lexicalised, and less commonly used than its fully lexicalised Indonesian equivalent *babak.*
d) Preferred equivalents in the case of multiple equivalents. The ranking is indicated by the translation order number.
Differences in usage indicated by contrasting pragmatic labels. For instance, in an English-Dutch lexicon the word *bicycle* (neutral) can be translated as *fiets* (neutral), *rijwiel* (formal, oldfashioned).
The following example is based upon the Dutch-Indonesian lexical database and illustrates some of the possible translation contrasts:

Rijst [n] 1. [rijstplant of ongepelde rijstkorrels (rice plants or unhusked rice grains)] * *padi* ; [gepelde rijstkorrels (husked rice grains)] **beras* ; [gekookte/gestoomde rijst (cooked or steamed rice)] **nasi* ; rijst koken (cook rice) *memasak/menanak nasi;* kleffe rijst (sticky rice) ** *babak* , twee pakken rijst (two packs of rice) *dua pak beras*…..
Geluk [n] 1. [gunstig verloop van omstandigheden (favourable course of circumstances)]
(1)keberuntungan, (2)untung, (3)rezeki … **2.** …

(*) Partial conceptual equivalency
(**) Contrast in lexicalisation degree
(1,2,3)Translation equivalents ordered by usage

Ex. 1 : Extract from the lemmas *rijst (rice)* and *geluk (happiness)*

In section 4 we show how this rich information can be modeled in LMF.

3. Conversion into LMF

The Lexical Markup Framework (Francopoulo et al. 2006; ISO-24613:2007, still under development) provides a common model for the creation and use of lexicons. The framework provides specifications for monolingual and multilingual lexical resources. It consists of a core package which describes the basic hierarchy of information in a lexical entry and various extensions designed for the description of specific lexicons. In this paper we assume a basic knowledge of the LMF model and we will only focus on the parts that are relevant to the current discussion. For more details on the model in general the reader is referred to (ISO 24613: 2007).
The first step in the conversion process involved the definition of our model for an LMF conformant lexicon. This meant that we had to choose which extensions would be the most appropriate for our purpose. Closest to the type and content of our Lexical Resources (LRs) is the 'NLP multilingual notations extension' which – according to the LMF model - can be used for bilingual and multilingual resources. In Section 4 of this paper we present this issue in more detail.
For interchange and interoperability the use of uniform data category names and definitions is necessary. Therefore, we had to translate the original data category names into correspondent ISO 12620 data category names.
This involved the decomposition of values in minimal units, explicitation of implicit values (e.g. null value for part of speech -> noun), and translation of data category names (e.g. description -> explanation). Only in few cases the data category definitions did not match between the source and target data: for example, the source data distinguishes between lexical and grammatical collocations whereas the ISO standard does not further subdivide the class of collocations.
Finally, once we had decided on our LMF model, the data could be converted. As we are dealing with a variety of languages using different encodings we first adopted

UTF-8 as our standard file format[1] (conform to LMF specifications). This facilitated comparison of the resources in different phases of the conversion process by different people. The whole conversion process comprised two steps which were performed using one set of Perl scripts for all language combinations. First, the SGML output generated by the OMBI database was converted into valid XML correcting structural errors in the input data (mainly by adding closing tags as needed). During this process, an effort has been made to keep a link with the original dataset by preserving the ids from the source code in the conversion process[2]. On the one hand, this allows to check and correct possible errors afterwards, on the other, keeping a clear link between the two files language A to language B opens up the perspective of using the data in a multilingual resource with the RBN as a pivot. In a second step, the resulting XML was converted into LMF.

We use XSLT stylesheets to display the lexical information on the screen, very much like the original OMBI-rtf output which looks like a printed translation dictionary.

4. Representing Bilingual Lexicons in LMF

The LMF-standard offers two different approaches for the description of bilingual lexical data: the Machine Readable Dictionary (MRD) extension and the NLP multilingual notations extension. Below we give a short description of each of those extensions followed by a discussion of the extension for bilingual lexicons that we propose.

4.1. LMF's extension for Machine Readable Dictionaries

The MRD extension is designed to represent monolingual and bilingual data which are compiled in the first place for human use. The data are represented in a source to target language format. However, the MRD extension only takes into account complete translation equivalents and does not account for other than complete translation equivalents. Therefore, it is not fine-grained enough with regard to the description of the translation equivalents for our purposes.

4.2. LMF's extension for NLP multilingual lexicons

The NLP multilingual notation (NLP-multilingual) extension focuses on the representation of equivalents of two or more languages and special attention is paid to the representation of partial vs. full conceptual equivalence (see fig. 2). The extension introduces a pivot approach creating a kind of interlingua by using notions like Sense Axis, Sense Axis Relation and Transfer Axis. The Sense Axis and Sense Axis Relation deal with semantic contrasts between the languages; the Transfer Axis deals with differences in syntactic features. As our data focus on semantic differences , we do not take the Transfer into consideration.

Figure 2 (for more details see ISO/TC 37/SC 4 Rev.15: p. 52) represents the partial equivalence relation between the French *fleuve* ('river that flows into the sea') and the English *river* ('stream of water that flows into the sea, or in another stream of water'). The sense axes represent an language independent concept or 'sense' which link the French and English word senses. As complete equivalent senses share the same sense axis, SA2 is linked to both the French *rivière* and the English *river*; SA1 is not linked directly to an English sense because of the lack of a fully equivalent concept in English.

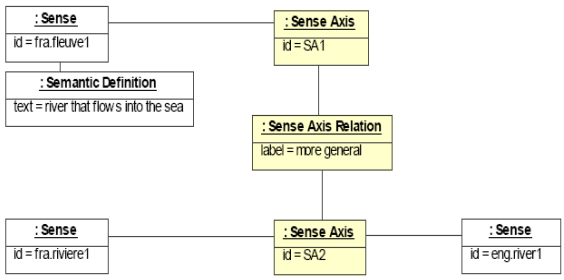


Fig. 2 : representation of partial translation equivalency according to the LMF multilingual extension

The notion of an interlingua is necessary to prevent an explosion of links when trying to link multiple languages. However, strictly speaking, an interlingua is not needed to describe two languages, but only to represent complex links in a multilingual lexicon. The decision to consider bilingual lexicons as "the simplest configuration of the multilingual extension" and by consequence to introduce an interlingua for the description of bilingual data has important implications:

- For each source-target word pair at least one sense axis is created; in the case of bilingual dictionaries – which do not need an interlingua – these axes do not have any use and form only empty elements.
- The use of an interlingua implies the distribution of the data among different files: one for the complete description of the source language, one for the complete description of the target language and one for the interlingual data. Cross references from the source language sense units to the language independent sense axis units and then to the target language sense units combine the data of the different files. By consequence, tools and procedures are needed to produce source to target language oriented versions of the data.

A more fundamental point is that the kind of contrasts which can be described within the language independent interlingua are contrasts of a conceptual nature only. The sense axes and sense axis relations refer to language independent concepts and are not suitable for representing the language dependent contrasts which may exist between two particular languages. So it is not clear how differences with regard to pragmatics, to the degree of lexicalization status, or to the notion of preferred equivalents, which typically exist between two specific languages and which are present in the source data of our

[1] Files were converted using the iconv tool (see http://ftp.gnu.org/pub/gnu/libiconv/libiconv-1.11.tar.gz).

[2] In the original SGML output generated by the OMBI database, this information was not preserved.

bilingual LRs, might be expressed within the multilingual extension.

4.3. A Compromise model

Therefore, our LMF compliant model is a compromise of the MRD and the multilingual extensions. It is illustrated in figure 2.

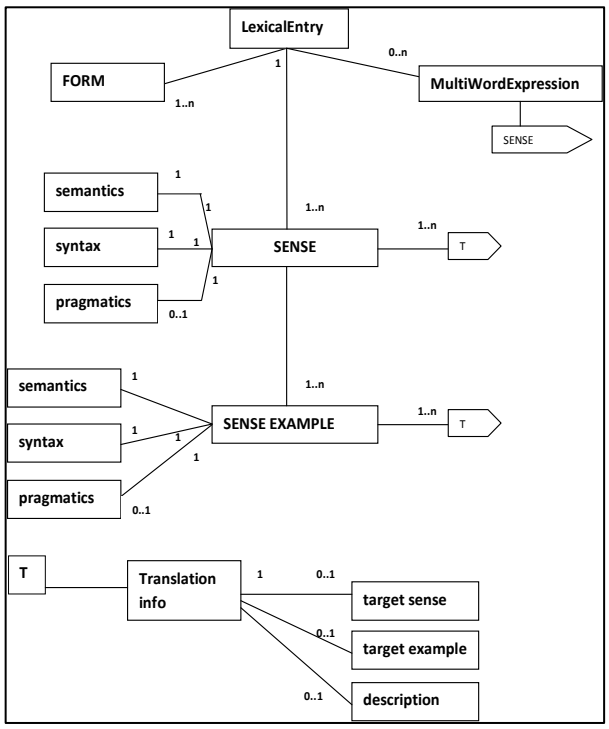
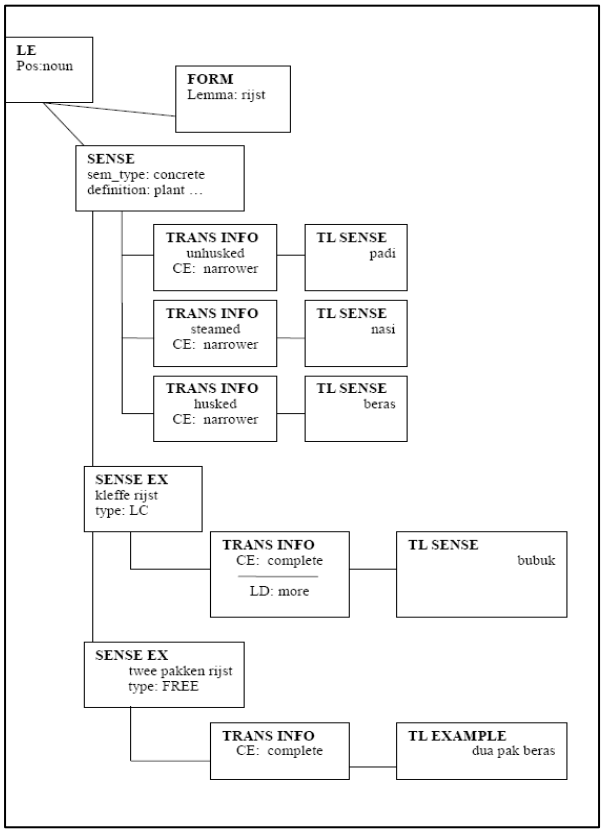


Fig. 2 OMBI-LMF model

First, there is no use of an interlingua. By introducing an interlingua the data structure becomes unnecessarily complex. As the accessibility of the data is one of our priorities, we decided against an interlingua.

Second, the representations are source target oriented. We believe that source target language oriented versions of the data are easy to understand and therefore easy to use not only for humans but also with respect to computational applications like machine translation and computer aided translation. Therefore, our representation of the data is oriented from source to target language and is stored in one XML-file. It contains a full description of source language A and a limited description of the target language B. At form, sense and example level ids refer to the XML-file which includes the complementary file, i.e. the bilingual lexicon oriented from source language B to target language A.

We introduced the use of translation information units which express both conceptual differences (similar to the contrasts by the sense axis relations) and language bound usage contrasts. The following figure shows that these units relate the source language sense and source language examples directly to its equivalent target language sense and/or example.



CE = conceptual equivalence
LC = lexical collocation
LE = lexical entry
LD = degree of lexicalisation
TL = target language

Fig. 3 representation of translation information in OMBI-LMF Bilingual Extension

We believe that the LMF standard should be more explicit about the representation of language dependant usage contrasts. In contrast to multilingual data, most existing bilingual lexical data contain rather subtle and fine-grained information about the degree of translation equivalence between the words of the involved languages and we feel that it is important to be able to represent this information in a standardised way. If these contrasts cannot be included in one of the current LMF extensions, this is an indication that a separate extension for bilingual NLP lexicons is needed.

5. Metadata

It is clear that conversion to a standard only contributes to the technical aspects of the availability of the data; the findability, identification and linguistic characterisation are of equal importance. Therefore good metadata are crucial. Especially, within the framework of the Dutch HLT agency, metadata are essential for archiving, distribution and maintenance of the data. The metadata that we use are based on a combination of current standards (IMDI, OLAC, DC) as none of the available proposed standards seemed to provide an unequivocal treatment of metadata at the levels we are interested in (i.e. archiving, identification and linguistic characterisation).

In our opinion, the available metadata are not geared specifically enough to the identification and linguistic characterisation of bilingual lexical resources. For instance, IMDI does not distinguish between monolingual, bilingual and multilingual lexica because it claims that bilingual and multilingual lexica can be broken down into monolingual lexica. Consequently, the possibility to indicate that the bilingual resources contained a lot of information on translation equivalence between the languages, was missing and has been added to our metadata. Figure 4 shows an extract of the metadata for the Dutch Arabic LR.

Fig. 4 metadata bilingual lexicon

6. Conclusion

In this paper, we discussed some of the issues involved in the standardisation of a set of bilingual lexical resources according to LMF. To be able to represent the fine-grained semantic and pragmatic distinctions available in our source data (degree of translation equivalence, preferred equivalents, etc.) we proposed an LMF extension specifically for bilingual lexicons. The model is a compromise of the LMF extensions for MRDs and for NLP multilingual lexicons. Due to the specific nature of monolingual, bilingual and multilingual lexical resources, we believe that it is important for standardisation efforts to make a distinction between them and to provide separate standards.

7. References

A. Dictionaries

Prisma Groot Woordenboek Nederlands-Deens. G. Laureys (ed.), Gyldendal and Het Spectrum, Copenhagen and Utrecht (2003)

Nederlands-Indonesisch Woordenboek Susi Moeimam and Hein Steinhouwer, KILTV, Leiden (2004).

Prisma Groot Woordenboek Nederlands-Portugees/ Portugees-Nederlands. M.C. Augusto (ed.), Het Spectrum and Verbo, Utrecht and Lisbon (2004).

Vertaalwoordenboek Nederlands-Arabisch/Arabisch Nederlands. Jan Hoogland (ed.), Bulaaq, Amsterdam, (2003)

B. Other literature

Francopoulu, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet and C. Soria (2006) *Lexical Markup Framework* Proceedings of the fifth International Conference on Language Resources and Evaluation, Genova.

ISLE Meta Data Initiative http://www.mpi.nl/IMDI/

ISO 12620: 2003 Language Resource Management – Data Categories – Data Category Selection for Electronic Lexical Resources (draft version), ISO Switzerland.

ISO 24613: 2007 Language Resource Management – Lexical Markup Framework (draft version), ISO Switzerland.

Maks, I. 2007. 'OMBI: The practice of Reversing Dictionaries'. *International Journal of Lexicography* 20.3.

Martin, W. 2007. 'Government Policy and the planning and production of bilingual dictionaries: the 'Dutch' approach as a case in point'. *International Journal of Lexicography* 20.3.

Martin, W. and Tamm A. 1996. 'OMBI: an editor for Constructing Reversible Lexical Databases' in M. Gellerstam et al. (eds.), *Euralex '96 Proceeding I-II,* Goteborg University.

OLAC Open Language Archives Community http://www.language-archives.org/

Vliet, H. van der. 2007. 'The Referentiebestand Nederlands as a Multipurpose Lexical Database'. *International Journal of Lexicography* 20.3.