Standards going concrete: from LMF to Morphalou

Laurent Romary, Susanne Salmon-Alt, Gil Francopoulo

Standards going concrete: from LMF to Morphalou

Laurent Romary
Laboratoire Loria-INRIA
B.P. 239
54506 Vandœuvre Les Nancy
France
Laurent.Romary@loria.fr

Susanne Salmon-Alt
ATILF-CNRS
B.P. 30687
54063 Nancy
France
Susanne.Salmon-Alt@atilf.fr

Gil Francopoulo
INRIA-Syntax
B.P. 239
54506 Vandoeuvre Les Nancy,
France
Gil.francopoulo@wanadoo.fr

Abstract

Lexical resources are key components for applications related to human language technology. Various models of lexical resources have been designed and implemented during the last twenty years and the scientific community has now gained enough experience to design a common standard at an international level. This paper thus describes the ongoing activity within ISO/TC 37/SC 4 on LMF (Lexical Markup Framework) and shows how it can be concretely implemented for the design of an on-line morphological resource for French in the Morphalou project.

1 Introduction

Lexical resources play a crucial role in most applications related to human language technology. They may be used by both human readers and automatic processors for a wide range of activities that require an even wider variety of lexical structures. Some applications may demand broad linguistic coverage, where the word is the entry point and all the possible senses are attached to them, whereas other applications could require a concept-based organization of the lexical data, from which the relevant words (or terms) may be derived. Some applications barely need more then a simple list of words, whereas other may require a precise morpho-syntactic, syntactic and semantic description of the various lexical entries.

Furthermore, the huge cost of creating and maintaining a lexical resource in any of these domains requires that they should not be designed in isolation but that they may potentially be linked with one another for mutual enrichment.

As a consequence, we believe that there is a strong need for more widely accepted methods for specifying lexical structures, so that the conditions under which the corresponding databases can exchange data are precisely defined. Moreover, it seems that enough knowledge has been gathered across the years to contemplate the idea that such technical principles and methods could be the source of an international standard that would preserve the possibility of both describing various types of formats and ensuring interoperability among them.

This paper will present such a methodology as currently under discussion in the context of ISO committee TC 37/SC 4 in its on-going project called LMF (Lexical Markup Framework, which will become the future ISO 24613 standard). This international context also provides us with a unique opportunity to experiment with these most recent proposals in the context of the concrete necessity to deploy an open morphological dictionary for French (the Morphalou project). We will centre the discussion here on mapping the modelling principles that have been agreed upon so far in the LMF project with the actual requirements associated with the design of a morphological lexicon, hoping that it may lead to similar activities on lexical modelling in the future.

2 Standards for lexical resources

Before describing the ongoing standardization efforts within the LMF project, it is essential to get an idea of the actual background available to us in lexical representation at large and see how LMF may build upon, or rather receive input from, other past or ongoing standardization activities.

Lexical structures can classically be viewed according to the way they organize the relation between words and senses: either senses are considered as subdivisions of the lexical entry (the semasiological view of lexical data, which is the one usually applied in print dictionaries) or on the contrary, it is assumed that words (or “terms”) are described as ways of expressing *a priori* concepts, (the onomasiological view).

The onomasiological view has formed the basis for most previous standardization efforts since it is at the focus of many applied contexts. This trend started quite a while ago when the first standards for thesaurus representation were issued in the documentary field (ISO 2788 and ISO 5964). Those standards basically organize lexical matter as hierarchies of terms (e.g. broader-narrower terms), with the possibility of adding some basic lexical information (e.g. equivalences). More recently, the terminological field has provided

more elaborate standards within ISO committee TC 37, starting from the definition of an initial SGML/XML-based representation for terminologies (ISO 12200), and progressing on to the design of a flexible platform for specifying terminological structures (ISO 16642). The main problem with the onomasiological view is that even if it is well suited for providing homogeneous lexical descriptions within an application domain, it is hardly extensible when broader linguistic coverage is required.

In contrast, the semasiological view allows an exhaustive survey of lexical content for a given language. In particular, it provides the basis for any classical editorial (or print) dictionary, but the wide variety of possible dictionary formats seems to have hampered the development of international standards in this domain. The two main initiatives that can be cited here are on the one hand the ISO 1951 standard dedicated solely to the representation of dictionary entries, and on the other hand, the seminal work done within the TEI[1] on print dictionaries, which, even though it has already been applied to some large scale projects such as the OED[2], the *Deutsches Wörterbuch*[3] or the Anglo-Norman dictionary[4], has never been considered by publishers in particular as a real international standard. As a consequence, many relevant projects such as the TLFi[5] (Dendien & Pierrel, 2003) have designed their own proprietary structure for the description of their lexical archives.

If one moves away from classical dictionaries proper and considers lexical resources dedicated to the domain of NLP, there are numerous projects that have worked toward the definition of standardized lexical structures in the domain of NLP (Multext for basic morphological lexica; Genelex, Simple, Isle/Mile for complex multilingual entries; OLIF 1&2 for translation lexica, etc.), but none of them has lead to a standard that reflects a wide international consensus and that is effectively maintained by an authoritative body.

From a more theoretical point of view, it has been shown that such lexical structures can be modelled as feature structures (Ide et alii, 1995; Veronis & Ide, 1992), leading to inheritance properties within entries (Ide et alii, 2000), as partially implemented in the TEI Print Dictionary chapter (Ide &Veronis, 1995). It has also been shown that, with respect to describing the micro-structure of such lexica, at least three configurations are possible: 2-layered, 3-layered and 7-layered models. In the 2-layered approach, following Ferdinand de Saussure (1974), a word is described by a signifier/signified pair, corresponding to a morphological/semantic description. The syntactic behaviour of the word is then systematically attached to the semantic description. This is the approach that has been retained for LMF. In the 3-layer approach (Antoni-Lay et alii, 1994), a word is described by three units: a morphological, a syntactic and a semantic unit as in Genelex or Eagles. It should be noted that due to the fact that the syntactic unit is a mandatory connection between morphology and semantics, such a model is necessarily heavy and complex. In the 7-layered approach (Mel'cuk et alii, 1995), a word is described by various units in surface phonology, deep phonology, surface morphology, deep morphology, surface syntax, deep syntax and semantics. This approach imposes a heavy burden on the lexical description task.

Let us stress here the necessity of guaranteeing that the methods used to describe onomasiological and semasiological structures shall not be completely different, so that it is possible (as required by industrial applications in particular) to combine various kinds of lexical resources, but also to open the way for lexical architectures to combine concept-based and word-based descriptions as evidenced in the EDR dictionary[6], the Papillon project[7], or IBM's TransLexis resource.

3 The Lexical Markup Framework project

The LMF proposal, as currently being developed in ISO committee TC 37/SC 4, is conceived as a generic platform for the specification of lexical structures at any level of linguistic description. As such, it does not provide one single model, but rather a mechanism by which implementers combine elementary lexical subsystems to design models that can be both as close as possible to their needs and comparable to any other lexical models based on the same principles and, possibly, on the same components.

The underlying data model for LMF follows the general principles of the linguistic annotation scheme design stated in Ide & Romary, 2003 and implemented in the context of ISO standard 16642 for the representation of terminological data (Romary, 2001). Those principles provide a mechanism for combining a given structural

[1] Text Encoding Initiative (http://www.tei-c.or)

[2] http://dictionary.oed.com/

[3] http://www.DWB.uni-trier.de

[4] http://www.mhra.org.uk/

[5] http://www.atilf.fr/_ns/produits/tlfi.htm

[6] http://www2.crl.go.jp/kk/e416/EDR/index.html

[7] http://www.papillon-dictionary.org/

metamodel that informs the general organization of a certain level of linguistic information (morphology, syntax, etc.) with elementary descriptors (so-called *data categories*). Data categories reflect basic linguistic concepts (e.g. /part of speech/, /grammatical number/, /paucal number/, etc.) and allow for recording language-specific properties independently of linguistic level specific models. In order to share data categories within the community, on-going work (in ISO/TC 37) is in the process of deploying an on-line registry[8] of them, especially for use in conjunction with the other standardization activities.

According to these principles, LMF consists of the following elements:

- a core metamodel (i.e. the *structural skeleton* shared by any linguistic description at the lexical level);
- mechanisms for attaching lexical extensions (see below) to the core metamodel in order to build up more complex metamodels;
- mechanisms for selecting data categories used for lexical description and for determining how they relate to a metamodel;
- mechanisms for expressing any combination of the core metamodel and data categories as XML structures, i.e. by deciding to implement a given data category (/gender/) as an XML element rather than as an attribute and by providing the corresponding *vocabularies* ('gen', 'gender', 'genre');
- methods for describing how to extend LMF to analyze, design, and describe a variety of more specific lexical resources.

As shown in Figure 1, the core metamodel of LMF is organized as a purely hierarchical structure built upon the following components:

The *Lexical database* component gathers up all information related to a given lexicon;

The *Global information* component groups together all the metadata (e.g. version, contributors, up-date, etc.) that can be globally attached to the lexicon (see 4.4);

The *Lexical entry* component comprises the elementary lexical unit in a lexical database. This component can, of course, be iterated, but no specific constraint is expressed as to its level of granularity in a lexical database (e.g. proper treatment of homonyms), since this depends highly on languages and local editorial practices;

The *Form* component groups together all the general graphical or phonetic descriptions attached to the lexical entry (reference orthographic form, transliteration, hyphenation, pronunciation, etc.);

Finally, the *Sense* component is the one that actually organizes the lexical entry since it can be both repeated and further subdivided into senses. In a word-to-sense lexical structure, it is indeed thought that this central way of organizing a lexical entry should be part of the metamodel.

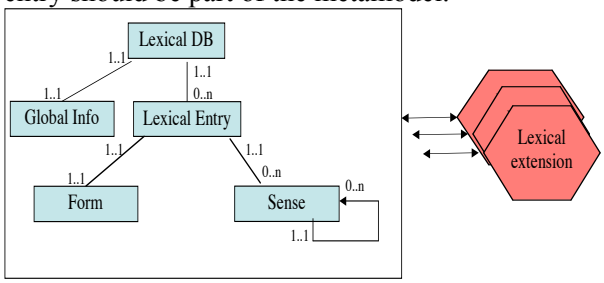

Figure 1: Core model and lexical extensions in LMF

In order to specify more complex models than would be expressible with just the core metamodel, LMF introduces the notion of lexical extensions. Those extensions correspond to clusters of components dedicated to the representation of a specific type of lexical information (e.g. morphology, syntactic constructions, transfer patterns (socalled *interlingua*), and theory dependant lexicographical approaches such as Mel'cuk et al. 1995 or Véronis, 2000). Each lexical extension is characterized by an anchor component, which is either a component of the core metamodel or of another lexical extension when more complex combinations are being considered (e.g. description of morphological operations used to extend a simple morphological lexical extension).

The future LMF standard as such should not provide a specific list of data categories to be used for lexical descriptions. This would by far be too complex given, as we have seen, the potential variety of applications. It is thus expected that implementers will systematically refer to the ISO/TC 37 data category registry to find the adequate descriptive background for their own purposes. Still, we can outline the basic types of data categories that one could encounter in an LMF based application, namely:

- data categories that may be considered as rather specific to the domain of lexical description: these are typically those attached to the Form component (/pronunciation/, /syllabification/, /stress pattern/ etc.) or to the Sense component (e.g. /definition/, /example/, /etymology/, etc.). Some of these categories have already been partially described in the 'old' ISO 12620:1999 standard, but a more precise list should be compiled as the work on LMF is being completed;
- data categories that relate to a specific level of linguistic description such as morphology, syntax, etc. The strategy here is to avoid defining

[8] An experimental on-line data category registry is accessible under http://syntax.loria.fr

ad hoc descriptors dedicated to lexical structures and to enforce coherence with other standardization activities by adopting those associated with the development of related standards. For instance, data categories such as /grammatical category/, /grammatical gender/ or /grammatical case/ should be shared between POS tagging applications and corresponding lexical descriptions;

- data categories corresponding to metadata descriptors used to document the production and maintenance of a lexical database, a lexical entry and probably, of any component in a lexical structure (see 4.4).

To conclude this brief presentation of LMF, which can only be considered to be a snapshot of the ongoing discussions about it, it is important to consider how it provides a whole standardization spectrum for implementers who will want to apply it for their own purposes. At a first level, they can limit themselves to the core model, to standardized lexical extensions and to the data categories that are available in the DCR. Doing so, they will have the certainty of being fully interoperable with any other implementation that has adopted the same scope. If necessary, it is possible for implementers to define some proprietary data categories or maybe their own lexical extensions, knowing that the corresponding part of their lexical model will probably require more work if they wants to interchange data with other applications. Still, such a strategy is probably the optimal one in the current stage of LMF, since, for instance, we do not know yet which lexical extensions will be sufficiently consensual to be further adopted as international standards. This is indeed the spirit in which the Morphalou project has been established, i.e. to design a simple morphological lexical extension to the LMF core principles and see how it could be validated when confronted with the real development of a lexical resource. In the long run, we do expect that some combinations of the core metamodel and some standardized lexical extensions may also be seen as possible future standards when they match specific industrial needs (e.g. transfer lexica à la OLIF) or existing practices (e.g. TEI Print Dictionary format).

4 An LMF-based model for a morphological lexicon

4.1 Requirements for a morphological lexicon

Morphological dictionaries typically associate inflected word forms (for example plural nouns or past tense verb forms) with values for relevant morphological features, such as gender and number for adjectives or person and tense for verbs. In addition, there is often a link to one particular word form, conventionally chosen as being the lemma. Those dictionaries are basic resources in the field of NLP (needed for any application based on tagged and/or lemmatized input data) and in the field of computer-assisted language acquisition.

Most existing morphological resources for NLP (*MulText*, Véronis 1999; *LEFFF*[9], Clément & Sagot) occur as text files, whose lines display the inflected word form, one or more morphological tags (relative to a given tag-set) and the lemma. This kind of representation, directly inspired by one specific type of usage of such resources (i.e. morphological tagging) takes the inflected form as an entry point. At the same time, the morphological point of view is an extensional one, in the sense that the resource explicitly contains the list of all inflected forms for one lemma. Furthermore, the linguistic concepts underlying the morphological description are not directly transparent and accessible, since the tags are generally synthetic tags for a set of values relative to a set of relevant features. Finally, if any metadata (such as the contributor or the last update) are associated with such a resource, they are often encoded in proprietary formats and there is no possibility to parameterize their scope to various description levels of the lexicon.

Starting from these observations, we tested LMF as a formal framework for the design of a morphological dictionary for French, based on existing data originally compiled during the digitization of a wide coverage French dictionary (*TFLi*). From a theoretical point of view, the aim of this experiment is to test the suitability of LMF at a quite simple level of lexical description. On the practical side, we wish to generate a resource that is accessible on-line and that implements the standardization proposals of ISO/TC 37/SC 4, and that is application-independent, well documented, extensible and provides the possibility to add further lexical description levels, such as syntactic and semantic information. Therefore, we have tried to overcome the aforementioned shortcomings of current morphological dictionaries by structuring the data around lemmas rather than around inflected forms, by proposing a data model that combines the co-occurrence of extensional and intensional morphological information (lists of inflexions vs. reference to inflexion classes or paradigms) and by paying special attention to the issue of the metadata necessary to qualify the identification of the source data (origin, contributor, up-date, etc.) and the status of the data (validated by an editorial committee, testified in a corpus, etc.).

[9] http://atoll.inria.fr/~lclement/lefff/

4.2 The lexical model of Morphalou

The underlying lexical model of the Morphalou project is a direct application of the LMF principles with the sole addendum of a simple lexical extension dedicated to the description of morphology. This extension can be directly linked to the lexical entry component of the core metamodel. It associates a single morphological description (*Morphology* component) to each lexical entry. This morphological description is made up of two sub-components:

- a *Paradigm* component that refers to or possibly describes the inflexion rules that govern the flexional behaviour of the entry;
- an *Inflexion* component that groups together zero up to *n* inflected forms related to the lexical entry.

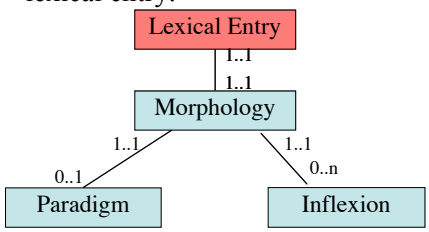

Figure 2: Lexical extension for morphology

As stated in section 3, to build up a full model for a concrete lexical database, one needs to associate a selection of data categories anchored at the different components of the metamodel (core metamodel + morphological lexical extension).

To the *Lexical entry* component, we basically associate to this component the data categories /lemma/ and /grammatical category/. A /key form/ is used in order to uniquely identify the entry within the lexical database. Possible orthographic variants may be recorded as /spelling variant/'s. Finally, depending on editorial choices, one could also decide to attach /gender/ information here, for example for nouns, in the case that gender variation is not considered as inflexional variation, as opposed to adjectives.

To the *Inflexion* component: beside /word form/, which identifies the actual inflected form in the component, it is necessary to associate the set of morphological features to provide a unique specification of the inflexion. The corresponding data categories are complementary to the general grammatical category of the entry: /gender/ and /number/ for adjectives; /tense/, /person/, /number/ and /mood/ for verbs, etc. Appendix 1 provides a complete list of the data categories we have considered for the first version of the database;

- *Paradigm* component: here we essentially need a /paradigm identifier/ to identify the inflexion class to which the lexical entry belongs. In order to integrate further data categories for the description of the inflexion rules, we still need to investigate linguistic practices for different language families.

4.3 Implementing the model: basic examples

Example 1 implements, in a generic XML format (GMT, see Romary 2001), a simple lexical entry and its morphological extension for the French noun *chat* ('cat'). The data categories associated with the lexical entry are /lemma/, /grammatical category/ and /key form/, respectively taking the values of *chat*, *noun* and *chat_1*. The morphology component contains the identification of the plural inflexion paradigm for regular French nouns (/fr-s-plural/) and the complete list of inflected word forms with associated morphological features, i.e. /number/.

Example 1

```
<struct type='lexical entry'>
   <feat type='lemma'>chat</feat>
   <feat type='grammatical category'>common
    noun </feat>
   <feat type='key form'>chat_1</feat>
   <struct type='morphology'>
      <struct type='paradigm'>
         <feat  type='paradigm  identifier'>
         fr-s-plural</feat>
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>chat</feat>
         <feat type='number'>singular</feat>
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>chats</feat>
         <feat type='number'>plural</feat>
      </struct>
    </struct>
</struct>
```

In the case that spelling variants exist such as *cheik* vs. *cheikh* (Example 2), these are referred to in the lexical entry component by means of the data category /spelling variant/ and an associated pointer to the /key form/ of the related lexical entry. Additional mechanisms such as unification may be envisaged in order to avoid duplication of the lexical information that is independent from this variation (syntactic or semantic information, for example).

Example 2

```
<struct type='lexical entry'>
   <feat type='lemma'>cheikh</feat>
   <feat type='spelling variant'
       target='#cheik_noun'</feat>
   <feat type='grammatical category'>
         common noun</feat>
   <feat type='key form'>cheikh_1</feat>
   <struct type='morphology'>
      …
   </struct>
</struct>
```

Example 3 to Example 5 – *afghan*, 'afghani', used as masculine and feminine noun and as an adjective – shows how data categories, here /gender/, can be used in a flexible way. Depending

on editorial practices, the implementers may, for example, chose to attach this feature to the lexical entry for nouns, and to the inflexion component for adjectives. They will thus consider masculine and feminine forms of a noun as different lexical entries (*afghan_1* vs. *afghane*), while grouping variations for adjectives into one single gender (*afghan_2*).

Example 3

```
<struct type='lexical entry'>
   <feat type='lemma'>afghan</feat>
   <feat type='grammatical category'>
         common noun</feat>
   <feat type='gender'>masculine</feat>
   <feat type='key form'>afghan_1</feat>
   <struct type='morphology'>
      …
      <struct type='inflexion'>
         <feat type='word form'>afghan</feat>
         <feat type='number'>singular</feat>
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>afghans</feat>
         <feat type='number'>plural</feat>
      </struct>
    </struct>
</struct>
```

Example 4

```
<struct type='lexical entry'>
   <feat type='lemma'>afghane</feat>
   <feat type='grammatical category'>
         common noun</feat>
   <feat type='gender'>feminine</feat>
   <feat type='key form'>afghan_2</feat>
   <struct type='morphology'>
      …
      <struct type='inflexion'>
         <feat type='word form'>afghane</feat>
         <feat type='number'>singular</feat>
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>afghanes</feat>
         <feat type='number'>plural</feat>
      </struct>
    </struct>
</struct>
```

Example 5

```
<struct type='lexical entry'>
   <feat type='lemma'>afghan</feat>
   <feat type='grammatical category'>
         adjective</feat>
   <feat type='key form'>afghan_2</feat>
   <struct type='morphology'>
      …
      <struct type='inflexion'>
         <feat type='word form'>afghan</feat>
         …
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>afghans</feat>
         …
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>afghane</feat>
          …
      </struct>
      <struct type='inflexion'>
         <feat type='word form'>afghanes</feat>
         …
      </struct>
    </struct>
```

4.4 Integrating metadata descriptors

One important issue for the management, updating and distribution of lexical databases is the appropriate management of metadata, related either to the identification of data sources or to the characterization of the data.

Our proposal is based on several international initiatives related to the definition of descriptors for language data collections (cf. OLAC[10], IMDI[11]). We currently identify those descriptors that may be relevant for lexical databases, such as the language identifiers (ISO 16620) or the 'roles' defined in OLAC (*depositor, developer, researcher, annotator, sponsor,* etc.). Concerning data characterization, existing standards (ISO 16620) also contain an inventory of possible useful descriptors related to the updating process (*origination date, input date, modification date, approval date, withdrawal date*, etc.).

Additional information should be more specifically related to the morphological extension: One could for example wish to keep track of morpho-syntactic tags (relative to a given tagset, such as *Multext*) currently used to refer to certain inflexions (see Example 6). Other useful metadata would be information about testimony and frequency of inflected forms in corpora, completeness of an inflexion list (relevant for defective verbs such as *pleuvoir* ('to rain') or indication of special usages (diachronic, diatopic or diastratic variation).

Example 6

```
<struct type='inflexion'>
   <brack>
      <feat type='POS tag'>Nms__</feat>
      <feat type='tagset'>Multext</feat>
   </brack>
   <feat type='word form'>chat</feat>
   <feat type='number'>singular</feat>
<struct/>
```

4.5 Morphalou : current state

The basic model described in this paper (apart from inflexion paradigms and metadata descriptors, currently under definition) has been used to build an electronic lexical database of inflected forms for French[12]. It contains 539413 inflected forms distributed over 68075 lemmas, converted from data previously collected at the ATILF laboratory. The whole database is encoded in XML. Since we envisage on-line access and the ability to up-date the data, we devoted particular attention to the interfaces and to documentation. The database is searchable through the web, via a graphical interface or direct XPath queries. The

[10] http://www.language-archives.org/

[11] http://www.mpi.nl/IMDI/

[12] http://loreley.loria.fr/morphalou/

graphical interface allows for lemmatization of a given form and generation of all inflected forms for a given lemma, whereas the XPath requests allows for combining search criteria over any combination of features and strings (for example, all lexical entries for common nouns having an inflected form containing the string *aba*). The next steps are the development of a JAVA API and web services to integrate search results directly into NLP applications and the development of an editorial line for efficient and coherent update of the database. Preliminary updating experiments based on freely accessible morphological databases such as LEFFF and ABU[13] are currently running and reveal the most important problems to be tackled (conversion of the input format, efficient comparison of two XML files, linguistic validation procedures and interfaces for submitted data, fusion of lexical data).

5 References

Antoni-Lay MH., Francopoulo G., Zaysser L. (1994). A generic model for reusable lexicons : *The Genelex project*. Literary and Linguistic Computing.

Dendien J., Pierrel J.-M. (2003). Le TLFi et le logiciel Stella, au centre d'un ensemble de ressources informatisées pour l'étude du français. *Traitement Automatique des Langues*, 44(2), 11-37.

Ide N., Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, 29(2), 167-179.

Ide N., Kilgarriff A., Romary L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000*. Stuttgart, 113-126.

Ide N., Le Maitre J., Véronis J. (1995). Outline of a Model for Lexical Databases. *Current Issues in Computational Linguistics: In Honour of Don Walker*, Pisa, 283-320.

Ide N., Romary L. (2003). Outline of the International Standard Linguistic Annotation Framework. *ACL Workshop on Linguistic Annotation: Getting the Model Right*, Sapporo, 1-5.

ISO 12200:1999, Machine-readable terminology interchange format (MARTIF) — Negotiated interchange.

ISO 12620:1999, Data categories.

ISO 16642:2003, Terminological markup framework.

ISO 1951:1997, Lexicographical symbols and typographical conventions for use in terminography.

ISO 2788, Guidelines for the Establishment and Development of Monolingual Thesauri.

ISO 5964, Guidelines for the Establishment and Development of Multilingual Thesauri.

Mangeot-Lerebours M., Sérasset G., Lafourcade M. (2003). Construction collaborative de base données lexicales multilingues : le projet Papillon. *Traitement Automatique des Langues*, 44(2), 151-176.

Mel'cuk I., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Duculot, Bruxelles.

Romary L. (2001) Towards an Abstract Representation of Terminological Data Collections - the TMF model, TAMA 2001 – Terminology in Advanced Microcomputer Applications, Antwerp.

Saussaure F. de (1974). *Cours de linguistique générale*. Payot, 1974.

Véronis J. (1999). *Multext-Corpora. An annotated corpus for five European languages* [CD-ROM, ELRA-ELDA].

Véronis, J. (2000). Sense tagging: Don't look for the meaning but for the use. *Computational Lexicography and Multimedia Dictionaries (COMLEX'2000)*, Greece.

Véronis J., Ide N. (1992). A feature-based model for lexical databases, *14th International Conference on Computational Linguistics (COLING'92)*, Nantes (France), 588-594.

6 Acknowledgement

The work presented in this paper has received support from the national RNTL/Outilex project, the INRIA corporate action Syntax and the Morphalou project (CPER Lorraine). Many thanks to Monte George, Sue-Ellen Wright and Kornel Bangha for their valuable comments.

7 Appendix : Data categories of the Inflexion component of Morphalou (04/2004)

Component	*Data Category*	*Conceptual Domain*
Entry	/lemma/	String
	/spelling variant/	String
	/grammatical category/	/common noun/
		/verb/
		/adjective/
		/adverb/
		/interjection/
		/onomatope/
		/function word/
Inflexion	/word form/	String
	/mood/	/indicative/
		/conjunctive/
		/conditional/
		/past participle/
		/present participle/
		/infinitive/
	/tense/	/present/
		/imperfect/
		/simple past/
		/future/
	/person/	/first person/
		/second person/
		/third person/
	/gender/	/feminine/
		/masculine/
	/number/	/singular/
		/plural/

[13] http://abu.cnam.fr/DICO/mots-communs.html