OMBI bilingual lexical resources: Arabic-Dutch / Dutch-Arabic

Carole Tiberius, Anna Aalstein, Instituut voor Nederlandse Lexicologie
Jan Hoogland, Nederlands Instituut in Marokko (NIMAR)

In this paper we present the OMBI reversible bilingual lexical resources for Dutch-Arabic and Arabic-Dutch. These bilingual resources have been derived from a bilingual lexical database which has originally been produced with OMBI, a special tool for creating and editing bilingual dictionaries. Printed dictionaries have been published on the basis of this database (Hoogland et al. 2003) and now the data has been converted to LMF (Maks et al. 2008) to ensure future interchangeability and interoperability.

1. Introduction

OMBI-Arabic-Dutch and OMBI-Dutch-Arabic are bilingual lexical resources which are available from the Dutch HLT Agency for language and speech technology (known as TST-Centrale) at the INL (Instituut voor Nederlandse Lexicologie).[1] These bilingual resources were originally compiled within the framework of the project "Woordenboek Nederlands-Arabisch, Arabisch-Nederlands, Nijmegen"[2] in the period of 1998 till 2002 at the Radboud University of Nijmegen. This project was part of a large government initiative in the Netherlands and Flanders in the 1990s aimed at improving and stimulating the production of bilingual dictionaries and lexical databases with Dutch as source or target language (Martin 2007:222). The goal of this initiative was to develop multifunctional and reusable electronic lexical databases. In total, 13 dictionary projects have been completed and 22 volumes have been produced. Among those, the printed dictionaries for Arabic and Dutch (Hoogland et al 2003) which have been published in 2003 by Bulaaq, Amsterdam. A few projects are still ongoing.

2. The OMBI database editor

Most of the data of these bilingual lexicographic projects, including the data of Arabic-Dutch and Dutch-Arabic, were compiled using the dictionary tool OMBI (Omkeerbare Bilinguale Bestanden = Reversible Bilingual Lexical Databases) which was specifically designed for creating and editing rich multi-purpose bilingual resources (Maks 2007, Martin and Tamm 1996). One of the most distinctive features of this tool is the reversal of source and target language at sense level. Thus from the same lexical database, two bilingual dictionaries can be derived.

The OMBI bilingual lexical resources have a rich information structure. They contain information on lemma, word form, part of speech, pragmatic labels, collocations, idioms, free text definitions, lexicographic comments, descriptions, but also information on semantic type, example types, complementation patterns, to name just a few elements. Furthermore, there is detailed information on translation equivalency with regard to both conceptual and usage differences. As such, these resources are, apart from being used for printed dictionaries, also particularly appropriate for use in computational applications such as machine translation, computer-assisted translation, cross-lingual information retrieval, and information technology in general.
The downside of the OMBI lexical databases is that they do not comply with current standards like unicode or XML. Therefore, the data from the OMBI databases are currently

[1] http://www.inl.nl

[2] The editorial committee of the Nijmegen Arabic Dictionaries consisted of three arabists: Kees Versteegh, Manfred Woidich and Jan Hoogland. See also http://www.let.kun.nl/wba/

being converted at the Dutch HLT Agency from their SGML-format into unicode and an XML conformant to the Lexical Markup Framework (ISO 24613 2008) and ISOcat[3] specifications, assuring future interchangeability and interoperability of the data.

Although the databases for the different languages have in principle the same set up, each database has its own peculiarities due to the specific nature of each language. Below we describe the conversion process for OMBI-Dutch-Arabic and OMBI-Arabic-Dutch and focus on the specific characteristics of the OMBI database for Arabic.

3. OMBI-Arabic-Dutch and OMBI-Dutch-Arabic

3.1. The database

As mentioned, the OMBI database for Arabic and Dutch was originally compiled within the framework of the project "Woordenboek Nederlands-Arabisch, Arabisch-Nederlands, Nijmegen" in the period of 1998 till 2002 at the Radboud University of Nijmegen. It has been constructed using a text corpus and related tools (frequency count, concordancy programme). The corpus consisted of texts from various sources and fields, both fiction and non-fiction literature.

The project was set up to produce a dictionary that could both be used for text understanding and text production. The focus is on Modern Standard Arabic. There is no information on pronunciation since pronunciation can be directly determined by the spelling of a word.

Specific attention has been paid to collocations and examples to provide usage information. The grammatical behaviour of function words (demonstratives, adverbs etc.) is extensively illustrated in example sentences and expressions. Furthermore, a detailed distinction is made between various meanings of polysemic words (using synonyms or semantic field labels to define the different meanings). Finally, unpredictable grammatical information is entered for all different parts of speech (e.g. for verbs stem I: imperfect vowel, masdar; for nouns: broken plurals, diptotic plurals, gender if not clear from the external form of the word; for adjectives: broken plurals, irregular feminine forms, diptotism etc.). The table below gives an overview of the data in OMBI-Arabic-Dutch and OMBI-Dutch-Arabic:

	OMBI-Arabic-Dutch	**OMBI-Dutch-Arabic**
Lexical entries	24237	37706
Meanings	31051	44024
Translation equivalents	46920	59650
Examples	37973	48501
Examples translation equivalents	37919	35261
Descriptions	14318	29487
Idioms	74	402

Table 1. Overview of data in the bilingual lexical resources

3.2. Conversion of the data

The data in an OMBI lexical database can be exported as an SGML file for each language pair. Thus from the OMBI database for Arabic and Dutch, two files were exported, i.e. OMBI-Dutch-Arabic and OMBI-Arabic-Dutch. The original OMBI export facility was slightly changed as an effort was made to keep a link with the original dataset by preserving the unique identifiers from the source code in the conversion process. In the SGML output generated by the original OMBI database, this information was not preserved.

[3] http://www.isocat.org/index.html

The resulting files (including the unique IDs) formed the input for the conversion process. As OMBI was not unicode compliant, the original (Windows Arabic) character encoding was first converted into UTF-8. The unicode files were then converted into XML-LMF using a set of Perl scripts (Maks et al. 2008). First, the SGML format was converted to XML, using minimal processing. This means that the implicit structure of the data was not made explicit at this stage, but was left untouched in a 'relatively flat XML structure'. Next, this XML structure has been interpreted and converted into a more structured XML-LMF format. Below we illustrate this with the entry for ٱئْتِلافيّ 'coalition'. Figure 1 shows the XML structure of the entry.

```
- <FORM>
    ٱئْتِلافيّ # adj
  - <GRAPHEMICS>
      <ROOT>ألف</ROOT>
      <ROOTORDER>13</ROOTORDER>
      <HYPHENATION>ٱئْتِلافيّ</HYPHENATION>
    </GRAPHEMICS>
  - <MORPHOLOGY>
      <CODE>1</CODE>
    </MORPHOLOGY>
  - <LEXICAL_UNIT>
      pol
    - <SYNTAX>
        <SUBCAT>"(syntactic subcat)"</SUBCAT>
      - <SUBCAT_PRAG>
          <STYLE>formal</STYLE>
        </SUBCAT_PRAG>
      </SYNTAX>
    - <SEMANTICS>
      - <PRAGMATICS>
          <STYLE>formal</STYLE>
        </PRAGMATICS>
      </SEMANTICS>
      <DESCRIPTION>coalitie-</DESCRIPTION>
    - <EXAMPLE>
        "حُكومة ٱئْتِلافيّة"
        <NON_LEX />
        <TYPE>free</TYPE>
      - <TRANSLATE>
        - <FORM>
            coalitieregering # noun
            <LU>"regering door coalitiepartijen"</LU>
          </FORM>
          <CONC_EQUI>"complete equivalent"</CONC_EQUI>
        </TRANSLATE>
      - <TRANSLATE>
        - <FORM>
            coalitiekabinet # noun
            <LU>"kabinet van meerdere partijen"</LU>
          </FORM>
          <CONC_EQUI>unmarkAN</CONC_EQUI>
        </TRANSLATE>
      - <SHARED>
        - <FORM>
            حُكومة # noun
            <LU>"مَجْلِس الوُزَراء"</LU>
          </FORM>
        </SHARED>
      </EXAMPLE>
    </LEXICAL_UNIT>
  </FORM>
</xsl:stylesheet>
```

Figure 1. XML format of the entry for ٱئْتِلافيّ 'coalition'

This entry has been slightly edited to make it fit (ID numbers have been omitted). Figure 2 presents the resulting entry in LMF.

```
<LexicalEntry LE-id="7" LE-homonymnr="1" sy-pos="adj">
 <LE-admin osrcfuid="119415"></LE-admin>
 <Form-A>
  <LemmatisedForm-A writtenForm="ائْتِلافِيّ"></LemmatisedForm-A>
  <ReferredRoot writtenForm="ألف" rootOrder="13"></ReferredRoot>
  <Morpho-syntax-A mor-type="" mor-code="1" mor-diptotic="" mor-comparisonType="" mor-declinability="" flec-flectionalType="" sy-adverbialUsage="" sy-gender="" sy-
  position=""></Morpho-syntax-A>
 </Form-A>
 <Sense-A S-id="14" S-seqnr="1" sem-gloss="pol">
  <S-admin o-srcluid="119416" o-srcfuid="119415"></S-admin>
  <Semantics sem-type="" sem-shift=""></Semantics>
  <Syntax-A>
   <Sy-complementation></Sy-complementation>
  </Syntax-A>
  <Pragmatics prag-connotation="" prag-geography="" prag-subjectField="" prag-style="formal" prag-origin="" prag-socGroup="" prag-chronology=""></Pragmatics>
  <Translations-Sense>
   <Description Descr-id="18" description="coalitie-">
    <Descr-admin o-did="638094"></Descr-admin>
   </Description>
  </Translations-Sense>
  <Examples>
   <Example Ex-id="19" Ex-seqnr="1" canonicalForm="حُكومة ائْتِلافِيّة" textualform="">
    <Ex-admin o-srcexid="119498" o-srcluid="119416" o-srcfuid="119415"></Ex-admin>
    <Syntax-Ex sy-category="" sy-type="free"></Syntax-Ex>
    <Semantics-Ex exDefinition=""></Semantics-Ex>
    <Pragmatics prag-connotation="" prag-geography="" prag-subjectField="" prag-style="" prag-origin="" prag-socGroup="" prag-chronology=""></Pragmatics>
    <Translations-Ex>
     <Translation-Ex TrEx-id="6" TrEx-seqnr="1" TrEx-equivalent="coalitieregering" TrEx-pos="noun" TrEx-degreeOfEquivalence="complete equivalent">
      <TrEx-admin o-transid="119499" o-tarfuid="119415" o-tarluid="119416" o-tarexid="119498" tr-gloss="regering door coalitiepartijen" tr-form="coalitieregering" tr-
      pos="noun"></TrEx-admin>
     </Translation-Ex>
     <CrossRef crossRef-form="حُكومة" crossRef-pos="noun" crossRef-comment="مَجْلِس الوُزَراء"></CrossRef>
    </Translations-Ex>
   </Example>
  </Examples>
 </Sense-A>
</LexicalEntry>
```

Figure 2. LMF format of the entry for ائْتِلافِيّ ‘coalition’

In order to avoid redundancy in the data, not all information elements from the original database are processed in both directions, i.e. Arabic-Dutch and Dutch-Arabic. This is similar to what has been done for the printed version of the dictionaries. It concerns information elements that are marked respectively by ‘unmarkAN’ (to be ignored in Arabic-Dutch) and ‘unmarkNA’ (not to be included in Dutch-Arabic) in the database. For instance, in the entry above, the informal Dutch term, *coalitiekabinet* (‘coalition cabinet’) which is marked as ‘unmarkAN’ in the XML does not occur in the Arabic-Dutch part (see Figure 2).

In both OMBI-Dutch-Arabic and OMBI-Arabic-Dutch the lexical entries are ordered alphabetically, the same order as in the SGML export from the database. However, in the printed version of the dictionary only the lexical entries of Dutch-Arabic are ordered alphabetically. The lexical entries of the printed version of Arabic-Dutch are ordered under a particular root (the value in the attribute ‘ReferredRoot’). For instance, *kataba, maktab, kitab* are all listed under the root *KTB*. Within a root, the order is determined by the value in the attribute ‘rootOrder’. It is possible to generate an ordering according to root on the basis of the values in ‘ReferredRoot’ and ‘rootOrder’. However, this falls outside the scope of the responsibilities of the HLT Agency, as the goal is not to create a digital copy of the printed resources but to ensure interchangeability and interoperability of the data in the future.

Another distinctive feature of Arabic is that adjectives and nouns can be diptotic, meaning that they get two endings instead of the usual three. In the data a special element was introduced to capture this fact, i.e. the element ‘mor-diptotic’. This element has been added to the conversion scripts for Arabic-Dutch. The adjective ائْتِلافِيّ is not diptotic and thus the value of mor-diptotic is empty.

The above entry shows another interesting element, i.e. the shared examples. These are examples which consist of two or more words, and as such occur under more than one lexical entry in the dictionary. They are cross-referenced in the data. The conversion scripts were extended to include this information.

4. Concluding remarks

In this paper we have presented the OMBI bilingual lexical resources for Dutch-Arabic and Arabic-Dutch. They are part of a larger set of bilingual lexical resources which are available at the Dutch HLT Agency. The main strength of the resulting bilingual computational resources is the high quality of the input data, which exceeds that of most existing computational resources, since it is based on the work of a team of professional lexicographers. In addition, most of these bilingual resources use the same Dutch component as a base, which offers interesting perspectives for linking the resources to each other following the hub and spoke model (Martin 2007).

References

Hoogland, J., Versteegh, K. and M. Woidich (eds.). (2003). *Woordenboek Nederlands Arabisch / Arabisch Nederlands*. Uitgeverij Bulaaq, Amsterdam.

ISO 24613:2008 *Language Resource Management -- Lexical Markup Framework*, ISO Geneva, 2008.

Maks, I. (2007). 'OMBI: The practice of Reversing Dictionaries'. In *International Journal of Lexicography* 20.3. 259-274.

Maks, I., C. Tiberius, and R. van Veenendaal. (2008). 'Standardising bilingual lexical resources according to the Lexicon Markup Framework'. In *Proceedings of LREC-2008.* Marrakech, Morocco. 1723-1727.

Martin, W. (2007). 'Government Policy and the planning and production of bilingual dictionaries: the 'Dutch' approach as a case in point'. In *International Journal of Lexicography* 20.3. 221-238.

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. 675-685.

Vliet, H. van der. (2007). 'The Referentiebestand Nederlands as a Multipurpose Lexical Database'. In *International Journal of Lexicography* 20.3. 239-258.