HAL
open science

LMF for a selection of African Languages

Chantal Enguehard, Mathieu Mangeot

To cite this version:

Chantal Enguehard, Mathieu Mangeot. LMF for a selection of African Languages. Francopoulo, Gil. LMF: Lexical Markup Framework, theory and practice, Hermès science, pp.8, 2013. hal-00959228

HAL Id: hal-00959228

https://hal.science/hal-00959228

Submitted on 1 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chapter 7[1]

LMF for a selection of African Languages

Chantal Enguehard

LINA, 2 rue de la Houssinière, BP 92208,

44322 Nantes Cedex 03, France

Mathieu Mangeot

GETALP-LIG, 41 rue des Mathématiques, BP 53

F-38041 GRENOBLE CEDEX 9

7.1. Introduction

Electronic resources are scarce regarding less-resourced languages, so it is wise to take published dictionaries and convert them into a standard format usable by automated tools for natural language processing. We introduce the notion of less-resourced languages and then discuss the methodology of conversion that we have defined and implemented. The fourth part presents examples of conversion from the initial published format to the LMF format. The last part describes some difficulties encountered when representing certain information into LMF format.

[1] Chapter written by Chantal Enguehard and Mathieu Mangeot

7.2. Less-resourced languages

7.2.1. Definition

Although a precise inventory of all the existing natural languages is difficult to achieve, there are currently approximately 6,000 languages spoken by humans, but only 200-300 are written. The transition from oral to written is complex and can not be limited to a simple transcription of sounds. It is necessary to conduct studies to achieve a linguistic description of the language in order to determine the transcription system to be used, to choose the most appropriate signs, then to write the spelling and syntactic rules, etc. Finally, languages are more or less well-resourced in terms of their support by tools: adapted keyboard, spell-checker, speech synthesis, machine translation, etc. A classification based on the estimation of the electronic resources and tools defines three classes: well-resourced languages or τ-languages (e.g.: English, French), languages with moderately-resourced languages μ-languages (e.g.: Portuguese or Swedish), and less-resourced languages or π-languages (e.g.: Bambara or Kanuri) [Berment].

The term less-resourced languages covers contrasting situations. We mention here three of them:

— it is the official language of a country as is Irish (or Gaelic Irish) in Ireland.

— it is a language without official status, that became a regional language: for example Basque and Breton in France; Ladin in Italy, Cornish in the United Kingdom.

— it is a national language of a country whose official language (used at school, or to write the laws) is different and often comes from a former colonizer state [Calvet]. This is the case of African languages on which we have worked and that are spoken in Niger, Mali and Burkina Faso. In these three countries, the official language is French.

7.2.2. Socio-economic context

We focus on five African languages: Bambara, Kanuri, Hausa, Zarma and Tamajaq. They are less-resourced languages which socio-economic context is characterized by limited resources:

— there are few linguists having a less-resourced language as their mother tongue and exercising their professional activity on that language.

— the budget for the development of linguistic resources is low.

The governmental investment dedicated to language planning and, in particular, the development of electronic language resources are therefore very limited. The few studies that are conducted are characterized by a discontinuity in the time and spatial spread, which affects their sustainability and reuse [Streiter].

7.2.3. linguistic resources

Because of the scarcity of linguistic research, descriptions of these languages are incomplete and many questions remains. There are few dictionaries and they are generally not made by professional lexicographers. In addition, it is unusual to revise and make corrections on a published dictionary. This contrasts sharply with the published dictionaries of well-resourced languages like French or English. For example, Larousse or Harrap's are firms employing dozens of professionals who regularly review their dictionaries for several decades. Therefore, the dictionaries on which we worked contain numerous errors or incompleteness and are likely to evolve.

7.2.4. Building electronic lexical resources

Developing lexical resources *ex nihilo* requires large budgets, qualified and available professionals, and the ability to lead a project for several years. These conditions can not be met in many countries. However, there are some published dictionaries (often bilingual) that can be exploited to build a first version of an electronic resource in a few weeks and at low cost.

7.2.4.1. Dictionaries written by a single author

Many of the dictionaries written by a single author are bilingual because their author, originally from another language, aims to promote a language. Some were written by clerics in charge of people evangelism in colonized countries ("pères blancs" in Africa, Portuguese Jesuits in Asia). For example, we worked on the Bambara-French dictionary of Father Charles Bailleul [Bailleul]. There are also dictionaries developed by literate people, often linguists, wishing to serve their mother tongue. This is the case of elementary Hausa-French dictionary written by Abdou Minjinguini [Minjinguini] and the monolingual zarma dictionary written by Issoufi Alzouma Oumarou [Oumarou].

7.2.4.2. Dictionaries built by projects

Dictionaries built by projects have several authors. The group of authors usually defines some principles about the structure and the definition of closed lists of values such as grammatical classes.

For example, we worked on dictionaries written in five national languages West-Africa languages for the DiLAF project [Enguehard b].

7.3. From published dictionaries to LMF

7.3.1. Objectives

Our goal is to convert published dictionaries to make them available to the natural language processing (NLP) scientific community. We choose LMF as final format because it is an ISO standard that favors the re-utilization of the data (this is a key-point when working on less-resourced languages as stated also by the RELISH project [Windhouwer]). The actual conversion of several dictionaries with thousands of entries constitutes an experimentation in order to test the operationality of this format and, optionally, to suggest improvements.

7.3.2. Methodology

Lexicographers and NLP experts must collaborate to convert a published dictionary into a structured electronic format. Thus we define the tasks performed by each collaborator of such a project.

The conversion methodology we defined proceeds in several steps and requires successive transformations of the published dictionary to three different XML files called *copy*, *pivot* and *target* formats. We also take into account the fact that lexicographers will revise and develop the produced resources.

The *copy format* is a structural copy of the published dictionary in a valid XML format. The transformation of the published dictionary to the *copy* format is performed by lexicographers[2] with the support of NLP experts. This step requires solving many problems, including the conversion of special characters to Unicode,

[2] As each dictionary includes thousands of entries that would be tedious to manually tag, the conversion methodology includes the training of lexicographers in handling regular expressions so that they are able to automate themselves a part of this task.

the identification of each information part, the definition of a set of markup tags and finally the explicit tagging of information by placing tags [Enguehard b]. When a first valid version of the *copy* format is available, various checks are performed using simple programs (counting the number of occurrences of each tag, checking the embeddedness of the markups, counting the number of closed lists values like parts of speech, etc.) and errors are reported to lexicographers who can make the corrections. The use of a CSS stylesheet associated with the display of the *copy* format also allows a browser to introduce facilities for valuable consultation: the relationship of synonymy and antonymy are represented by href links, which allows to easily control their consistency. Finally, the markup tag names are often expressed in the language of the dictionary which facilitates the appropriateness of the new format. The *copy* format does not alter the structure of the original format but improves readability by explicitly labelling every part of information.

The *pivot format* respects the normative core of LMF. It is obtained by structural changes of the *copy* format by applying an XSLT[3] program. It may be necessary, for example, to change the place of a morphological information that was described in a semantic block. Most important changes may be necessary like the merge of two lexical entries, or the separation of a lexical entry with two semantic blocks into two lexical entries with a single semantic block. These treatments are performed by perl programs. Markup tag names are preserved from the *copy* format.

The *target format* follows the syntax of the informative part of the LMF standard. It is obtained by processing the *pivot* format with an XSLT program. As the *pivot* format meets the standard LMF format, the transformations from the *pivot* format towards the *target* format are limited to changing the name of an element, to add an additional level element with a child, and to convert a text node in an attribute value (see examples below). NLP experts develop conversion programs to process the transformation from *copy* format to *pivot* format, and from *pivot* format to *target* format. When they conceive these programs they get the opportunity to detect new errors and inconsistencies that are reported for subsequent corrections.

Finally, the *copy* format dictionary is aimed to disappear in favour of the *pivot* format dictionary. It can be easily understood by lexicographers as they themselves chose the markup tag names. On the contrary, the *target* format dictionary is more difficult to understand and to modify. The *pivot* format dictionary can be then uploaded on an online lexical resources management platform such as Jibiki [Mangeot c] in order to be readable and editable online by lexicographers who will be able to correct and enhance it directly (by adding new lexical entries, various information, translations, examples, etc..). It would then be easy to generate a new *target* format dictionary by processing again the adequate program on the *pivot* format dictionary.

---

[3] eXtensible Stylesheet Language Transformations.

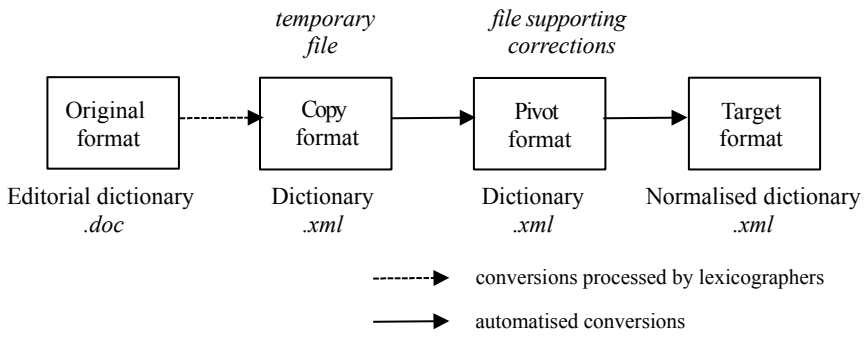

Figure 7.1. *Conversion process*

7.4. Illustrations

Here are some examples of the implementation of the above methodology.

7.4.1. Definition of the copy format

The *copy* format defined by the lexicographers is close to the initial structure of the dictionary. This is to make explicit the nature of the information: definition, lexical label, phonetic, synonyms, French equivalent, etc...

Four of the five dictionaries on which we have worked are intended for an audience of students in elementary courses. They are written primarily in the dictionary language, only the presence of one or several French equivalents for each input gives them a bilingual character. In these dictionaries, lexical categories are expressed in the language of the dictionary (see Table 1) and may vary according to the characteristics of the language. For example, in Kanuri, the class of each verb is specified; in Zarma most verbs are defined as transitive or intransitive.

Language	Lexical Category	Abbreviation	English equivalent
Hausa	kamantau	k.	adjective
Kanuri	alama njoma	alnj	adjective
Zarma	taka sifa	tsif.	adjective

Haussa	suna	s.	noun
Kanuri	cu	cu.	noun
Tamajaq	Isən tənte	sn. tnt.	feminine noun
Tamajaq	isən yey	sn. yy.	masculine noun
Zarma	ma	m.	noun
Kanuri	nufatan yawa	nuy.	quantity adverb
Zarma	dimma teebare	dteeb.	quantity adverb
Hausa	amsa kama	ak.	ideophone
Kanuri	manda coktuwuma	cok.	ideophone
Zarma	teeraci kubandiko	teerk.	transitive verb
Zarma	teeraci kubandi si	teerks.	intransitive verb
Kanuri	kalma kəndoye	kkye.	impersonal verb
Kanuri	kalma kəndoye 2	kkye2.	2nd class verb

Table 1. *Examples of parts-of-speech*

The information contained in lexical entries are different according to the dictionary because of the language represented or the choices made by the dictionary authors. For instance, in the Tamajaq dictionary, an annexation state is indicated for some lexical entries, while this information does not exist in other languages, and phonetics is not specified; in the Hausa dictionary colloquial expressions and many variants spelling are reported. Some examples of markup names are presented in Table 2. As the names of lexical categories were written in the language of each dictionary, it seemed natural to define also markup names in the same language.

Language	Tag name	English equivalent
Hausa	ma_ana	definition
Kanuri	maana	definition
Tamajaq	almaɣna	definition
Zarma	feeriji	definition
Kanuri	maana_tiloa	synonym
Tamajaq	anammelu	synonym
Zarma	himacare	synonym
Kanuri	bowodu	phonetic
Zarma	ciiyaŋ	phonetic

Hausa	makwatanci	French equivalent
Kanuri	kalakta	French equivalent
Tamajaq	təfaransist	French equivalent
Zarma	bareyaŋ	French equivalent
Hausa	salon_magana	phrase
Hausa	yare	variant
Kanuri	kənyakkuye_ tilola	third pers. sing. of 2nd class verbs
Tamajaq	əsəfsəs	annexion state

Table 2. *Examples of tag names (copy and pivot formats)*

7.4.2. From original format to copy format

Two examples, one Kanuri, the other Tamajaq can illustrate this first step.

The figures 7.2 and 7.3 show two entries. In their original version, the special characters initially entered with an artisanal font[4] are not readable. In the Unicode version these special characters have been transformed to meet Unicode.

Bannadu2 [baˆnnaˆðu˜] *kkye2.* Diwiro yal alamdu. *Gənanjun bannaje, ku tadanju rakce kəlanju rojiwawo. Mt.:* låˆnðu˜.*[Fa.:* e˜duquer(mal)*]*

bannadu2 [bànnàḍú] *kkye2.* Diwiro yal alamdu. *Gənanjun bannaje, ku tadanju rakce kəlanju rojiwawo. Mt.:* lạ̀nḍú.*[Fa.:* éduquer(mal)*]*

Figure 7.2. *Kanuri lexical entry* bannadu *(2) in initial published format then in Unicode format*

äœaruf *sny. pardon* ⊠ Agamay n pkpnni dpffpr erk ärät. Musa as ypwät empji-net dpffpr pnki ypgmäy dä£-as äœaruf. *An:* tptubt. *Sf:* ä . *Gt:* äœuruf. *TW:* tpsureft

ăṣaruf cat=sny. pardon ‣ Agamay n əkənni dəffər erk ărăt. Musa as yəwăt eməji-net dəffər ənki yəgmăy dăɣ-as ăṣaruf. *An:* tətubt. *Sf:* ă . *Gt:* ăṣuruf. *TW:* təsureft.

Figure 7.3. *Tamajaq lexical entry ăṣaruf in initial published format then in Unicode format*

[4] Lots of artisanal fonts have been created before Unicode when there were no code for special characters. In these fonts, the glyphs of some unused characters are replaced by the glyphs of a special character [Enguehard a].

The transformation of the Tamajaq special characters was specially complicated because the artisanal font that was used changes the glyphs of the letter 'p' (missing in the Tamajaq alphabet [République du Niger]) into the glyphs of the letter 'ə'. As the character 'p' is susceptible to occur in French equivalents, using a regular expression was essential for rapid replacement of 'p' in 'ə' only in parts written in Tamajaq.

The same lexical entries, transformed into the *copy* format are shown in Figures 7.4 and 7.5.

```
<article>
    <kalma lamba="2">bannadu</kalma>                                  lexical entry number 2
    <bowodu>[bànnàḏú]</bowodu>                                        phonetic
    <naptu_curo_nahauyen>kkye3.</naptu_curo_nahauyen>                 part of speech
    <maana>Diwiro yal alamdu.</maana>                                 definition
    <misal>                                                           example
        <version təlam="kau">Gənanjun bannaje, ku tadanju             example in Kanuri
rakce kəlanju rojiwawo.</version>
        <version təlam="fa">Durant son jeune âge il l'a mal           equivalent of the
éduqué, aujourd'hui son fils n'arrive pas à se prendre en             example in French
charge.</version>
    </misal>                                                          synonym
    <maana_tiloa>lə̀nḏú</maana_tiloa>                                 equivalent in French
    <kalakta təlam="fa">éduquer (mal)</kalakta>
</article>
```

Figure 7.4. *Kanuri lexical entry* ***bannadu*** *(2) in copy format*

The lexicographer that transforms ***bannadu*** corrected the part of speech in kkey3 (third class verb) and added a French equivalent of the example.

```
<albab>                                                               article
    <təzugəst>ăṣaruf</təzugəst>                                       lexical entry
    <təmuṣt>sn. yy.</təmuṣt>                                          part of speech
    <təfaransist>pardon</təfaransist>                                 equivalent in French
    <almaɣna>Agamay n əkənni dəffər erk ărăt.</almaɣna>               definition
    <əlmisal>Musa as yəwăt eməji-net dəffər ənki yəgmăy dăɣ-as       example in Tamajaq
ăṣaruf.</əlmisal>
    <anammelu>tətubt.</anammelu>                                      synonym
    <əsəfsəs>ă.</əsəfsəs>                                             annexion state
    <igət>ăṣuruf.</igət>                                              plural
    <tastəqW>təsureft.</tastəqW>                                      Tawəlləmət variant
</albab>
```

Figure 7.5. *Tamajaq lexical entry ăṣaruf in copy format*

7.4.3. From copy format to pivot format

The lexical entries in Figures 7.4 and 7.5 are automatically transformed into the *pivot* format in which appear explicitly a lemma block (with its spelling and pronunciation) and a semantic block.

```
<article id="bannadu2">                                                  article with identifier
    <lemme>
        <kalma lamba="2">bannadu</kalma>                                 lexical entry number 2
        <bowodu>bànnàḍú</bowodu>                                         phonetic
    </lemme>
    <naptu_curo_nahauyen>kkye3.</naptu_curo_nahauyen>                    part of speech
    <bloc-semantique>
        <kalakta təlam="fa">éduquer(mal)</kalakta>                       equivalent in French
        <maana>Diwiro yal alamdu.</maana>                                definition
        <misal>                                                          example
          <version təlam="kau">Gənanjun bannaje, ku tadanju              example in Kanuri
rakce kəlanju rojiwawo.</version>
          <version təlam="fa">Durant son jeune âge il l'a mal éduqué,    equivalent of the
aujourd'hui son fils n'arrive pas à se prendre en charge.</version>      example in French
        </misal>
        <maana_tiloa>lànḍú</maana_tiloa>
    </bloc-semantique>                                                   synonym
</article>
```

Figure 7.6. *Kanuri lexical entry* bannadu *(2) in pivot format*

Adjustments can be carried out directly by the lexicographers in this format. In the Kanuri example, the synonym is designated by its phonetic and should be replaced by an article identifier.

```
<albab id="ăṣaruf">                                                      article with identifier
    <lemme>
        <təzugəst>ăṣaruf</təzugəst>                                      lexical entry
    </lemme>
    <təmuṣt>sn. yy.</təmuṣt>                                             part of speech
    <əsəfsəs>ăṣaruf</əsəfsəs>                                            annexion state
    <igət>ăṣuruf</igət>                                                  plural
    <tastəqW>təsureft</tastəqW>                                          Tawəlləmət variant
    <bloc-semantique>
        <təfaransist>pardon</təfaransist>                                equivalent in French
        <almaɣna>Agamay n əkənni dəffər erk ărăt.</almaɣna>              definition
        <əlmisal>Musa as yəwăt eməji-net dəffər ənki yəgmăy dăɣ-        example in Tamajaq
as ăṣaruf.</əlmisal>
        <anammelu>tətubt</anammelu>                                      synonymous
    </bloc-semantique>
</albab>
```

Figure 7.7. *Tamajaq lexical entry ăṣaruf in pivot format*

In the Tamajaq example, and additional program automatically replaced the notation of the annexion state by a single vowel *'ă'* (meaning that the first vowel of the lemma must be replaced by *'ă'* to determine the annexion state) by the new form of the lemma *ăṣaruf*.

7.4.4. From pivot format to target format

The conversion from the *pivot* format to the *target* format is automatically processed by XSLT programs. There is one program per dictionary.

Examples of Figure 7.6 and 7.7 are automatically transformed into target format that meets the syntax of the informative part of the LMF standard.

```
<LexicalEntry id="bannadu2">                                        article with identifier
    <Lemma>
        <feat att="writtenForm" val="bannadu"/>                     written form
        <feat att="phoneticForm" val="bànnàḍú"/>                    phonetic
    </Lemma>
    <feat att="partOfSpeech" val="kkye3."/>                         part of speech
    <Sense id="1">
        <Equivalent>
            <feat att="language" val="fra"/>                        equivalent in French
            <feat att="writtenForm" val="éduquer(mal)"/>
        </Equivalent>
        <Definition>
            <feat att="writtenForm" val="Diwiro yal alamdu."/>      definition
        </Definition>
        <Context>                                                   example
            <TextRepresentation>
                <feat att="language" val="kau"/>                    example in Kanuri
                <feat att="writtenForm" val="Gənanjun bannaje,
ku tadanju rakce kəlanju rojiwawo."/></TextRepresentation>
            <TextRepresentation>
                <feat att="language" val="fra"/>
                <feat att="writtenForm" val="Durant son jeune âge il  equivalent of the
l'a mal éduqué, aujourd'hui son fils n'arrive pas à se prendre en charge."/>  example in French
            </TextRepresentation>
        </Context>
        <SenseRelation targets="lạ̀nḍú">
            <feat att="type" val="synonym"/>
        </SenseRelation>
    </Sense>                                                        synonymous
</LexicalEntry>
```

Figure 7.8. *Kanuri lexical entry* bannadu *(2) in target format*

```
<LexicalEntry id="ăṣaruf">                                          article with identifier
    <Lemma>
        <feat att="writtenForm" val="ăṣaruf"/>                      written form
    </Lemma>
    <feat att="partOfSpeech" val="sn. yy."/>                        part of speech
    < W o r d F o r m        w r i t t e n F o r m = " ă ṣ a r u f "  annexion state
contextualVariation="annexion"/>                                    plural
    < W o r d F o r m        w r i t t e n F o r m = " ă ṣ u r u f "
grammaticalNumber="plural"/>                                        Tawəlləmət variant
    <Equivalent>
        <feat att="language" val="ttq"/>
        <feat att="writtenForm" val="təsureft"/>
    </Equivalent>
    <Sense id="1">
        <Equivalent>                                                equivalent in French
            <feat att="language" val="fra"/>
            <feat att="writtenForm" val="pardon"/>
        </Equivalent>                                               definition
        <Definition>
            <feat att="writtenForm" val="Agamay n əkənni dəffər
erk ărăt."/>
        </Definition>                                               example in Tamajaq
        <Context>
            <TextRepresentation>
                <feat att=language="tmh"/>
                <feat att="writtenForm" val="Musa as yəwăt eməji-
net dəffər ənki yəgmăy dăɣ-as ăṣaruf."/>                           synonym
            </TextRepresentation>
        </Context>
        <SenseRelation targets="tətubt">
            <feat att="type" val="synonym"/>
        </SenseRelation>
    </Sense>
</LexicalEntry>
```

Figure 7.9. *Tamajaq lexical entry ăṣaruf in target format*

7.5. Difficulties and proposals

The actual conversion of dictionaries into LMF format was the opportunity to meet the difficulties that we detail below. We also include solutions or elements of reflection.

7.5.1 Data category

7.5.1.1 Language names and associated ISO 639-3 codes

It may be that difficult to identify a language and the associated ISO 639-3 code. Currently, ISO refers to the ethnologue[5] website that relies on a small number of studies mainly carried out by the staff of the Summer Institute of Linguistics (SIL). For example, the page dedicated to the language Tamahaq (for Tamajaq) includes only one bibliographic reference (an article on the music of the Tuareg) and cites no extract of text. However, a significant number of academic research have been conducted on this language and should be included in the bibliography. Thus, we suggest to enrich this languages catalog with some academic research articles.

7.5.1.2 Parts of speech list

We encountered parts of speech that are not included in the Parts of speech list of the ISO Data Category Registry (DCR)[6]. For example, the part of speech "ideophone" appears in the list of parts of speech in Hausa and Kanuri dictionaries. It is also used in Somali [Assowe]. For this latter language, there are also other parts of speech ("verbal affix", "focus marker", "sentence marker", etc.) that are missing in the ISO list. Zarma language does not distinguish between masculine and feminine, but distinguish between the definite and indefinite, etc. Thus, it appears necessary to enrich this parts of speech list or to allow a modular definition of this list with a sublist for each language.

7.5.2 LMF structure

7.5.2.1 Absence of macrostructure

The LMF standard represents a lexical resource in a unique file (see Figure 7.10). Thus, it is not possible to represent complex dictionaries macrostructures and their links between volumes, such as Papillon pivot structure [Mangeot b] or PIVAX structure [Mangeot d]. In [Mangeot a], we define a volume as an alphabetically ordered set of entries of the same language and a dictionary as a set of volumes. An entry of one volume can be linked to an entry of another volume.

```
<LexicalResource>
    <GlobalInformation entrySource="Prolex"/>
    <Lexicon languageSymbol="fra">
```

[5] http://www.ethnologue.com

[6] http://www.isocat.org/rest/dcs/119.html

```
        <LexicalEntry partOfSpeech="noun">...</<LexicalEntry>
         <LexicalEntry partOfSpeech="noun">...</<LexicalEntry>
                …..
            <LexicalEntry partOfSpeech="noun">...</LexicalEntry>
      </Lexicon>
</LexicalResource>
```

Figure 7.10. *The beginning of the ProLex lexicon in LMF format [Maurel]*

7.5.2.2 Objects of different nature at the same level

We think that, in order to be clearly understandable, an XML format should avoid to put objects of different nature at the same indentation level. The siblings of an element must be of the same nature. The LMF format does not respect this principle. In figure 7.10, the object <GlobalInformation> which is a meta-information about the lexicon is a sibling of the object <Lexicon> which is the resource itself.

7.5.2.3 Informative part of LMF

— Free text

In XML, it is customary to include items from closed lists as attribute values, and frame the free texts by markup tags. This general principle is not respected in the informative part of LMF since all the information is stored in textual attributes. This choice has the effect of prohibiting the minimal information display via a browser for example.

— Examples of use representation

The dictionaries we worked on being bilingual, we faced the problem of representing information in different languages. In the general structure of an article, the lexical entry is clearly distinguished from its equivalents in other languages. In contrast, the representation of the equivalent of an example in the same form as the example itself, only by specifying another language, does not distinguish the example itself from its translation (see Figure 7.11). LMF offers the possibility to represent a translation with the "Multilingual notations extension" mechanism which makes the assumption that each equivalent exists in the dictionary of its language. But this is not always the case. For instance, the French equivalent of the Kanuri lexical entry *bannadu* is *éduquer (mal)* which is not a French lexical entry. This phenomena is common when two languages concern cultures with differences in food, religion, cooking, dressing, etc. and because there are distinct linguistic structures[7]. Thus we decide to add a convention to read the occurrences of

[7] Here are some examples issued from the Kanuri-French dictionary (with a translation in English): *adinnamdu - aller vers l'est* (to go towards east); *albayi - pochette touareg* (Tuareg bag); *asar - troisième prière* (third pray); *bare - il ne faut pas* (it is not allowed to);

TestRepresentation: “When the language of the TextRepresentation is different from the language of the dictionary, the TextRepresentation is a translation of the TextRepresentation expressed in the language of the dictionary”.

```
<Context>
    <TextRepresentation>
        <feat att="language" val="kau"/>
        <feat att="writtenForm" val="Gənanjun bannaje, ku tadanju rakce kəlanju rojiwawo."/>
    </TextRepresentation>
    <TextRepresentation>
        <feat att="language" val="fra"/>
        <feat att="writtenForm" val="Durant son jeune âge il l'a mal éduqué, aujourd'hui son fils
n'arrive pas à se prendre en charge."/>
    </TextRepresentation>
</Context>
```

Figure 7.11. *A usage example and its equivalent in another language*

Finally, an example may need to be explained, a simple translation being not enough to make it understandable. This is the case of many Bambara proverbs in the dictionary [Bailleul]. The author has often included a loan translation and an explanation giving the meaning of the proverb. We choose to simply represent such an explanation by using the “explanation” category (see an example in Figure 7.12).

proverb: jalaki tɛ baji la
literal translation: on ne condamne pas l'eau du fleuve[8]
explanation: c'est de ta propre faute ![9]

```
<Context>
    <TextRepresentation>
        <feat att="language" val="bam"/>
        <feat att="writtenForm" val="jalaki tɛ baji la."/>
    </TextRepresentation>
    <TextRepresentation>
        <feat att="language" val="fra"/>
        <feat att="writtenForm" val="on ne condamne pas l'eau du fleuve."/>
        <feat att="explanation" val="c'est de ta propre faute !"/>
    </TextRepresentation>
</Context>
```

Figure 7.12. *Example of a proverb in Bambara and its representation*

— orthographic variants

basi - mets à base de mil (dish made with millet).

[8] In English: River water can not be condemned.

[9] In English: It is your fault!

Less-resourced languages are sometimes written for a short time and orthographic forms may vary. Also, some words have different spellings. They are neither synonyms nor geographical variants.

7.5.3 Adding annotations

Dictionaries on which we have worked are incomplete, often being the first version, they contain errors. In addition, their use by NLP researchers should raise new linguistic questions. Thus, it appears necessary to provide the ability to add annotations that could be collected later and addressed to the concerned linguists community. Annotations about inaccuracies of the dictionary can easily lead a linguist to make new corrections (e.g.: an entry marked as synonymous has three meanings, or a synonymous is missing in the dictionary). Annotations about more fundamental problems could feed the thoughts of the linguists community (e.g.: some words seem to hesitate between two lexical categories and are labeled with both).

7.6. Conclusion

The actual conversion of multiple published dictionaries into the LMF format has put into practice the DiLAF methodology of conversion we defined. This methodology is suitable for less-resourced languages and integrates the limitations in working time and financial resources. The final conversion into LMF allows to distinguish limitations regarding the completeness of the list of parts of speech and the consequences of structuring information in the form of attribute values. We have identified some desirable developments for the future as an opportunity to enrich the list of parts of speech or the definition of new markup tags to annotate dictionaries evolution. After practicing the LMF standard [LMF] for encoding our dictionaries, we think that LMF would gain in usability with a simple examplified tutorial of how to encode an existing resource into LMF.

Bibliography

[Assowe] Assowe, Houssein Ahmed. 2011 Etude linguistique et approches de l'étiquetage morphosyntaxique du somali. Mémoire de Master 2. Université Michel de Montaigne Bordeaux 3.

[Bailleul] Bailleul, C. 1996. Dictionnaire bambara-français, édition 1996.

[Berment] Berment, V. 2004. Méthodes pour informatiser des langues et des groupes de langues peu dotées. Ph.D. thesis, Université Joseph Fourier.

[Calvet] Calvet L.-J. 1996. Les politiques linguistiques. Paris, PUF.

[Enguehard a] Enguehard, C. 2009. Les langues d'Afrique de l'Ouest : de l'imprimante au traitement automatique des langues . Sciences et Techniques du Langage, 6, p.29-50,.

[Enguehard b] Enguehard, C., Kané, S., Mangeot, M., Modi, I., Sanogo, M.L., Issouf Modi. 2012. Vers l'informatisation de quelques langues d'Afrique de l'Ouest, JEP-TALN-RECITAL 2012, Atelier TALAf 2012 : Traitement Automatique des Langues Africaines, p. 27-40.

[LMF] Lexical Markup Framework. 2008. ISO/TC 37/SC 4 N453 (N330 Rev.16) ISO FDIS 24613:2008.

[Mangeot a] Mangeot, M. 2001. Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Ph.D. thesis, Université Joseph Fourier, 280 p.

[Mangeot b] Mangeot, M. & Kuroda, K. 2003. Interlinguistic Divergences in Papillon Multilingual Dictionary. Proc. of ASIALEX 2003, Meikai University, Urayasu, Chiba, Japan, 27-29 August 2003, p. 156-162.

[Mangeot c] Mangeot, M. & Chalvin, A. 2006. Dictionary Building with the Jibiki Platform: the GDEF case. Proc. of LREC 2006, Genoa, Italy, 23-25 May 2006, p. 1666-1669.

[Mangeot d] Mangeot, M. & Hong-Thai Nguyen, H-T. 2009. Building lexical resources: towards programmable contributive platforms Proc. IEEE-RIVF 2009, DaNang, VietNam, 14-16 July 2009, Vol 1/1, p. 84-92.

[Maurel] Maurel D. & Bouchou B. 2013. Prolmf A multilingual dictionary of proper nouns and their relations. in this volume.

[Minjinguini] Minjinguini, A. 2003. Ƙaramin ƙamus, na hausa zuwa faransanci - Dictionnaire élémentaire hausa-français.

[Umaru] Umaru, I. A. 1997. Zarma ciine - kaamuusu kayna. Editions Alpha.

[République du Niger] République du Niger. 1999. Alphabet tamajaq, arrêté 214-99.

[Streiter] Streiter, O., Scannell, K. P. & Stuflesser, M. 2006. Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers. Machine Translation, vol. 20 n°3, March.

[Windhouwer] Windhouwer, M., Petro, J., Nevskaya, I., Drude, S., Aristar-Dry, H. & Gippert, J. 2013. Creating a serialization of LMF: the experience of the RELISH project. in this volume.

Acknowledgements

The DiLAF project to convert published dictionaries to LMF format is funded by the Fonds Francophone des Inforoutes of the Organisation Internationale de la Francophonie.