Lemma-oriented dictionaries, concept-oriented terminology and translation memories

André Le Meur*, Marie-Jeanne Derouin**
* Laboratoire RESO-CNRS, Université de Rennes 2
6, Avenue Gaston Berger, F-35043 Rennes, France
Andre.lemeur@uhb.fr
** Langenscheidt Fachverlag
Mies-van-der Rohe-Strasse 1, D-80807 München, Germany
marie-jeanne.derouin@langenscheidt.de

Abstract

Market surveys have pointed out translators' demand for integrated specialist dictionaries in translation memory tools which they could use in addition to their own compiled dictionaries or stored translated parts of text. For this purpose the German specialist dictionary publisher, Langenscheidt Fachverlag in Munich has developed a method and tools together with experts from the University Rennes 2 in France and well known Translation Memory Providers.

The conversion-tools of **dictionary entrie**s ("lemma-oriented") in **terminological entries** ("concept-oriented") are based on lexicographical and terminological ISO standards: ISO 1951 for dictionaries and ISO 16642 for terminology. The method relies on the analysis of polysemic structures into a set of data categories that can be recombined into monosemic entries compatible with most of the terminology management engines on the market.

The whole process is based on the TermBridge semantic repository (http://www.genetrix.org) for terminology and machine readable dictionaries and on a XML model "LexTerm" which is a subset of Geneter (ISO 16642 Annex C). It illustrates the interest for linguistic applications to define data elements in semantic repositories so that they are reusable in various contexts.

This operation is fully integrated in the editorial XML workflow and applies to a series of specialist dictionaries which are now available.

1. Specialist dictionaries for the actual professional translators

The specialised translation volumes are steadily increasing (about 14% per year) 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 optimize their work-flow at a maximum and are looking for time-sparing tools and strategies.

In the past decade, 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. On the other hand, terminology management and translation memory providers (TMS) have equipped the professional translators with most valuable translation management tools which enable them to compile their own dictionaries and to store their validated translated text-segments for reuse in future translations. These tools are regularly improved in new versions which offer users a better comfort.

In order to meet the growing demand for an efficient global solution in matter of translation tools, TMS providers and bilingual specialist dictionary publishers have decided to propose 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.

This ambitious project is a joint challenge for TMS providers and specialist dictionary publishers. They have the same target group but completely different approaches in data management and product marketing. For both of them it means cooperation with other partners and hence a profound need for standardisation in matter of data-modelling and presentation.

2. From lemma-oriented dictionaries to concept oriented terminological dictionaries

The integration of bilingual specialist data in a translator's workbench requires concept-oriented data. For a bilingual specialist dictionary publisher this requirement means an entirely new procedure as lemma-oriented XML-data are mainly used as a unique source for print and electronic dictionaries. Compared to the terminological specialist dictionary funds, the lexicographical bilingual specialist one is huge. In our case, more than a million of lemma-oriented data had to be converted to ensure that the main subject fields could be taken into account.

The development of a method, concrete XML-models, encoding and finally a converter – described in 3 could not be carried out in house as lexicographers have very little experience with terminology. This is the reason why the publisher entrusted data-modelling experts from the University of Rennes 2 in France with this project.

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, every specialist dictionary is available in two versions: a lemma-oriented one for print and electronic dictionaries which are proposed separately and a concept-oriented one for all possible integrations in translation-tools. The described conversion has been fully integrated in the previous editorial work flow and allows publishing specialist dictionaries in all possible publishing devices from a single source.

3. Transforming lemma-oriented data into concept oriented entries. Methodology

3.1. 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 "aerosphere" 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

3.2. 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) developped 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 constituents of machine readable dictionaries have their counterpart in terminology repositories.

The second task has been to define a strategy for converting structures. Rules have been established in order to cluster linguistic units having the same meaning by grouping "referring entries" (like "aerating root") with the entry to which they refer (the first meaning of "pneumatophore" in our example)..

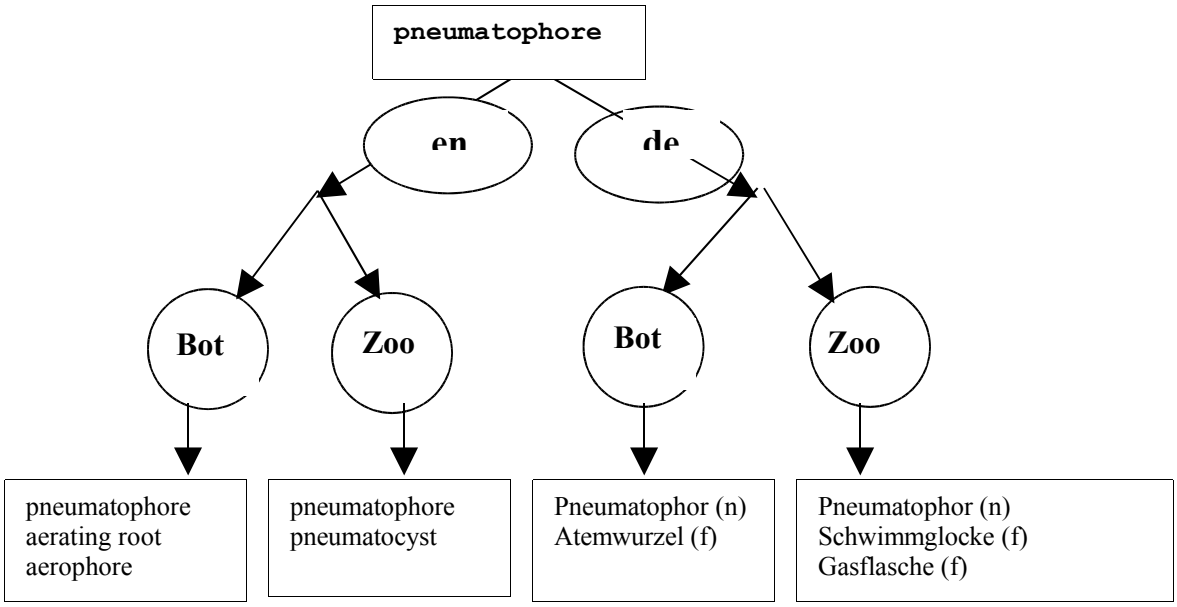

Figure 2: clustering of the synonyms

Another structural issue has been the factorization of lexicographical elements (for instance in Figure 1, the note "der Siphonophoren" being displayed after three translations applies to all of them). This structure is frequent in lexicography (it is encoded by a "Block" structure in ISO 1951 see figure 4, lines 21-35) but standardized terminological formats don't support such a feature. Consequently the solution has been to replicate this type of information when necessary, that is for each term in this case (see figure 5, lines 15-26)..

3.3. Building an XML subset: LexTerm

In order to be platform and application independent, 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 a 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://www.genetrix.org/dtd/LexTermV1-2.dtd) which is publicly available so that anybody can produce data compatible with the translation memory providers import routines.

3.4. 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.

It has been then possible to generate "terminological entries" according to ISO principles by

- grouping synonyms with their main headword. For instance "pneumatocyst" is a synonym of "pneumatophore" in its second meaning (Zoology). The result of this first step (figure 2) is to group within a unique ISO 1951 "sense" all the designations of a "unit of knowledge" (ISO 704 and ISO 16642 terminology)
- splitting each sense in a monosemic unit. For our example, the result (figure 3) is two terminological entries belonging to two different domains : "Botanic" and "Zoology".

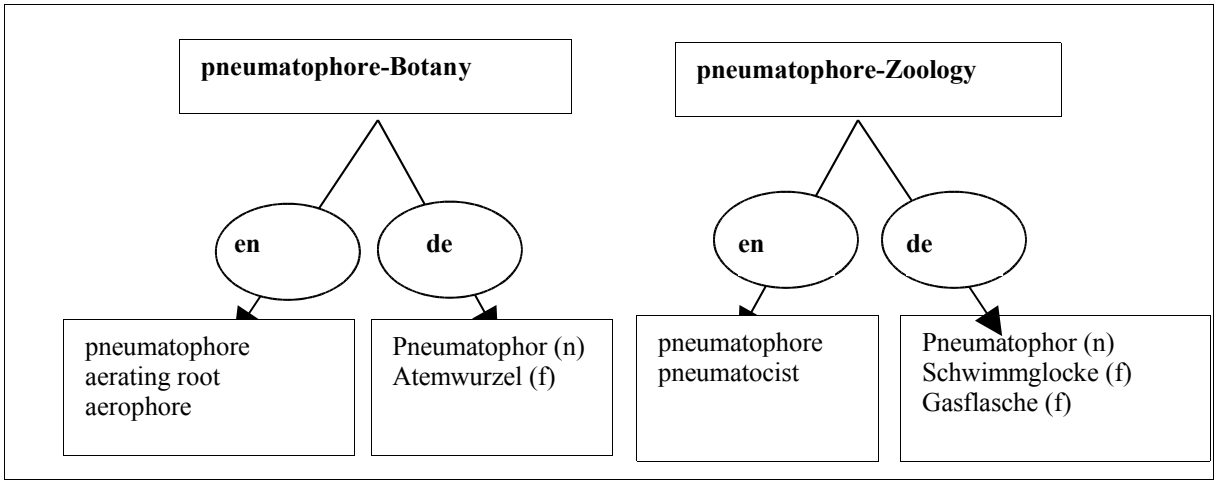

Figure 3: splitting concepts

3.5. 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)

```
1.<?xml version ="1.0" encoding="ISO-8859-1" ?>
2.<!DOCTYPE Dictionary SYSTEM
  'http://www.genetrix.org/common/XmLex/XmLex_V00.dtd'>
3.<Dictionary version = 'XmLex_V00'>
4.  <DictionaryEntry identifier = 'ID-pneumatophore'
5.     sourceLanguage = 'en'
6.     targetLanguage = 'de'>
7.    <Headword>pneumatophore</Headword>
8.    <SenseGroup>
9.      <SubjectField>Zoology</SubjectField>
10.      <TranslationCtn>
11.        <Translation>Pneumatophor</Translation>
12.        <PartOfSpeech value = 'noun'/>
13.      </TranslationCtn>
14.      <TranslationCtn>
15.        <Translation>Atemwurzel</Translation>
16.        <GrammaticalGender value = 'feminine'/>
17.      </TranslationCtn>
18.    </SenseGroup>
19.    <SenseGroup>
20.      <SubjectField>Zoo</SubjectField>
21.      <TranslationBlock>
22.        <TranslationCtn>
23.          <Translation>Pneumatophor</Translation>
24.          <PartOfSpeech value = 'noun'/>
25.        </TranslationCtn>
26.        <TranslationCtn>
27.          <Translation>Schwimmglocke</Translation>
28.          <GrammaticalGender value = 'feminine'/>
29.        </TranslationCtn>
30.        <TranslationCtn>
31.          <Translation>Gasflasche</Translation>
32.          <GrammaticalGender value = 'feminine'/>
33.        </TranslationCtn>
34.        <Note>der Siphonophoren</Note>
35.      </TranslationBlock>
36.    </SenseGroup>
37.  </DictionaryEntry>
38.</Dictionary>
```

Figure 4: XML-encoding of the dictionary entry "pneumatophore" conforming ISO 1951 (revised)

Figure 5 contains a result of the conversion: the two monosemic entries corresponding to the dictionary entry “pneumatophore”.

As an instance of the LexTerm dtd it validates against the URL previously seen. These entries are compatible with most of the Translation Memory Systems on the market.

```
<?xml version='1.0' encoding='iso-8859-1'?>
<!DOCTYPE Geneter SYSTEM 'http://www.genetrix.org/dtd/LexTermV1-2.dtd'>
<Geneter version = 'GeneterV0.8' profile = 'LexTermV1-2'>
<TerminologicalEntry identifier='pneumatophorZoology'>
 <SubjectField>zoology</SubjectField>
 <LanguageCtn value='en'>
  <TermCtn>
   <Term>pneumatophore</Term>
  </TermCtn>
  <TermCtn>
   <Term>pneumatocyst</Term>
  </TermCtn>
</LanguageCtn>
 <LanguageCtn value='de'>
  <TermCtn>
   <Term>Pneumatophor</Term>
   <Note>der Siphonophoren</Note>
  </TermCtn>
  <TermCtn>
   <Term>Schwimmglocke</Term>
   <GrammaticalGender value = 'feminine'/>
   <Note>der Siphonophoren</Note>
  </TermCtn>  <TermCtn>
   <Term>Gasflasche</Term>
   <GrammaticalGender value = 'feminine'/>
   <Note>der Siphonophoren</Note></TermCtn>
</LanguageCtn>
</TerminologicalEntry>
<TerminologicalEntry identifier='pneumatophorBotanic'>
 <SubjectField>botanic</SubjectField>
 <LanguageCtn value='en'>
  <TermCtn>
   <Term>pneumatophore</Term>
  </TermCtn>
  <TermCtn>
   <Term>aerating root</Term>
  </TermCtn>
  <TermCtn>
   <Term>aerophore</Term>
  </TermCtn>
</LanguageCtn>
 <LanguageCtn value='de'>
  <TermCtn>
   <Term>Pneumatophor</Term>
   <Note>der Siphonophoren</Note>
  </TermCtn>
  <TermCtn>
   <Term>Pneumatophor</Term>
   <GrammaticalGender value = 'neuter'/>
  </TermCtn>
  <TermCtn>
   <Term>Atemwurzel</Term>
   <GrammaticalGender value = 'feminine'/>
  </TermCtn>
</LanguageCtn>
</TerminologicalEntry>
</Geneter>
```

Figure 5: XML encoding of the two terminological entries corresponding to the headword “Pneunatophore” conforming to ISO 16642

4. Towards integrated bilingual specialist dictionaries in translation memory tools

The resulting concept-oriented specialist bilingual data undergoes then a last conversion in the TMS-provider exchange format. For security reasons, the content of the data is finally coded and write protected in order to fulfil the copyright rules.

Six 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 will be available on the market by the end of 2006.

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.

In the future 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.

Conclusion

The solution offered by this experimental project presents two major advantages: an added value to TMS-tools looking for high quality specialist dictionary content and for the specialist dictionary publisher a better adequacy to the translation market demand.

The described global solution for the translation industry illustrates the merging of complementary know-how of its different actors: language industry, dictionary publishers as content provider and university research. Moreover it stresses the importance of standardisation in matter of data-modelling. Last but not least it points out the growing convergence of Lexicography and Terminology for the future issues of language communication.

5. References

ISO 704:2000, *Terminology Work: Principles and Methods*

I

SO 16642:2003 *Computer applications in terminology -- Terminological markup framewor*k (available in English only)

ISO 1951 (FDIS) *Presentation/Representation of entries in dictionariesw*

XmLexLib : Xsl libraries for ISO 1951conformant lexicographical data : http://www.genetrix.org/common/docs/lexicography/XmLexLib.pdf