Ongoing Changes in Lexicographical International Standards: Report on the Revision of ISO 1951 Lexicographical Symbols and Typographical Conventions for Use in Terminography and Proposals for the first Draft: Presentation/Representation of Entries in Dictionaries

Marie-Jeanne Derouin
Managing Director
Langenscheidt Fachverlag
Postfach 40 11 20
D-80711 München
Germany
marie-jeanne.derouin@langenscheidt.de

Dr. André Le Meur
Maître de conférences en Informatique
Laboratoire RESO - CNRS - UMR 6590
Université de Rennes 2
6, Avenue Gaston Berger
F-35043 Rennes, France
andre.lemeur@uhb.fr

Abstract

The two authors of this paper belong to the expert commission of the standardisation bodies in France (AFNOR) and in Germany (DIN) and are, within the ISO/TC37/SC2, project leader and expert for the revision of the ISO-standard 1951. In this paper, they will report on:
- the needs for the revision of this standard, for example: inadequacy of the existing standard to take into account the computer-based lexicography and necessity of harmonising lexicographical sources for reuse and exchange purposes.
- the ongoing revision of ISO 1951: the decisions which have been taken during the last ISO/TC37/SC2 meeting 2001 in Toronto and information on the first proposals for the future revised ISO 1951 which have been compiled during the beginning of this year 2002 by the expert commission. This last part will lay stress on the significance of a secure and efficient data-format underlying the superficial representation of entries in dictionaries and will show the relationship between terminological and lexicographical international standards with its links and limits.

Introduction

In this paper, we will first report on the needs for revision of the existing national and international standards in lexicography which list from the inadequacy of these existing standards to take into account the computer-based lexicography to the considerable need for harmonising lexicographical sources for reuse and exchange purposes.

We will then give information on the revision of ISO 1951 which has been decided after the results of a feasibility study which has been carried out in more than 15 countries. Its new title: *Presentation/representation of entries in dictionaries* indicates that a larger scope has been chosen and that the artificial separation between general language and specialised dictionaries is no longer valid, as far as the representation of entries is concerned. This part of the paper will report on the decisions which have been taken during the last ISO/TC37/SC2 meeting 2001 in Toronto and on the first proposals for the future revised ISO 1951 which have been compiled in this year 2002 by experts from Austria, Belgium, Canada, Finland, France, Germany, Greece, United Kingdom and Ukraine. This last part will lay stress on the significance of a secure and efficient data-format underlying the superficial representation of entries in dictionaries and will show the complementary relationship between terminology and lexicography with regard to international standards with its strength and limits.

1 The Need for the Revision of the Existing Standard ISO 1951

The first edition of ISO 1951: *Lexicographical symbols and typographical conventions for use in Terminography* was published in 1973. According to its introduction, it "deals with the use of lexicographical symbols and typographical conventions in terminological entries in specialized dictionaries in general and standardized vocabularies in particular". Although the revised edition of 1997 aims "to harmonize the use of symbols and typographical conventions in terminography by taking into account theoretical and scientific traditions as well as the development of computer hardware and software", it lists mainly typographical conventions, language and country codes, grammatical information and lexicographical entries without giving any recommendation regarding the organisation of lexicographical entries and taking into account the computer-based dictionary manuscript and its various uses and reuses on different print and electronic devices. And this is exactly what people involved in the dictionary making process actually need.

The numerisation of data revealed two important facts: the lack of consistence of the structure of these data within a dictionary or a publishers' programme and its inadequacy for electronic devices. Manuscripts have now to be considered as single sources for multiple print and electronic devices and must therefore be structured in order to fit to the largest possible publishing devices, to be bundled with other titles, or to allow extraction of entries for concise editions. Consequently most dictionary publishers are now redefining a common structure (DTD) for their titles and also the lexicographical conventions for their authors and editors.

Lexicographical symbols which were only used for paper dictionaries in order to spare room, such as "tilde" tend to disappear. On the other hand, entries are getting more and more complex as electronic devices allow us to store large amounts of information and users ask for it. Reference to subdomains, supplementary information for disambiguation of translations, references to norms, illustrations, synthetic voice etc. are progressively being included in dictionaries.

With regard to the distinction between comprehensive and concise editions, every dictionary entry or part of a entry receives distinguishing marks. Consequently, there is a need for recommendations with regard to the superficial structure of dictionary entries which is visible to compilers of entries and dictionary users and also with regard to the deep structure which underlies the superficial structure and allows different applications of the data.

In the same time the need for exchanging lexicographical data is growing: it is getting more and more difficult to find dictionary authors as they have to provide more information according to strict lexicographical rules for the same amount of money and publishers have started to import data from owners of terminological collections and non-commercial dictionaries, either in industrial companies or in institutions. A part of their business is also to export data in order to create new titles with other language combinations or mix-products

such as a language course linked with a bilingual dictionary, or a bilingual specialist dictionary linked to a general language dictionary or an encyclopaedia.
In all these exchanges of data, the need for harmonising data and for an efficient data-format underlying the superficial entries is considerable and urgent in order to avoid losing information and precious time and keep the costs of dictionary making at a reasonable level.

Last but not least, the experience of merging general and specialist dictionaries has proved that there is very little difference between the deep structure of these dictionaries and that it would be wise to publish a standard for the lexicographical entries of all types of dictionaries: monolingual, multilingual, based on general language and specialist vocabularies which will take all the structural parameters into account.

2 Report on the Revision of the Standard ISO 1951 on Lexicographical symbols and typographical conventions for use in terminography

The first steps towards the revision of this standard started in 2000 when the German ISO-delegation reported in London on the ongoing updating of the equivalent national DIN-standard 2336 *Lexikographische Zeichen für manuell erstellte Fachwörterbücher* according to the needs we have presented in the first part of this paper. Consequently it has been decided to check whether the needs expressed in Germany can apply to most countries in the world or not and a feasibility study has been carried out in every ISO-member country. A questionnaire has been sent to lexicographers in universities or special schools, specialized dictionary authors, specialized dictionary publishers, terminology department of industrial companies and national or international bodies.

The results of the feasibility study show that most countries insist on the fact that ISO 1951 does not anymore meet the current needs in lexicography. In Sweden, for example, the ISO-standard has not been adopted as a national standard and it has not had any impact on current lexicographical and terminographical practice. The Nordic Association for Lexicography applies its own model for the presentation of entries in specialized dictionaries, terminological vocabularies and databases. In most countries there is an urgent need for a standard for representing and exchanging data of special languages which should take in account the needs of the computer-based lexicography in order to get a consistent representation of the entries and therefore homogeneous dictionaries.

The French AFNOR wishes that the redefined standard should define a solid XML-based format (see below an example) for representing and exchanging data so that each collaborating partner would need one single and export routine. According to AFNOR, the scope of the standard should be larger that the only “specialized dictionaries” and it would be worth enlarging it to general- monolingual and multilingual-dictionaries.

In Germany, the revision of the DIN 2336 with its new title and scope, *Darstellung von Einträgen in Fachwörterbücher und Terminologie Datenbanken* is nearly finished. The German DIN is ready to propose the German revised standard as basis for the development of the revised ISO 1951.

During the Toronto ISO-meeting 2001, according to the positive feasibility study, a resolution to revise the ISO-standard 1951 has been approved. The revised standard will apply to general and specialized dictionaries and give a specific model for lexicography. Its objective is to facilitate the management, use, reuse and exchange of data for dictionaries. Its new title is: *Presentation/Representation of entries in dictionaries.*

Experts from nine countries Austria, Belgium, Canada, Finland, France, Germany, Greece, United Kingdom and Ukraine have started to develop this revision on the basis of the last draft of the new revised German standard in November 2001. The following first comments will underlay the proposals of the first draft of the future ISO 1951 standard which will be submitted at the next ISO-meeting in Vienna in August 2002.

a) Although the forthcoming new German DIN-standard 2336 provides a variety of possible layouts for presenting data in different electronic environments it appears to be too much focused on the print specialised of dictionaries since only one subclause is devoted to the presentation of databank entries. It is restricted to the presentation issues concerning typographical characters and conventions and types of entry arrangement, without working on the systematical structuring of the presented lexical information, such as sequence of information. Consequently it can serve as basis for starting the revision but will have to be significantly extended and restructured.

b) The future new ISO-standard 1951 will have to cover the different options of organisation and management of data for print and electronic environments. In other terms, a formal model independent from presentation of data is needed. This model should be build in order to obtain any layout (and particularly the DIN 1336) through stylesheets, and to fulfil the requirements for electronic editing, storing, querying and dissemination.

c) It will have to cover a wide range of lexicographical resources such as general and specialised dictionaries, monolingual and multilingual, Machine Readable Dictionaries (MRDs) etc.

d) Uniformity at the exchange of data should be ensured. Except for the specifications for typographical conventions, already described in the present ISO-standard 1951, we need a more generic data exchange format. Moreover, a DTD should be initiated so as the creators and the users of the lexical collections to be confident that can (re)produce and use unambiguously parts or the whole of the included information. That DTD should also cater for optionality issues of the data, combination of data categories, which may influence the presentation options providing a structured generic exchange format.

e) For that, the experts will have to take into account the published inventories of data categories or format specifications for dictionaries and lexicons like TEI[1] (Text Encoding Initiative), EAGLES[2], EAGLES-ISLE[3], ISLE[4], IMCI[5] and other works related to this matter such as Pierre Corbin's EURALEX 2002 paper on "*Composants lexicographiques et contenus informationels des dictionnaires*".

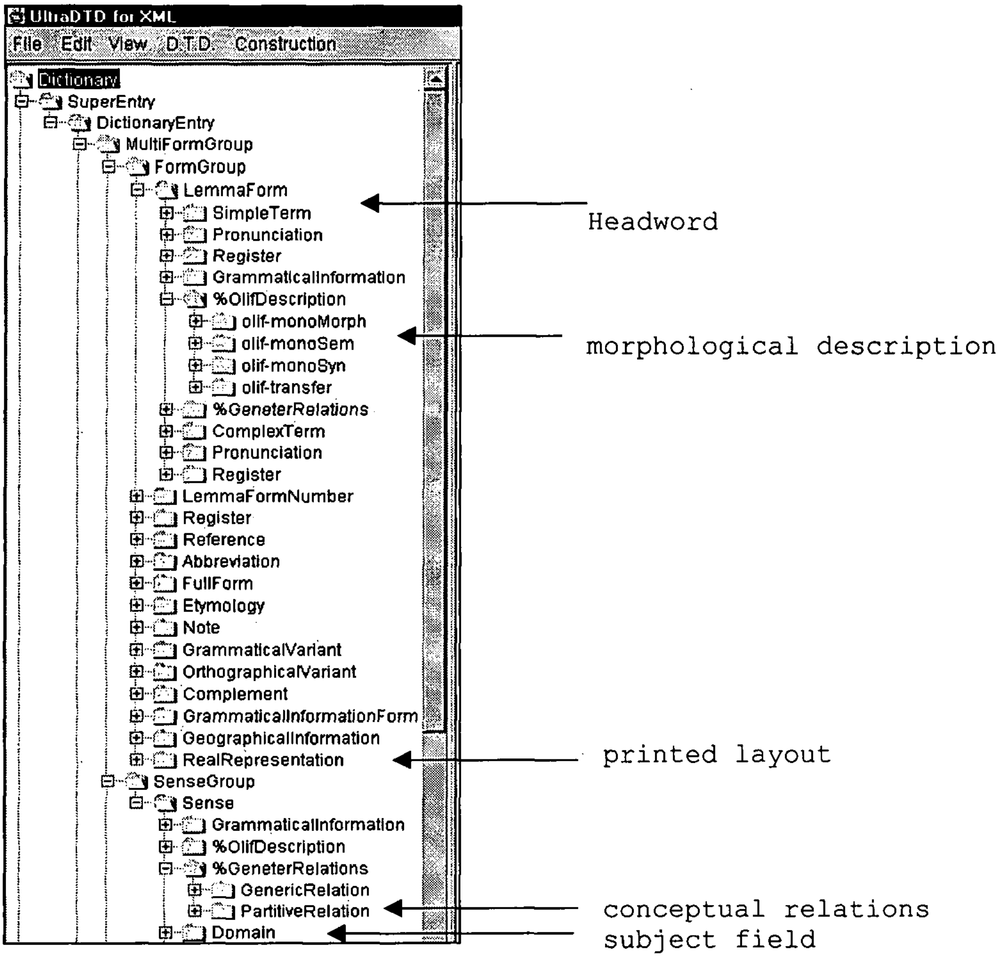
f) Moreover lexicographical description models have to be compatible with other linguistic ressources like machine oriented lexicons[6] and concept oriented terminologies[7].

First steps : towards a new formal representation of entries in dictionaries

In a previous paper [DEROUIN, LE MEUR 2000] a first inventory of data categories has been presented, based on the observation of seven technical dictionaries. This inventory considers now thirty technical, general, bilingual or monolingual dictionaries[8]. It shows that more than sixty elements are required in order to represent all the informations we can find in dictionaries, that many elements (administrative information for instance) are commun to all linguistic resources and that an accurate description of many elements can be borrowed to existing more specialized formats. For instance ontological relations can be borrowed to concept oriented terminological formats and morphological, syntactical and semantical can be borrowed to machine oriented lexicons. A first draft of a formal dictionary model can be built ont these bases. It takes into account most of the structural features that are described in the previouly mentioned analysis (TEI, ISLE, etc.). The example below shows how a classical entry of a technical german-english technical dictionary maps on this structure.

Läufer *m* **1**. (*El*) rotor *m*, induit *m* (*bei Gleichstrommaschinen*); **2** (Strm) rotor *m*, roue *f* mobile (*s.a.* Laufrad 1.); **3**. curseur *m* (*z. B. einer Spinnmaschine*); **4**. couette *f* (coiffe *f*) vive, coulisse *f* vive (*Stapellauf*); **5** garant *m* (*Tau*); **6**. panneresse *f* (*Mauerwerk*); **7**. coulure *f* (*Ansttrichfehler*); **8**. *s.* Cursor

FIGURE 1 : Sample



```
Dictionary
SuperEntry
  DictionaryEntry
    MultiFormGroup
      FormGroup
        LemmaForm
          SimpleTerm
          Pronunciation
          Register
          GrammaticalInformation
          %OlifDescription
            olif-monoMorph
            olif-monoSem
            olif-monoSyn
            olif-transfer
          %GeneterRelations
          ComplexTerm
          Pronunciation
          Register
        LemmaFormNumber
        Register
        Reference
        Abbreviation
        FullForm
        Etymology
        Note
        GrammaticalVariant
        OrthographicalVariant
        Complement
        GrammaticalInformationForm
        GeographicalInformation
        RealRepresentation
      SenseGroup
        Sense
          GrammaticalInformation
          %OlifDescription
          %GeneterRelations
            GenericRelation
            PartitiveRelation
          Domain
```

Headword

morphological description

printed layout

conceptual relations

subject field

Figure 2 :Tree structure of an entry

This figure illustrates the general outline of a lexicographical entry for Machine Readable Dictionnaries which keeps the traditional features of printed dictionnaries such as printed layout (see [DEROUIN, LE MEUR 2000]) but is enriched with morphological, syntactical and semantical features coming from Translation Oriented Lexicons (Olif) as well as with ontological relations coming from Terminology Markup Languages (ISO 16642 - annex C Geneter).

The full encoding and tools for validation and presentation (XSL stylesheet) are available at http://www.genetrix.org/lexicography/

XML encoding

```
<?xml version ="1.0" encoding="ISO-8859-1" ?>
<!-- ======================================= -->
<!DOCTYPE Geneter SYSTEM 'http://www.genetrix.org/dtd/GeneterV06.dtd' [] >
<Geneter >
<LEX:Dictionary>
<LEX:DictionaryEntry id='boch2' sourceLanguage = 'de'
                                targetLanguage = 'en'>
<LEX:FormGroup>
<LEX:LemmaForm>
<LEX:SimpleTerm>Läufers</LEX:SimpleTerm>
<olif:monoMorph>
          <olif:inflection>
            <olif:paradigm>
              <olif:inflectedForm>
                <olif:form>Läufers</olif:form>
                 <olif:monoMorph>
                  <olif:case>g</olif:case>
                  <olif:number>sg</olif:number>
               </olif:monoMorph>
              </olif:inflectedForm>
              <olif:inflectedForm>
                <olif:form>Läufern</olif:form>
                <olif:monoMorph>
                  <olif:case>g</olif:case>
                  <olif:case>d</olif:case>
                  <olif:number>pl</olif:number>
                </olif:monoMorph>
              </olif:inflectedForm>
            </olif:paradigm>
          </olif:inflection>
        </olif:monoMorph>
</LEX:LemmaForm>
</LEX:FormGroup>
<LEX:SenseGroup>
<LEX:Sense id='boch3'>
        <GenericRelation value = 'superordinateConcept'>motor
        </GenericRelation>
<LEX:TranslationGroup>
<LEX:TranslationEntity>
<LEX:Translation>
<LEX:SimpleTerm>rotor</LEX:SimpleTerm>
</LEX:Translation>
</LEX:TranslationEntity>
</LEX:TranslationGroup>
</LEX:Sense>
</LEX:SenseGroup>
</LEX:DictionaryEntry></LEX:Dictionary></Geneter>
```

Endnotes

[1] Text Encoding Initiative P4. Part 12 Printed dictionaries (http://www.tei-c.org/P4X/)

[2] Synopsis and Comparison of Morphosyntactic Encoded in Lexicons and Corpora, (http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/morphsyn.html)

[3] Preliminary Study of the Structure of Lexicon Entries, http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html

[4] Survey of Major Approaches Towards Bilingual/Multilingual Lexicons: http://lingue.ilc.pi.cnr.it/EAGLES96/isle/clwg_doc.html

[5] Metadata Elements for lexicon descriptions (http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf)

[6] Open Lexicon Interchange Format :www.olif.net

[7] Terminology Markup Framework (http://www.loria.fr/projets/TMF)

[8] LEX : Elements for a formal representation of lexicographical data categories -AFNOR - X03 A - G1 N7 (http://www.genetrix.org/lexicography/texts/Lex-en.doc)

References

[DEROUIN; LE MEUR 2000] Derouin, M.-J., Le Meur A. 2000. European Co-operation in standardisation of lexicographical resources and merging of existing specialised dictionaries for Internet Purposes, *Proceedings of the Ninth Euralex International Congress, Euralex 2000*