[*Translating and the Computer 28*, November 2006 (London: Aslib, 2006)]

Integrated bilingual specialist dictionaries - the LexTerm initiative

Dr. André Le Meur

Andre.lemeur@uhb.fr

Marie-Jeanne Derouin

marie-jeanne.derouin@langenscheidt.de

Abstract

Dictionary publishers dealing with millions of bilingual and multilingual data have to find answers to the following two main issues: accommodating the needs of a still large dictionary users community for traditional printed or electronic dictionaries and meeting the need of the professional users beyond Machine Readable Dictionaries on CD-ROM or Online. For this purpose the German specialist dictionary publisher, Langenscheidt Fachverlag (LFG) proposes a global solution together with experts from the University of Rennes 2 and well-known Translation Memory providers. LexTerm is a methodology for reusing lexicographical data and building a bridge between dictionaries and terminology.

1. Meeting the actual professional translators' needs

The specialised translation volumes are steadily increasing especially in the technical, economical and juridical fields. In the same time, companies and institutions which are the main order-givers have to cut down costs and in many cases the budget for language communication undergoes a very strict control. As a consequence, a lot of easy and repetitive translation tasks are achieved by non-professionals or translation software and professional translators are given the most difficult texts to translate for which they have to charge at a reasonable rate in order to keep getting orders. In this context, they have to optimise their work-flow at a maximum and are looking for time-sparing tools and strategies.

In the past decade, terminology management and translation memory providers have equipped them with most valuable translation management tools which enable them to compile their own dictionaries and to store their validated translated text-segments in order to reuse them in future translations. These tools are regularly improved in new versions which offer users a better comfort. On the other hand, dictionary publishers who have been for years the main tool providers for professional translators have digitalized their dictionary data and published several generations of electronic dictionaries, regularly adding new features in order to provide updated bilingual specialist terms.

The next challenge is to propose now a global solution with a unique tool which will ensure translators the reuse of their stored data together with an easy and quick access to the specialist terminology they never had translated before and therefore need. This unique tool: translation memory with “à la carte” integrated specialist dictionaries will save time-consuming internet researches and browsing in print or electronic dictionaries.

2. The proposed solution

TMS providers and bilingual specialist dictionary publishers have the same target group but completely different approaches in data management and product marketing. The global solution we propose to translators means some major changes in the dictionary publishers' data processing and in TMS product concept. With the introduction of digital support, dictionary lifecycle has been considerably extended. The original manuscript has now become a unique source that can be accessed many times in order to be reused and even integrated in other language applications. For such data manipulations as integrating lexicographic data in terminological tools, contents have to be structured according to standards recognized by other professionals in order to avoid time-consuming and expensive data manipulations. The revised ISO standard 1951 due to be published next year aims to bridge the traditional methods of dictionary-making with the above mentioned future-oriented ones

One of the main challenges of this project has been to convert thousands of lemma-oriented lexicographical data from bilingual specialist dictionaries into concept-oriented terminological data to be integrated in the translator's workbench. A method, concrete XML-models, encoding and finally a converter- described in 3. - have been developed by data-modelling experts from the University of Rennes 2 in France. As TMS providers wish to licence data from different specialist dictionary publishers and specialist dictionary publishers offer their data to different TMS providers, a concept-oriented data representation based on ISO standards has been chosen for smoothly data-exchanges. From now on this final data conversion has been added in the previous editorial work flow which allows publishing specialist dictionaries in all possible publishing devices from a single source.

3. The LexTerm initiative: a methodology for transforming lemma-oriented data into concept-oriented terminological entries

Example of lexicographical entries

Figure 1 shows four typical entries from an english-german technical dictionary.

- Entry 1 and 2 are “referring entries”. They only indicate that “aerating root” and “aerophore” are synonyms of “pneumatophore” in its first meaning.
- Entry 3 indicates that “pneumatocyst” has its own equivalents in the domain of botany but it is a synonym of “pneumatophore” in its second meaning (zoology)
- Entry 4 indicates that “pneumatophore” has two meanings according to the domain where it is used and that it has two german translations for the domain of botany and three translations for the domain of zoology

1. **aerating root** *s.* pneumatophore 1.
2. **aerophore** *s.* pneumatophore 1.
3. **pneumatocyst** 1. *(D: Bot)* Pneumatozyste *f,* Luftkammer *f (in einem Pneumatophor);* 2. *s.* pneumatophore 2.
4. **pneumatophore** 1. *(D: Bot)* Pneumatophor *n,* Atemwurzel *f; 2. (D: Zoo)* Pneumatophor *n,* Schwimmglocke *f,* Gasflasche *f (der Siphonophoren)*

Fig 1 Typical entries

Mapping methodology

This way of structuring data is “lemma driven”: each linguistic unit appears in the nomenclature of the dictionary, which is convenient for alphabetic access to the entries by a reader. For converting such data into a “concept oriented” structure acceptable by usual terminology management systems we had first to identify and map data elements from one system to the other and in a second time to convert lemma oriented structures (one linguistic unit and all its meanings) to concept oriented structure (one meaning and all its designations).

The TermBridge semantic repository (http://www.genetrix.org) developed in relation with the french Normalangue project has been used for the first task of identification and mapping of the data elements. It contains all the data elements and permissible values found in ISO 12620, which are specific to terminology (like “Term”) and, for lexicography, data elements and permissible values found in the new version of ISO 1951 (such as “Headword”).

Correspondence has been established between data elements (for instance what is in lexicography a “Headword” or a “Translation” is in terminology a “Term”) but most of the elements are identical (“Part of speech” or “Grammatical gender” for instance). The conclusion was that all the observed constituants of machine readable dictionaries have their counterpart in terminology repositories.

The second task has been to convert structures. Rules have been established in order to cluster linguistic units having the same meaning by grouping referring entries with the entry to which they refer.

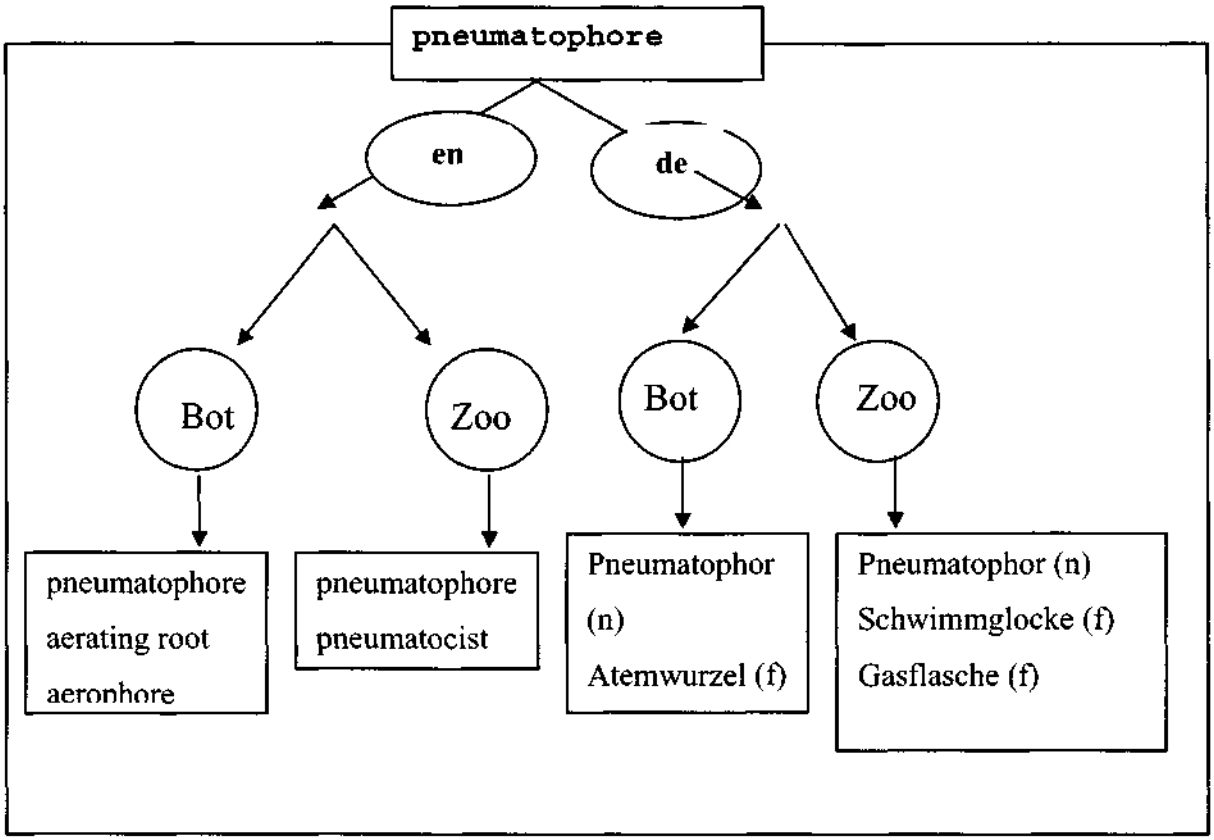

Figure 2: clustering of the synonyms

Another structural issue was that of the factorization of lexicographical elements (for instance in Figure 1, the note “der Siphonophoren” being after three translations applies to all of them). This structure is frequent in lexicography but standardized terminological formats don't admit such a feature. Consequently the solution was to replicate this type of information when necessary, that is for each term in this case.

Building an XML subset: LexTerm

In order to be platform and application indépendant, it was agreed with the translation memory providers, that the result of the conversion would be an XML format conformant to Geneter, a standardized Terminology Markup Language defined in ISO 16642 (Annex C) easy to import into any software via an XSL stylesheet. Geneter is “generic” which means that it takes into account all the terminological data categories defined in ISO 12620. A subset corresponding to the data categories and to the structures of technical dictionaries has been produced by applying the XML subsetting rules described in ISO 16642 C6. The result is the LexTerm model (http://wvvw.genetrix.org/dtd/LexTermVl-2.dtd) which is publicly

available so that anybody can produce data compatible with the translation memory providers import routines.

Conversion

Source data (Langenscheidt technical dictionaries) being encoded in XML, the conversion process consisted in transforming an XML tree into another XML tree according to the rules previously mentioned. Then it was possible to generate "terminological entries" according to ISO principles by separating each meaning with all its designations in the different languages. For our example, the result is two "concepts" belonging to different domains:

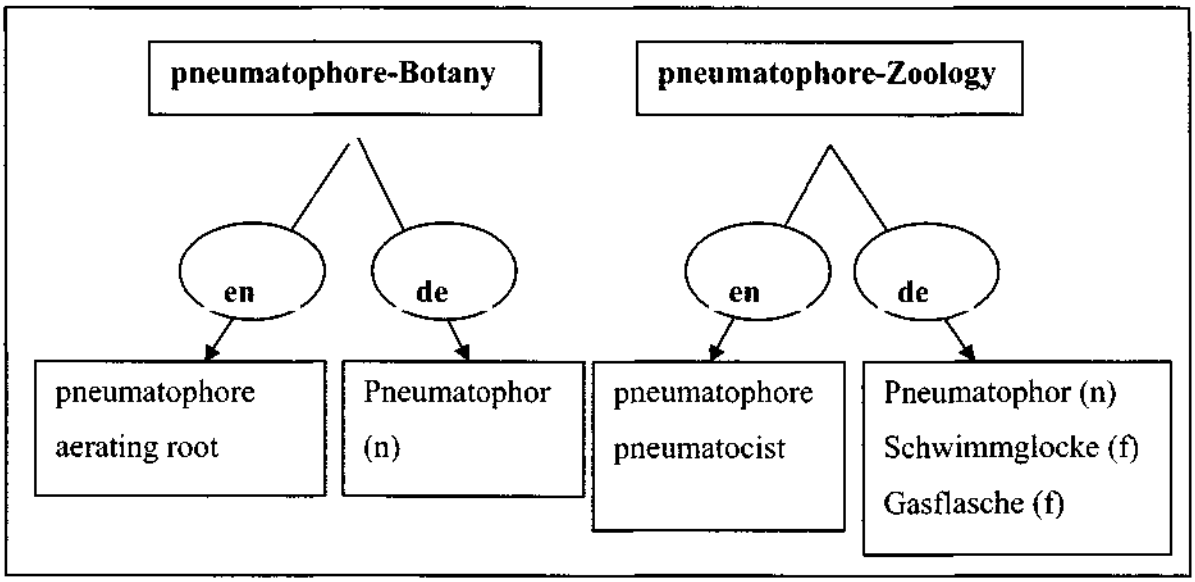

Figure 3: splitting concepts

Source and target XML encoding

The two following XML examples illustrate the parallelism and the divergences between the lexicographical model and the terminological model.

Figure 4 shows the XML encoding of the entry "**pneumatophore**" (Figure 1, line 4) conforming to ISO 1951 (note that it is not the real Langenscheidt encoding which is older than the revision of ISO 1951: structures are similar but their expressions are different)

```
<Dictionary version = 'LEXmlVOS' profile = 'LEXmlVOS' >
  <DictionaryEntry identifier = 'I1'
     sourceLanguage = 'en'
     targetLanguage = 'de'>
    <Headword>pneumatophore</Headword>
    <SenseGrp>
      <SubjectField>Bot</SubjectField>
      <TranslationCtn>
        <Translation>Pneumatophor</Translation>
          <PartOfSpeech value = 'noun'/>
        </TranslationCtn>
        <TranslationCtn>
          <Translation>Atemwurzel</Translation>
          <GrammaticalGender value = 'feminine'/>
        </TranslationCtn>
      </SenseGrp>
      <SenseGrp'>
        <SubjectField>Zoo</SubjectField>
        <TranslationCtn>
          <TranslationCtn>
            <Translation>Pneumatophor</Translation>
            <PartOfSpeech value = 'noun'/>
          </TranslationCtn>
          <TranslationCtn>
            <Translation>Schwimmglocke</Translation>
            <GrammaticalGender value = 'feminine'/>
          </TranslationCtn>
          <TranslationCtn>
            <Translation>Gasflasche</Translation>
            <GrammaticalGender value = 'feminine'/>
          </TranslationCtn>
          <Note>der Siphonophoren</Note>
        </TranslationCtn>
      </SenseGrp>
    </DictionaryEntry>
  </Dictionary>
```

Figure 4: XML encoding of “pneumatophore”

Figure 5 contains a result of the conversion: the conceptual entry based on the second meaning of entry 4 (botany).

As an instance of the LexTerm dtd, it validates against the URL previously seen and LexTerm being a subset of Geneter this sample validates against

http://www.genetrix.org/dtd/TermBridge_V00/dtd/TermBridge_V00.dtd.

```
<Geneter version = 'GeneterV0.8' profile = 'LexTermVl-2'>
<TerminologicalEntry identifier='mytest-9'>
 <SubjectField>Botany</SubjectField>
 <LanguageCtn value='en'>
   <TermCtn>
     <Term>pneumatophore</Term>
  </TermCtn>
  <TermCtn>
```

```
9.     <Term>pneumatocyst</Term>
10.      </TermCtn>
11.    </LanguageCtn>
12.     <LanguageCtn value='de'>
13.      <TermCtn>
14.       <Term>Pneumatophor</Term>
15.       <PartOfSpeech value = 'noun'/>
16.       <Note>der Siphonophoren</Note>
17.      </TermCtn>
18.      <TermCtn>
19.       <Term>Schwimmglocke</Term>
20.       <GrammaticalGender value = 'feminine'/>
21.       <Note>der Siphonophoren</Note>
22.      </TermCtn>
23.      <TermCtn>
24.       <Term>Gasflasche</Term>
25.       <GrammaticalGender value = 'feminine'/>
26.       <Note>der Siphonophoren</Note>
27.      </TermCtn>
28.    </LanguageCtn>
29.    </TerminologicalEntry>
30.    </Geneter>
```

Figure 5: XML encoding of the botany related entry

4. The product

6 specialist bilingual dictionaries in the language combination English-German/German-English in Electrical Engineering-Electronics, Architecture and Construction, Chemistry, Biology, Medicine, Business and Banking and 4 large technical dictionaries in English, French, Spanish and Italian including altogether more than 1.600.000 specialist terms in over 100 subject fields should be available on the market in 2007. Translators who either already work with a TMS tools or new users will be able to purchase the integrated dictionaries from their TMS provider and get easy and quick access to one of the largest collection of bilingual specialist dictionaries in Europe. A yearly update is planned for each subject field. The following example of screenshot shows a very user-friendly interface. The unknown specialist terms in the user's language are marked in the partly translated text. At the bottom of the display different possible translations with reference to the source, subject field, semantic and pragmatic information are proposed. The user needs only to choose one of them and it will be placed automatically in the translated text.

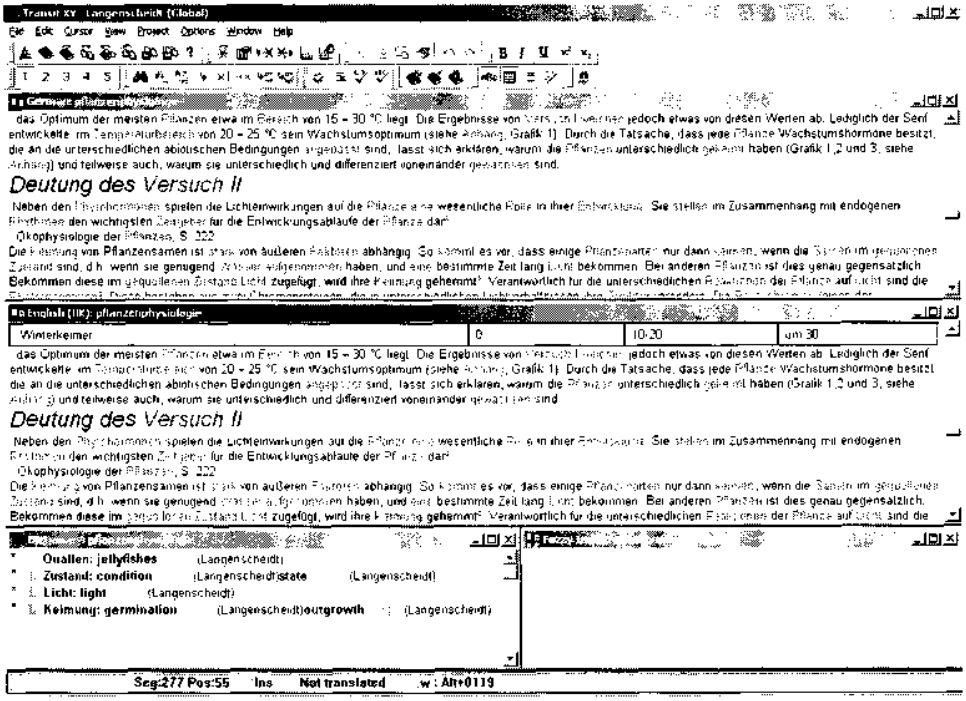

Figure 6: An example of a screenshot

5. Conclusion

Although we are aware of the limits of this tool which are in fact the limits of bilingual specialist dictionaries, we are proud of our contribution to the efficiency of professional translators' work. This solution is one of the best illustrations of the evolution from a specialist dictionary publisher to a language content provider in cooperation with the language industry and university research.

Last but not least the described "global solution" points out the growing convergence of Lexicography and terminology for the future issues of language communication.

References:

ISO 704:2000, *Terminology Work: Principles and Methods*
SO 16642:2003 *Computer applications in terminology— Terminological markup framework* (available in English only)
ISO 1951 (FDIS) *Presentation/Representation of entries in dictionaries*
XmLex Introduction: http://www.XmLex.net/lexicography/xmlexintro.pdf
XmLexLib: Xsl libraries for ISO 1951 conformant lexicographical data: http://www.XmLex.net/lexicography/XmLexWorkbench.rar