Dictionaries for Under-Resourced Languages: from Published Files to Standardized Resources Available on the Web

Mathieu Mangeot, Chantal Enguehard

Dictionaries for Under-Resourced Languages: from Published Files to Standardized Resources Available on the Web

MATHIEU MANGEOT[1], CHANTAL ENGUEHARD[2]
[1]*GETALP-LIG laboratory, 38041 Grenoble, France (mathieu.mangeot@imag.fr);*
[2]*LINA laboratory, 44000 Nantes, France (chantal.enguehard@univ-nantes.fr)*

Abstract.

Most work in the field of natural language processing focuses on well-resourced languages. However, much remains to be done on under-resourced ones: there are few dictionaries, parsers, etc. Nevertheless, when published dictionaries are available, it is sometimes possible to find the data files used to print the dictionary (usually in Word format). A conversion process can then be applied to these files in order to obtain standardized XML lexical data. Attention must be paid to specific problems such as a lack of standardization in the alphabets or the use of hacked fonts for displaying specific characters. Next, the standardized XML data can be imported into an online lexical resources management platform. It is then available online for lookup and editing. A final step can also be performed to automatically export the data into interchange formats such as Lexical Markup Framework or lemon in order to produce linked data.

Keywords. Dictionary; lexical database; Jibiki platform; XML; LMF; Bambara; Khmer; Wolof; Niger; National Language

1 Introduction

In the field of natural language processing (NLP), most work focuses on well-resourced languages (English, French, German, Japanese, etc.): there are lexicons, dictionaries, corpora, sophisticated tools such as lemmatizers or parsers, etc. and one

can find a lot of online resources. However, much remains to be done on under-resourced languages: there are few dictionaries, parsers, etc. (and their quality is often very low) even when some have a large number of speakers (there are, for example, 250 million Bengalis, 30 million Haoussas).

Being a speaker of a poorly endowed language means limited or no access to the linguistic resources of this language: dictionaries, but also textbooks, literary, media, etc. These unmet needs affect many aspects of life: web, education, health, culture, etc. (Osborn, 2011). This difficult access to the resources compromises the capacity to express one's thoughts: freedom of expression is at stake. In addition, it negatively affects development with respect to several dimensions: the economy, culture, technology, public health, etc.

NLP also suffers from this shortage because it is very difficult to develop NLP tools or research without any linguistic resources or with limited resources. Under-resourced languages remain *terra incognita* for NLP.

The lack of quality and sustainable resources for under-resourced languages is a bottleneck that creates a vicious circle: the lack of resources discourages research in NLP, prevents the development of populations (including in education), resulting in a low level of education and a lack of "language workers" (linguists, lexicologists, novelists, reporters, etc.). Finally there are not enough linguists and lexicologists to produce good quality resources...

The United Nations Educational, Scientific and Cultural Organization (UNESCO) has mentioned several times this dimension and the richness of linguistic diversity. It has called upon Member States "*to promote the preservation and protection of all languages used by peoples of the world*". In 2005, during the conference organized in Bamako (Mali), it added that language is a critical factor in the ability to communicate.

Finally, the access to linguistic resources appears as an ethical issue. Promoting and creating linguistics resources for under-resourced languages meets this need.

Thus, we consider that there are two major reasons for NLP to develop research and tools which help in the creation and promotion of resources for under-resourced languages: the human development of the populations and the scientific issues of NLP.

When printed dictionaries for these languages are available, it is sometimes possible to find the data files used to print them (usually in Word format). A conversion process can then be applied to these files in order to obtain standardized XML lexical data that is next put online for lookup and editing or exported into interchange formats such as Lexical Markup Framework (LMF) or lemon, in order to produce linked data.

If the data files cannot be found either because they were lost or non-existent (in the case of old dictionaries), it is also possible to perform an optical character recognition process on the scanned files from the printed books. The result of the OCR process is then saved into Word or ODF format and processed as with the previous data files. In order to cope with the remaining OCR errors, it is possible to use a lemmatiser if it exists for the processed language, like in the Japanese-French Jibiki.fr project[1] (Mangeot, 2016).

In the first section of this article, specific issues are tackled concerning available dictionaries for under-resourced languages, and the resources to which the conversion methodology was applied are presented. The following section details technical problems such as the lack of standardization in the alphabets or the misuse of unicode characters with hacked fonts. The third section focuses on the conversion process of a published dictionary into a standardized XML file with a discussion about the choice of the standard format, the description of the process itself and the different steps it involves. The last section describes the use of Jibiki, an online lexical resources management platform for lookup and editing of the previously converted dictionaries.

2 Dictionaries for under-resourced languages

2.1 SPECIFIC ISSUES OF UNDER-RESOURCED LANGUAGES

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 (eg: English, French), moderately-resourced languages or μ-languages (eg: Portuguese or Swedish), and under-resourced languages or π-languages (eg: Bambara, Kanuri or Khmer) (Berment, 2004).

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

— it is the official language of a country, as is Irish (or Irish Gaelic) in Ireland. It is also the case for languages spoken by the majority of the population such as Khmer in Cambodia.

— 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, 1996). 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.

The dictionaries presented in this article concern Khmer, the official language of Cambodia and also eight African languages from countries where French is the official language: Bambara, Fulfulde, Hausa, Kanuri, Tamajaq, Wolof and Zarma. They are under-resourced languages whose socio-economic context is characterized by limited resources:

— there are few linguists who have an under-resourced language as their mother tongue and who exercise their professional activity in 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 is 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, 2006).

Because of the scarcity of linguistic research, descriptions of these languages are incomplete and many questions remain.

Developing lexical resources from scratch requires substantial budgets, qualified and available people, and the ability to lead a project for several years, conditions that cannot be met in many countries where there are few dictionaries which are generally not compiled by professional lexicographers. Therefore, the dictionaries on which we have worked contain numerous errors or incompleteness and are likely to evolve.

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 over several decades.

However, there are some published dictionaries (often bilingual) which can be reused to achieve in only a few weeks, at low cost, a first version of an electronic resource. Collecting the digital files constitutes an important advance. However, in their absence, the dictionary can be keyed in again when only a printed copy has been collected.

Whatever its format (electronic or print), a dictionary represents a sum of important knowledge that can be recovered and reused. In all cases the authors or publisher of the initial dictionary should be involved in the project in order to obtain their agreement that the lexical resource that will be produced can be widely distributed in electronic form, visible on the Internet.

2.2 DICTIONARIES WRITTEN BY A SINGLE AUTHOR

Many of the dictionaries written by a single author are bilingual because their author, originally from a τ-language, aims to promote a π-language. Some were written by clerics in charge of the evangelism of populations in colonized countries ("pères blancs" in Africa, Portuguese Jesuits in Asia).

There are also dictionaries developed by literate people, often linguists, wishing to serve their mother tongue. This is the case of the elementary Hausa-French dictionary written by Abdou Minjinguini (Minjinguini, 2003) and the monolingual Zarma dictionary written by Issoufi Alzouma Oumarou (Oumarou, 1997).

The conversion methodology (described in section 4) has been applied successfully to the following two dictionaries.

2.2.1 The French-Khmer dictionary

abondant, e (fruits, riz…) — (pluie) (trempé-humide)	(daːel) sɑmboː (daːel) coːk-coam
abonnement (magazine) — (téléphone)	ciːəw-prāː cam kaː- bɑŋ-seːvaː
abonner (sur pied-nom (s'inscrire)-commercer)	coh-chmuəh-ciːəw
abonner (s') (à un magazine) — (au téléphone)	ciːəw-prāː cam bɑŋ-seːvaː
abord (adv.) (d'—)	mun-dɑmboːŋ / daːəm-laːəj
aborder (accoster) — (qqn) (appeler-arrêter) — (commencer-discuter-sur-sujet-un)	coːl-cāː t / coːl-cət haw-bɑŋchop phdaːəm-coːcɛːk-ɑmpɨː paŋə-haː- muːəj
aboutir (arriver à destination, déboucher) — (avoir-résultat) — (devenir) (aller-être)	təːw-dɑl miːən-loːttəphāː l təːw-ciːə

Figure 1: Excerpt from the French-Khmer dictionary in Word format

The French-Khmer dictionary[2] project (Richer et al., 2007), started in the late 1990s, was completed in 2006 by a small group of computer scientists gathered in the "Pays Perdu" NPO created by Denis Richer, a French ethnolinguist established in Siem Reap (Cambodia). This first version of the dictionary was published in spring 2007 and has 13,249 entries. Each entry is composed of a French headword, in some cases a part-of-speech and a list of word senses. Each word sense has a gloss in French and a translation in Khmer. The Khmer translation is noted in International Phonetic Alphabet and not in Khmer writing. The dictionary was originally encoded in Word format. An example of the original file can be seen in Figure 1.

2.2.2 The Bambara-French dictionary

The Bambara-French dictionary[3] of Father Charles Bailleul (1996 edition) includes more than 10,000 entries. This dictionary is primarily intended for French speakers wishing to improve Bambara but it is also a resource for Bambara speakers. In the

words of the author himself, the dictionary "plays the role of a working tool for literacy, education and Bambara culture." To date, it can be considered as the most comprehensive dictionary of the language. It is also used by specialists of other varieties of this language such as Dyula (Burkina Faso, Côte d'Ivoire) and Malinké (Guinea, Gambia, Sierra Leone, Liberia, etc.).

2.3 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.

The most recent dictionaries of this type are built with lexicography tools such as Linguist Shoebox/Toolbox[4] (Buseman et al., 2000) and FieldWorks Language Explorer (FLEx)[5] of the Summer Institute of Linguistics (SIL) or TshwaneLex (TLex)[6].

Even if these tools are able to export content to an XML structure, it is often the case that this operation has not been performed or the files produced by the lexicographical tools are not available. Only Word files can be used for printing.

The conversion methodology (described in section 4) has been applied successfully to the following six dictionaries.

2.3.1 The Niger SouTeBa dictionaries

In the DiLAF project[2] (Enguehard and Mangeot, 2014), we worked on five dictionaries written in five national languages of Niger and French. They have been produced by the Soutéba project (a program to support basic education) with funding from German cooperation and the support of the European Union. The software used for building these dictionaries is the SIL toolbox. They have a simple structure because they were designed for children in primary school classes in a bilingual school (education is given there in a national language and in French). Most terms of lexicology, such as lexical labels, parts-of-speech, synonyms, antonyms, genres, dialectal variations, etc. are noted in the language in question in the dictionary,

contributing to forge and disseminate a meta-language in the local language, a specialized terminology. The entries are listed in alphabetical order, even for Tamajaq (although it is usual for this language to sort entries based on lexical roots) because the vowels are written explicitly (this mode of classification was preferred because it is well known by children).

The Fulfulde-French dictionary[7] includes 4,305 entries. The orthographic form of the entry is followed by a part-of-speech. Next, follows a definition and an example in Fulfulde. Some grammatical information is mentioned, such as the plural form (mbahdi). The entry ends with a French gloss (far). Here is an example:

> **saabi** *jokk*. helmere jokkondiroore konngi ngam hollude hujja. *saabi o sooda ngawri waɗi o soonni nga'ari makko*. mbahdi : saabi. *far :* à cause de.

The conversion of this dictionary was particularly difficult because the original word files disappeared and the dictionary was re-keyed in by hand from a printed version. Thus many structuring errors could be found.

The Hausa-French dictionary[2] includes 7,823 entries. The orthographic form of the headword is followed by the pronunciation (tones are marked with diacritics placed on vowels) and part-of-speech. On the semantic level, there is a definition in Hausa, a usage example (identified by the use of italics), and the equivalent in French. Here is an example:

> **jaki** [jàakíi] *s.* babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba amma ta fi shi dogayen kunnuwa. *Ya aza wa jaki kaya za ya tafi kasuwa. Jin.:* n. *Sg.:* **jaka**. *Jam.:* **jakai, jakuna**. *Far.:* âne

The Kanuri-French dictionary[2] includes 5,994 entries. The orthographic form of the entry is followed by an indication of pronunciation relating to tones. The part-of-speech is shown in italics, followed by a definition, a usage example, a French translation and meaning in French. Additional information may appear as variants. Here is an example:

abərwa [àbə̀rwà] *cu.* **Kəska təngəri, kalu ngəwua dawulan tada cakkidə.** *Kəryende kannua nangaro, abərwa cakkiwawo.* [Fa.: ananas]

The Tamajaq-French dictionary[2] includes 5,205 entries. The orthographic form of the entry is followed by the part-of-speech and a gloss in French displayed in italics. For nouns, morphological information about the state of annexation is often included, the plural and gender are also explicitly stated. A definition and an example of usage follow. Other information may appear as variants, synonyms, etc. As Tamajaq is not a tonal language, phonetics does not appear. Here is an example:

əbeɣla *sn.* mulet ♦ **Ag-anɣer əd tabagawt.** *Ibeɣlan wər tăn-tăha tămalăɣa.* anammelu.: **făkr-ejăḍ.** təmust.: **yy.** iget.: **ibəɣlan**.

The Zarma-French dictionary[2] includes 6,916 entries. Each entry has an orthographic form followed by a phonetic transcription in which the tones are rated according to the conventions already set for the Kanuri. The part-of-speech specifies explicitly the transitivity or intransitivity of verbs. For some entries, antonyms, synonyms and references are indicated. A gloss in French, a definition and an example end the entry. Here is an example:

ɲagas [ɲágás] *mteeb.* ● *brusquement (détaler)* ● sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din ● *Za zankey di hansu-kaaro no i te ɲagas*

2.3.2 The Wolof database

The Wolof database (Cissé, 2013) is a multifunctional lexical database from which it is possible to extract a monolingual Wolof dictionary as well as a bilingual Wolof-French dictionary. Each word sense has its own entry in the database. Polysemic words will then have several entries. The database was built with the SIL Toolbox tool. Each entry has several administrative fields (or meta-information): status of the entry, comments, author of the entry.

3 Technical difficulties

3.1 LANGUAGES WRITTEN BUT POORLY STANDARDIZED

While the fact that a language is under-resourced has been defined solely in terms of its equipment in IT tools and resources, the linguistic knowledge of this language is often scarce: there are few studies on the language, they are inaccessible because they are not published in journals or conference proceedings, and they are not available online. Moreover, these languages are also not very present in schools, either as a subject of study or as a teaching language. However, some exceptions are worth mentioning:

In Niger, experimental schools were established in the 1980s. The teaching is entirely delivered in a national language in the first half of the primary cycle. During the second half of the cycle, French appears and the national language is studied as a subject. In the final year, the teaching takes place in French only. It will be the same for the rest of school: middle school, high school and higher education (Programme Décennal du Développement de l'Éducation, 2003).

In Ecuador, the Shuar people (called improperly Jivaro) was structured in a Federation of Shuar Centres in 1964. In the 1970s the Federation organized, among other initiatives, the establishment of primary schools in villages with support through radio programs. These schools are bilingual. Instruction is provided almost entirely in Shuar in the first two years to move towards parity in education in Shuar and Spanish in the final year (Calvet, 1987).

Beyond the success of these approaches to children's literacy, the creation of a school curriculum promotes the writing and editing of textbooks which constitute text corpora written by language specialists, sometimes linguists, with a good knowledge of the written language. Such corpora could be exploited by linguists to build lexical resources. Their size remains small.

Other national language texts emerge. They are written by journalists and authors (of stories, novels, etc), who have mostly no access to language resources. Therefore,

these texts are written in a somewhat standardized language with particularly many spelling variations. These corpora thus cannot be used as a data source to automatically build lexical resources.

In this context of deprivation, the reuse of a published dictionary is a first step that will accelerate the creation of usable resources for natural language processing applications. In some dictionaries, word spelling is standardized and complies with linguistic studies; in others, such as (Bailleul, 1996), variants are explicitly marked and located geographically, while the official spelling is reported in addition to the usual spellings. Furthermore, the definitions and examples of use are a corpus of sentences and many entries are accompanied by morphological information for calculating the different forms of the same entry.

3.2 SPECIAL CHARACTERS

Set up in the 1960s, long before Unicode, the alphabets of most African languages use special characters which were absent from the standard character tables at that time. Although the alphabets of languages on which we have worked (Enguehard, 2009) are mainly of Latin origin, new characters needed to note specific sounds in some languages with a single character have been adopted by linguists in a series of meetings. Thus, each of the alphabets of the African languages we have worked on includes at least one of these special characters: ɓ ɗ ə ɛ ɣ ƙ ɲ ŋ ɔ ʋ. The character b with hook (ɓ, Ɓ) appears in the Hausa alphabet (République du Niger, 1999a) while the character eng (velar n) (ŋ, Ŋ) appears in the Bambara, Tamajaq (République du Niger, 1999c) and Soŋai-Zarma (République du Niger, 1999d) alphabets. Characters composed of a Latin character and a diacritical mark have also been created (such as â, ă, ã, ḍ, ğ or š).

Targeted primarily for printing texts on paper, fonts displaying these special characters have been created by redrawing the glyphs of certain characters (Chanard & Popescu-Belis, 2001). These fonts have for decades allowed texts to be published in national languages, but they prohibit natural language processing on these texts (Enguehard, 2009). The habit of using them is being installed, and the source files of

the dictionaries we have collected use such fonts. It is therefore necessary to convert them to Unicode so that the character encoding meets the international standards.

3.3 CHARACTER CONVERSION TO UNICODE

Establishing replacement characters can be tricky if the original fonts are not available, which is the most common situation. In this case, it is preferable to have a printed version to be sure to establish proper conversions. This step can be more subtle when the same character is used to display different glyphs. For example, the ampersand & is redrawn as t with a dot below ṭ of the Tamajaq alphabet in the 'Albasa Tamjq' font, as d with a hook ɗ of the Hausa alphabet in the 'AlbasaRockwellhau' and 'Hausa' fonts and as open e ɛ of the Bambara alphabet in the 'Times New Bambara' and 'Arial Bambara' fonts. It may also happen that different fonts have identical names. Finally, the same character may be used within the same document to display different glyphs. For example in the Tamajaq-French bilingual dictionary (Programme de soutien à l'éducation de base, 2007), the p lowercase character (U+0070) was used as such in the parts of the entries in French, and redesigned as a schwa ə in the 'Tamajaq Literacy2 TT20.4 SILSop' font for the Tamajaq parts.

Table 1 shows part of the list for Zarma. There is no automatic method that will detect these problematic characters. It is imperative to look at the data.

Origin	*Unicode*
§	ã
$	ɲ
ù	ŋ
£	Ɲ

Table 1: Partial view of the Unicode correspondence table for Zarma.

Figure 2 shows two entries. In their original version, special characters initially entered with a hacked font are not readable. In the Unicode version these special characters have been transformed to comply with Unicode.

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 2: Tamajaq lexical entry "ăṣaruf" in published format then in Unicode format

3.4 DIGRAPH LEXICOGRAPHICAL ORDER

Digraphs can be easily typed using two characters but their use changes the sort order which determines the lexicographic presentation of dictionary entries. Thus, for Hausa and Kanuri, the digraph 'sh' is located after the letter 's'. So, in the Hausa dictionary, the word "sha" (drink) is located after the word "suya" (fried), and, in Kanuri, the word "suwuttu" (undo) precedes the noun "shadda" (basin).

These subtle differences can hardly be processed by software and require that digraphs appear as a proper sign in the Unicode repertoire. Some used by other languages are already there, sometimes under their different letter cases: 'DZ' (U+01F1), 'Dz' (U+01F2), 'dz' (U+01F3) are used in Slovak; 'NJ' (U+01CA), 'Nj' (U+01BC), 'nj' (U+01CC) in Croatian and for transcribing the letter " Њ " of the Serbian Cyrillic alphabet, etc.

It would be necessary to complete the Unicode standard with digraphs of Hausa and Kanuri alphabets in their various letter cases.

fy	Fy	FY
gw	Gw	GW
gy	Gy	GY
ky	Ky	KY
kw	Kw	KW
ƙy	Ky	KY
ƙw	Kw	KW
sh	Sh	SH
ts	Ts	TS

Table 2: Hausa and Kanuri digraphs missing in Unicode

3.5 CHARACTERS WITH DIACRITICS

Certain characters with diacritics are included in Unicode as a unique sign, whereas others can only be obtained by composition.

Thus, vowels with tilde 'a', 'i', 'o' and 'u' can be found in Unicode in their lowercase and uppercase forms while the 'e' with a tilde is missing and must be composed with the character 'e' or 'E' followed by the tilde accent (U+303), which can cause renderings which are different from other letters with tilde when viewing or printing (tilde at a different height for example).

Letter j with caron exists in Unicode as a sign ǰ (U+1F0), but its capitalized form J̌ must be composed with the letter J and the caron sign (U+30C).

The characters ẽ, Ẽ and J̃ should be added to the Unicode standard.

Concerning letter case change, word processors usually provide this functionality, but do not always realize it in the correct way. Thus, we have found during our work that the OpenOffice Writer software (3.2.1 version) fails in transforming 'ɾ' to 'ᴙ' from lowercase to uppercase or vice versa (the character remains unchanged) while Notepad++ (5.8.6 version) fails in transforming 'ǰ' to 'J̌'.

4 Conversion process into a standardized format

4.1 LEXICOGRAPHIC FOUNDATIONS

For recovered dictionaries (such as the DiLAF project, see the central format below), we do not know exactly which lexicographical choices led to the nomenclature and microstructure. In the DiLAF project, the aim is to make existing resources available to the public. The initial lexicographical choices of the authors are therefore virtually unchanged.

- Lexicographical order is implemented by a specific function directly on the database following the orders defined by the alphabets of the concerned languages;
- The macrostructure is not altered: each paper volume is represented by an electronic volume;
- The microstructure is sometimes slightly modified according to the dictionary.

Traditional lexicography distinguishes two vocables as homographs if they have no clear semantic link with each other (Polguère, 2008). But in practice, it frequently happens that dictionaries of the same languages follow a division into different homograph vocables.

We believe that this distinction is actually arbitrary. To our knowledge, there are also no specific criteria to evaluate a semantic link between two words that can be implemented easily with NLP tools, let alone for under-resourced languages. When it is possible, we have therefore chosen to group homograph vocables into a single entry. However, we distinguish entries with different parts-of-speech. The combination of a lemma and a part-of-speech therefore constitutes a single entry.

For lexical databases generated from retrieved data (such as the MotÀMot project, see the target format below), the Jibiki platform allows great freedom in the definition of the macrostructure (ref paper links). It is possible to design complex macrostructures composed of several layers of volumes related to each other (see PiVAX (Nguyen et al. 2007) or proAxie (Zhang and Mangeot, 2013) macrostructures). However, the foundations of a Jibiki structure are based on a traditional approach to designing dictionaries, the onomasiological approach: from the lexical unit to the word meanings and gathering word meanings into interlingual links (axies) at the interlingual pivot level. Axies are not concepts, although they tend to become ones.

The OntoLex W3C working group, and more generally linked data and lemon projects are based on a semasiological approach (from the concept to the word

meaning). As stated in the first point of the OntoLex declaration: “The mission of the Ontology-Lexicon community group is to develop models for the representation of lexica […] relative to ontologies. These lexicon models are intended to represent lexical entries containing information about how ontology elements […] are realized in multiple languages. It is certainly possible to define a macro-structure based on a semasiological approach with Jibiki, but the tool has not been designed for that and it has not yet been experimented. For a more detailed discussion, see (Zhang et al. 2014).

Another difference between these two approaches is that, for the onomasiological one, the aim is to build consistent and high-quality resources. This can be done only if thorough verifications are carried out on the nomenclature (choice of entries) and the data. Several possibilities exist: to indicate for each entry a quality level, to show the history of the origin of the data, to establish a process of revision / validation, etc. For the semasiological approach, the main goal is to obtain the broadest coverage in terms of quantity of entries and number of languages. That said, we believe that in the future these two approaches, complementary, should become closer or even merge.

4.2 CHOICE OF THE STANDARD

Our goal is to convert published dictionaries to make them available to the natural language processing (NLP) scientific community. Thus, the final format must be based on standards. We studied two main standards: The Text Encoding Initiative and the Lexical Markup Framework.

4.2.1 The Text Encoding Initiative

The TEI is led by a consortium gathering American and European public research organizations. Its goal is to define an exchange format for exchanging, creating and storing annotated texts with a standardized tagset. The latest version is P5, published in 2007. Each TEI encoded document must begin with a header described in chapter 2.

Chapter 9 focuses on dictionaries. To address the problem of the structuring of entries, the TEI provides a binary solution: an article can be represented by a *<entry>* element whose structure is very rigid and codified; the element *<entryFree>* may be preferred because it admits the insertion of any elements in any order in the entry.

In practice, the *<entry>* is too burdensome to use. Therefore, lexicographers prefer to use the *<entryFree>*, but it is too loose to allow effective standardization and exchange of data encoded using the TEI.

TEI has had real success in encoding corpora. This is unfortunately not the case for dictionaries. The proposed solution (dichotomy between the *<entry>* and *<entryFree>*) is not satisfactory. As a consequence, very few dictionaries are encoded with TEI.

4.2.2 Lexical Markup Framework: the normative part

Lexical Markup Framework (Romary et al., 2004) is a meta-model separating the lexical parts, grammatical and semantic. The main class is a *Lexical Resource*. It contains a class that describes the meta-information on the lexicon *Global Information* and one or more *Lexicons*. The lexicon contains one or more *Lexical Entries*. A lexical entry contains one or more *Forms* of the entry and one or more *Senses*. The forms contain spelling variants *Form representations*. The senses can in turn contain other senses recursively. They can contain *Definitions* that contain *Text Representations* and narrative descriptions or *Statements*. The *Representation* class allows one to link different *Form Representations* and their occurrences in a *Text Representation*. This meta-model became an ISO standard under the number 24613:2008 in November 2008 (Francopoulo et al., 2009).

We would draw the reader's attention to the fact that LMF is separated into two parts. The normative part describes the meta-model but does not specify the data representation scheme (which elements and attribute names to use). This is described in the informative part of LMF, which is not included in the standard itself. Thus, it is possible to obtain a resource using its own elements and attributes names but still following the normative part of LMF. Some people enjoy that freedom (and we are

among them), while others would have preferred the informative part to be included in the standard in order for all LMF resources to follow the same syntax.

4.2.3 Lexical Markup Framework: the informative part

The informative parts of LMF give precise information as to how to encode a lexical resource in XML: the structure that must be followed and the elements and attributes that must be used. As it can be suitable for an interchange format, we prefer not to use it as a working format for the following reasons:

- The name of the elements are in English. While it can be understood by the majority of the researchers in the world, it can be difficult to understand for the people working directly on the dictionary. When developing a resource, they are the most important people to take into consideration!
- Objects of different nature can be put at the same level. We think that, in order to be clearly understandable, an XML format should avoid putting 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. The object <*GlobalInformation*> which isa meta-information about the lexicon is a sibling of the object <*Lexicon*>, which is the resource itself.
- Free text is stored in attribute values. In XML, it is customary to include items from closed lists as attribute values, and frame the free texts by using 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.

4.3 PRESENTATION OF THE METHODOLOGY

4.3.1 The conversion process

The conversion methodology proceeds in several steps and requires successive transformations of the published dictionary to several XML files called copy, central,

target and export formats. We also take into account the fact that lexicographers will revise and develop the produced resources.

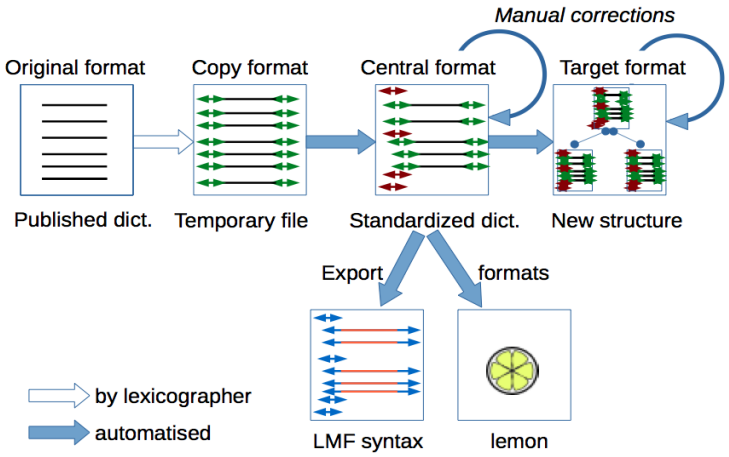

Figure 3: Conversion process

NLP experts develop conversion programs to process the transformation from copy format to central format, and from central format to target format. When they conceive these programs they get the opportunity to detect new errors and inconsistencies that are reported for subsequent correction.

The methodology is simple and based exclusively on freely available tools such as Open/Libre Office, NotePad++ and FireFox. In the following parts, we will briefly present the necessary steps of the methodology to transform one format to another. For more information, it is possible to refer to the technical manual stored on the DiLAF website (see the “Méthodologie” section of the “Projet” page). In the following section, the methodology steps necessary in order to obtain each format will also be explained.

4.3.2 The copy format

The copy format is a structural copy of the published dictionary in a valid XML format: the nature of each information part is identified and pieces of information are bracketed by XML markups according to this nature: it could be a part-of-speech, a

definition, an example, etc. The transformation of the published dictionary to the copy format is performed by lexicographers with the support of NLP experts. This step requires the resolution of many problems, including the conversion of special characters to Unicode, the identification of each information part, the definition of a set of markup tags and finally the explicit tagging of information by adding tags (Mangeot & Enguehard, 2013).

In order to bracket each information part composing an entry (definition, lexical label, phonetic, synonyms, translations, etc...), a set of elements must be chosen. This raises the question of the choice of the language used for the elements. English, the international language of research may be favoured. But in many cases, it is not present in the dictionaries. Furthermore, it is not mastered by all linguists working on the project. In the case of under-resourced language computerization projects, it is important to encourage partners to use terms in their own language (or mother tongue) to define the names of the elements. This may eventually lead to the creation of new terms that did not exist in these languages and contribute to the transfer of knowledge and ideas and, consequently, to scientific and technological development (Diki-Kidiri, 2004). From a political point of view, it participates in moving away from a post-colonial vision of the social status of these languages and contributes to their valorization.

For example, Table 3 shows the element names chosen for the Kanuri-French dictionary (Programme de soutien à l'éducation de base, 2004) in the DiLAF project:

Name of the element in Kanuri	English equivalent
kalma	headword
bowodu	pronunciation
naptu_curo_nahauyen	part-of-speech
maana	definition
misal	example
kalakta	translation
maana_tiloa	synonym
fərəm	antonym

bowodu_gade	variant
mane	cross-reference

Table 3: Element names chosen for the Kanuri-French dictionary. The conversion process from the published format to the copy format consists of four main steps:

1. Retrieving the published format in XML format (either Open Document Format or OpenXML from Microsoft). A text document in Open Document format (.odt) or Open XML (.docx) is in fact a zip archive containing several files and folders. The core document containing the text must be extracted from the zip archive. For a .odt document, the file is called content.xml; for a .docx document, the file is word/document.xml.

2. Applying an XSL stylesheet to simplify the XML. The ODF or OpenXML formats are very verbose and difficult to read. Furthermore, much of the information concerning styles is not useful for the conversion process. Therefore, a first simplification of the XML can be made with an XSL stylesheet. More precisely, the “text:p” elements are replaced by “p” elements, and “text:span” elements are replaced by the value of the “text:style-name attributes; headers, sections, columns breaks and page breaks are removed.

3. Tagging each information part by converting the XML with regular expressions. This part is the most important one and takes up most of the conversion time. One must recognize how each information part is tagged and replace those tags by the new tag set defined previously.

4. Checking the validity of the result file. First, the well-formedness of the XML file must be checked. A simple web browser can be used for this task, as long as it gives the line numbers where the errors are located. Once the file is well formed, the structural validity can be checked. Before this step, it is necessary to specify the structure of the copy format (either with a Document Type Definition or an XML schema). Then an XML parser is used to check the validity.

Annex 1 shows the result of the conversion process from the published format to the copy format for a Kanuri entry.

When a first valid version of the copy format is available, various checks are made using programs (counting the number of occurrences of each tag, checking the embeddedness of the markups, counting the number of closed list values such as parts of speech, etc.) and errors are reported to lexicographers, who can make the corrections. The copy format does not alter the structure nor the order of the information of the original format, but improves readability by explicitly labelling every information part. Finally, it is designed to disappear in favor of the other formats.

4.3.3 The central format

The central format respects the normative core of LMF. Consequently the original order of the information (that was still kept in the copy format) is changed. The structure of the entries follows the LMF meta-model but the tag names do not necessarily follow the informative part of LMF. It is obtained by applying an XSLT program that performs structural changes on the copy format. During this step, it is sometimes necessary to clarify the nomenclature.

The basic unit constituting an article may vary from one dictionary to another: the lexeme (lemmatised surface form without part-of-speech), the vocable, etc. When the basic unit is the vocable, homonym vocables have multiple entries. It is necessary to distinguish these entries in order to indicate possible links (synonymy, homonyms, etc.) between them. It is also necessary to identify each word sense of an article. It is therefore necessary to build a unique identifier for each article and each word sense in an article. When the nomenclature choices have not been respected throughout the dictionary, it is sometimes also necessary to perform certain changes in the nomenclature: merging two articles, for example when word senses of the same vocable have been described in two different articles; splitting an article into two, for example when an article includes two different parts-of-speech, etc. Finally, within the microstructure, restructuring is sometimes required, such as moving morphological information described in a semantic block to the form block.

These treatments are performed by perl programs. Markup tag names are preserved from the copy format.

Depending on the goals of the project, especially if there is a target format, the central format can be frozen and proposed for download without further modification. It represents the standardized electronic version of the published dictionary.

A detailed example of a Kanuri entry in central format is available in Annex 2.

4.3.4 The target format

The target format is specific to each project. Each target format is defined from the needs of a project. The central format is then converted automatically into the target format.

The resources in central formats are electronic versions of printed dictionaries, mainly monolingual or bilingual. The use of computers has helped to overcome the constraints of the paper form. The impossibility of inverting bilingual dictionaries led to a model having a "pivot" consisting of an axis (interlingual meanings). This leads to the definition of new macrostructures based on this pivot such as Papillon (Mangeot et al., 2003) and Pivax macrostructures (Zhang et al., 2014). Several lexical resources can be used and merged to enhance the quality and add information that is not available in the converted resource. The result is a new resource that will then be corrected and completed online by voluntary or paid contributors.

For example, in the case of the French-Khmer dictionary of the MotÀMot project (Mangeot, 2014), the French volume has been completed with information from two other existing resources: the pronunciation of the entries was taken from the FeM dictionary and the list of French entries from the GDEF dictionary (taken from the Morphalou lexicon, taken from the TLFi).

In the future, the target format will be the only one to evolve. It can then be uploaded onto an online lexical resource management platform 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, adding various information, translations, examples,

etc..). It would then be easy to generate a new export format dictionary by processing again the appropriate program on the target format dictionary.

4.3.5 Export formats

Export formats depend also on the needs of each project. But, as the central format respects the LMF normative part, it is easy to generate a format that follows the syntax of the informative part of the LMF standard. It is obtained by processing the central format with an XSLT program. The transformations are limited to changing the name of an element, to add an additional level with a child element, and to convert a text node into an attribute value.

Nevertheless, there is still an important issue: the data categories. In order to be 100% compliant with the LMF standard, the data categories such as the list of parts-of-speech must follow the ISOcat Data Category Registry (DCR)[8]. We encountered parts of speech that are missing such as "ideophone", appearing in the list of parts of speech in Hausa and Kanuri dictionaries. Thus, it appears necessary to enrich this list or to allow a modular definition of this list with a sublist for each language.

A detailed example of a Kanuri entry in LMF syntax format is available in Annex 3.

Linked data is a set of strongly interconnected graphs. In the case of data coming from dictionaries, the graphs are lexical networks where nodes represent the lexemes of one or more languages, and links represent the relationships between these lexemes (translation, synonymy, etc.). A lexical network can be monolingual or multilingual.

Although lexical networks have many advantages, they are not suitable for all usages. Usually, they are not browsable in alphabetical order. But we need that possibility to have an idea of the content of a lexical repository, whatever its nature. On the other hand, in a lexical network, the concept of volume is missing, which prevents the creation of a resource in a simple way when studying a new language. When importing previously existing dictionary data into a lexical network, the editorial responsibility of the dictionary editors is also lost. Most of the time, there is no guarantee of quality, nor a clear view of which entry should be in the network or not.

Once again, while it is important to produce linked data in order to provide machines with easy access to our data, this should not be directly used as a working format.

In order to produce linked data, the ideal export format is the lemon model[9]. It is very close to LMF (Eckle-Kohler et al., 2014). Apart from slight modifications for some element names, the biggest difference is the word senses. In LMF, there is a fixed set of senses whereas in lemon, each word sense is linked to an ontology concept. In the case of LMF, the *semasiologic* (from word to meaning) approach has been chosen whereas in the case of lemon, the *onomasiologic* one (from concept to word) has been chosen. In the case of bilingual dictionaries or multilingual databases, we prefer to follow the *semasiologic* approach and use *axes* (interlingual acceptions) instead of concepts in order to link word senses of different languages (Mangeot et al., 2003). This is the choice adopted by dbnary[10] (Sérasset, 2014), the team database for linked data.

5 Web access via a resource management platform

5.1 DESCRIPTION OF THE PLATFORM

Jibiki (Mangeot et al., 2006; Mangeot, 2006) is a generic platform for handling online lexical resources with user and group management. It was originally developed for the Papillon Project. The platform is programmed entirely in Java based on the "Enhydra" environment. All data is stored in XML format in a Postgres database. This website mainly offers two services: a unified interface for simultaneous access to many heterogeneous resources (monolingual or bilingual dictionaries, multilingual databases, etc.) and a specific editing interface for contributing directly to the dictionaries available on the platform.

Several lexical resource construction projects are using this platform successfully for consultation or edition (DiLAF[2], GDEF[11], Jibiki.fr[12], MotÀMot[1], Papillon[6], Pivax[13]).

The source code for this platform is freely available for download from github[14]. A docker image is also available[15].

5.2 IMPORTING A RESOURCE ON THE PLATFORM

The Jibiki lexical resource management platform is able to handle any resource provided that it is encoded in XML. Indeed, a system of common pointers in heterogeneous structures can manipulate the resources without changing their structure. Each pointer is indexed in a database and can perform a quick search. This system is called Common Markup Dictionary (CDM). There are predefined common pointers for lexical items that are commonly found in most dictionaries. It is also possible to define specific pointers for a resource.

Therefore, dictionaries in many formats can be imported (copy, central, target or LMF formats). It may be a structured dictionary following the recommendations of the TEI, such as the LMF standard, etc.

In order to import a resource into the Jibiki platform, it has to be described in XML metadata files. The dictionary metadata (authors, dates, licence, etc.) and macrostructure (languages, volumes, links between volumes) are described in the dictionary metadata file. Each volume and microstructure is described in a separate volume metadata file.

The entry microstructure of each volume is described using common pointers identifying the same kind of information (entry, entry id, headword, pronunciation, part-of-speech, definition, domain, word sense, translation, translation link, example, etc.). These CDM pointers use the XPath standard. They allow the resource to be handled without any format conversion that would involve a loss of information.

An HMTL interface simplifies the creation of the metadata files. The dictionary metadata file is filled in by hand. The volume metadata files are automatically generated by a program that analyses the XML data and produces several description files (number of entries, CDM pointers, XML schema of an entry, XSL stylesheet,

XML template for an empty entry, etc.). The information can then be corrected by the user via an HTML interface.

When the metadata files are ready, the resource can be automatically imported into the Jibiki platform. It is then instantly accessible for lookup and editing.

5.3 ONLINE LOOKUP AND EDITING

5.3.1 Lookup interfaces

Three different interfaces are available to the user:

- the generic lookup allows the user to look up a word or a prefix of a word in all the dictionaries available on the platform. The language of the word must be specified.

- the volume lookup allows the user to look up a word or prefix on a specific volume. In the left hand part of the result window, the volume headwords are displayed, sorted in lexicographical (alphabetical) order. An infinite scroll allows the user to browse the entire volume. In the right hand part of the window, the entries previously selected on the left are displayed.

- the advanced lookup is available for complex multi-criteria queries. For example, it is possible to look up an entry with a specific part-of-speech, and created by a specific author. On the left of the result window, the headwords of the matching entries are displayed, sorted in alphabetical order. A scroll bar allows the user to browse all the matching entries. On the right, the entries previously selected on the left are displayed.

5.3.2 Editing process

The editing module (Mangeot et al., 2004) is based on an HTML interface model instantiated with the lexical entry to be published. The model is generated automatically from an XML schema describing the entry structure. It can then be modified to improve the rendering on the screen. Therefore, it is possible to edit any type of dictionary entry provided that it is encoded in XML.

We would like to stress here the fact that although Jibiki provides all the tools necessary for a correction/revision/validation process of the data online, it is completely illusory to imagine that the quality of a dictionary will be improved by allowing potential voluntary contributors to act alone.

First, a real project must be defined with a community animator, a group of editors and validators must be selected based on their skills, and especially contributors must be motivated.

Second, the success of Wikipedia might lead us to think that the same can be obtained for the construction of a quality dictionary, but various experiences have shown us that it is not the case.

We also quote Larry Sender, founder of Wikipedia on the subject:

"To try to develop a dictionary by collaboration among random Internet users, particularly in a completely uncontrolled wiki format, now strikes me as a nonstarter."

Each Wikipedia article can be written by a specialist in his or her field, but for a general dictionary, it is not possible to find a specialist for some articles only. Only linguists who are specialists in the language as a whole can really help (after being trained in lexicography). Moreover, in the case of under-resourced languages, linguists who are specialists in these languages are few and far between. They are often very busy and cannot work on a project if not financed.

5.3.3 Remote access via an API

Once dictionaries are uploaded onto the Jibiki server, they can be accessed via a REST API. Lookup commands are available for querying indexed information: headword, pronunciation, part-of-speech, domain, example, idiom, translation, etc. The API can also be used for editing entries. The user must be previously registered on the website.

5.4 THE DILAF METHODOLOGY AND WEBSITE

Until now, we have focussed this article on the conversion methodology, but the DiLAF methodology is not limited to these technical subjects. It also includes the constitution of some documentation concerning each dictionary (*a minima*: origin of the dictionary, alphabet, parts of speech list, markup list) and the chosen licence. This documentation and the publication of our methodology is central to the scientific quality of the dictionaries.

We have applied our methodology to several dictionaries. It appears that the best results are obtained when a linguist and a computer scientist cooperate closely. This multidisciplinary work leads to scientific questioning and discussions and is also a good opportunity to detect inconsistencies that may occur in dictionaries. With such a team a dictionary can be converted and documented within one month of work. We stress that this methodology has been successful in converting entire dictionaries.

For the moment, the DiLAF website available at dilaf.org presents five dictionaries and their documentation: Bambara, Hausa, Kanuri, Tamajaq, Zarma. These dictionaries were downloaded 260 times between January 2014 and September 2015. Two additional ones are in progress (Wolof and Fulfulde).

The website is based on the Jibiki platform. People can have a look at the dictionaries or download them (in central format). It has been designed in accordance with our African partners: the interface is simple, free and designed to be used without any technical skills. We can see in Figure 4 that the website displays all the entries of the dictionary on the left hand side in order to incite the user to discover words s/he does not know, as a person usually does when leafing through a paper dictionary.

Figure 4: The entry "bannadu" on the DiLAF website

6 Conclusion

Faced with the lack of online, free, and good quality resources for under-resourced languages, we searched for an approach to help to transform the vicious circle caused by this shortage into a virtuous one.

We found that some good quality dictionaries are produced by local linguists or lexicographers, but that this knowledge usually remains unavailable online. We decided to carry out some research that could meet two objectives. First, we developed a methodology to transform a good quality dictionary (linguistically speaking) into an online resource that could be useful to both NLP specialists and populations. This methodology was designed to use only free tools and limited knowledge and to be off line. It is written simply (in French), in order to achieve the second goal: to make it possible for non-NLP specialists to put new dictionaries

online. In addition, by following this methodology, people would learn new pieces of knowledge they are not familiar with: using regular expressions, designing a DTD for an XML document, checking and correcting some non Unicode characters, etc.

In a context of poverty where speakers often have never seen a dictionary of their language, but where access to the Web is improving, the benefits for the population are considerable. Furthermore, the visibility of the results is an additional motivation for those involved in the conversion process.

The converted dictionaries are incomplete, and corrections are needed. However, they are a stepping stone to other developments: bilingual corpus construction; development of morphological analyzers; etc. The availability of these resources can motivate new researchers, and thus increase the potential for research.

Acknowledgements

The MotÀMot project that produced the French-Khmer dictionary has been partly funded by the Agence Universitaire de la Francophonie. We thank especially Vanra Ieng.

The DiLAF project that produced the Bambara-French, Hausa-French, Kanuri-French, Tamajaq-French and Zarma-French dictionaries has been funded by the Fonds Francophone des Inforoutes of the International Organization of Francophonie. We thank especially Soumana Kané, Issouf Modi, Michel Maï Moussa Maï, Mahamou Raji Adamou, Rakiatou Rabé, Mamadou Lamine Sanogo.

The ALFFA project that produced the Fulfulde-French dictionary is funded by the French National Agency for Research (ANR). We thank especially Mariam Barry and Alicia Boucard.

Appendix

Appendix 1: Kanuri entry *bannadu (2)* in copy format

```
<article>
    <kalma lamba="2">bannadu</kalma>                    lexical entry number 2
```

```
    <bowodu>[bànnàḍú]</bowodu>                                          phonetic
    <naptu_curo_nahauyen>kkye3.</naptu_curo_nahauyen>                   part of speech
    <maana>Diwiro yal alamdu.</maana>                                   definition
    <misal>                                                             example
        <version tələm="ka">Gənanjun bannaje, ku tadanju rakce          example in Kanuri
kələnju rojiwawo.</version>
        <version tələm="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>
    <maana_tiloa>lànḍú</maana_tiloa>                                    synonym
    <kalakta tələm="fa">éduquer (mal)</kalakta>                         equivalent in French
</article>
```

Informative elements in green were added during conversion from published format.

Appendix 2: Kanuri entry *bannadu (2)* in central format

```
<article id="bannadu2">                                                 article with identifier
    <bloc-vedette>
        <kalma lamba="2">bannadu</kalma>                                lexical entry number 2
        <bowodu>bànnàḍú</bowodu>                                        phonetic
    </bloc-vedette>
    <naptu_curo_nahauyen>kkye3.</naptu_curo_nahauyen>                   part of speech
    <bloc-semantique  id="bannadu2.1">
        <kalakta tələm="fra">éduquer(mal)</kalakta>                     equivalent in French
        <maana>Diwiro yal alamdu.</maana>                               definition
        <misal>                                                         example
          <version tələm="kau">Gənanjun bannaje, ku tadanju             example in Kanuri
rakce kələnju rojiwawo.</version>
          <version tələm="fra">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>
```

Structuring elements in red were added during conversion from copy format.

Appendix 3: Kanuri entry *bannadu (2)* in LMF syntax

```
<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="lànḍú">
            <feat att="type" val="synonym"/>
        </SenseRelation>
    </Sense>                                                           synonymous
</LexicalEntry>
```

Notes

References

Bailleul, C. (1996). *Dictionnaire bambara-français*.

Berment V. (2004). *Méthodes pour informatiser des langues et des groupes de langues « peu dotées »*. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France.

Buseman A., Buseman K., Jordan D., Coward D. (2000). *The linguist's shoebox: tutorial and user's guide: integrated data management and analysis for the field linguist*, volume viii. Waxhaw, North Carolina: SIL International.

Calvet, L.-J. (1987). *La guerre des langues*. Paris, Payot.

Calvet, L.-J. (1996). *Les politiques linguistiques*. Paris, PUF.

Chalvin A., Mangeot M. (2006) Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français. *Proceedings of the EURALEX 2006 conference*, Torino, Italy, 6-9 September, 6 p.

Chanard, C. Popescu-Belis, A. (2001). Encodage informatique multilingue : application au contexte du Niger. *Cahiers du Rifal* (cont. Terminologies Nouvelles), n. 22, pp 33-45.

Cissé, M.T. (2013) Le dictionnaire électronique unilingue wolof et bilingue wolof-français : une création continuée, Repères DoRiF n.3 - Projets de recherche sur le multi / plurilinguisme et alentours..., septembre 2013 http://www.dorif.it/ezine/ezine_articles.php?id=127

Diki-Kidiri M. (2004). Multilinguisme et politiques linguistiques en Afrique. *Colloque Développement durable, leçons et perspectives,* Ouagadougou, Burkina-Faso.

Eckle-Kohler J., McCrae J. P., Chiarcos C. (2014). lemonUby – a large, interlinked syntactically-rich lexical resource for ontologies. *Semantic Web Journal*.

Enguehard C. (2009). Les langues d'Afrique de l'ouest : de l'imprimante au traitement automatique des langues. *Sciences et Techniques du Langage*, Vol. 6, pp 29–50.

Enguehard C., Mangeot M. (2013) LMF for a selection of African Languages. In G. Francopoulo (ed.) *LMF: Lexical Markup Framework*, Hermès science, Paris.

Enguehard C., Mangeot M. (2014) Computerization of African languages-French dictionaries *Proceedings of Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL), LREC 2014 workshop*, Reykjavik, Island, 27 May 2014.

Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., Soria C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation,* Vol. 43, pp. 57–70. ISBN: 10.1007/s10579-008-9077-5.

Mangeot, M., Sérasset, G., Lafourcade, M. (2003) Construction collaborative de données lexicales multilingues, le projet Papillon. (Papillon, a project for collaborative building of multilingual lexical resources). In M. Zock and J. Carroll (eds). *Special issue of the TAL journal: Electronic dictionaries: for humans, machines or both?*, Vol. 44:2/2003, pp. 151-176. 2003.

Mangeot, M., Thevenin, D. (2004). Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project. *Proceedings of the COLING 2004 conference*, ISSCO, Genève, Switzerland, 23-27 August, vol 2/2, pp 1029-1035.

Mangeot M., Chalvin A. (2006). Dictionary building with the Jibiki platform: the GDEF case. *Proceedings of the Language Resources and Evaluation Conference 2006 (LREC)*, p. 1666–1669, Genova, Italy. Available from European Language Resources Association, Paris.

Mangeot M. (2006). Dictionary Building with the Jibiki Platform. Software Demonstration, *Proceedings of the EURALEX 2006 conference*, Torino, Italy, 6-9 September 2006, 5 p.

Mangeot M., Enguehard C. (2013). Des dictionnaires éditoriaux aux représentations XML standardisées. In N. Gala & M. Zock (eds.), *Ressources Lexicales : contenu, construction, utilisation, évaluation,* Linguisticae Investigationes Supplementa, John Benjamins Publishing, Amsterdam, Pays-Bas, 24 p.

Mangeot M. (2014). MotàMot project: conversion of a French-Khmer published dictionary for building a multilingual lexical system. *Proceedings of the Language Resources and Evaluation Conference 2014 (LREC)*, Reykjavik, Island, 28-30 May 2014 (to appear). Available from European Language Resources Association, Paris.

Mangeot M. (2016) Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary. *International Journal of Lexicography,* Volume 31, Issue 1, 1 March 2018, Pages 78–112; doi: 10.1093/ijl/ecw035; 35 p.

Mijinguini, A. (2003). *Dictionnaire élémentaire hausa-français*.

Nguyen H-T., Boitet Ch., Sérasset G. (2007) *PIVAX, an online contributive lexical database for heterogeneous MT systems using a lexical pivot*. Proc of SNLP 2007, December 2007, Pattaya, Thailand. pp.337-342.

Osborn, D. (2011). *Les langues africaines à l'ère du numérique*. Laval, Canada: Presses de l'Université Laval.

Oumaru, I. A. (1997). Zarma ciine - kaamuusu kayna. Editions Alpha.

Polguère A. (2008) *Lexicologie et s émantique lexicale. Notions fondamentales*. Paramètres, 304 pages. Les Presses de l'Universit é de Montréal, Montréal, 2e édition. Nouvelle édition revue et augmentée.

Programme Décennal du Développement de l'Éducation (PDDE), Niger. (2003). Enseignement bilingue, pages 121-132.

Programme de soutien à l'éducation de base (Soutéba). (2004). *Dictionnaire kanouri-français destiné pour le cycle de base 1,* Niamey, Niger.

Programme de soutien à l'éducation de base (Soutéba). (2007). *Dictionnaire tamajaq-français destiné à l'enseignement du cycle de base 1*, Niamey, Niger.

République du Niger (1999a). *Alphabet haoussa*, arrêté 212-99.

République du Niger (1999b). *Alphabet kanouri*, arrêté 213-99.

République du Niger (1999c). *Alphabet tamajaq*, arrêté 214-99.

République du Niger (1999d). *Alphabet zarma*, arrêté 215-99.

Richer, D., Keo, T., Vanra, I. (2007). *Dictionnaire Français-Khmer (en phonétique)*, D.R. Edition, ISBN-13:890-0-9.

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete: from LMF to Morphalou. *Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries, ElectricDict '04*, p. 22–28, Stroudsburg, PA, USA: Association for Computational Linguistics.

Sérasset, G. (2014) Dbnary: Wiktionary as a Lemon Based RDF Multilingual Lexical Resource. *Semantic Web Journal - Special issue on Multilingual Linked Open Data*, 2014. (to appear).

Streiter O., Scannell K., Stuflesser M. (2006). Implementing NLP projects for non-central languages: Instructions for funding bodies, strategies for developers. *In Machine Translation, volume 20*.

Zhang Y., Mangeot, M. (2013) Gestion des terminologies riches : L'exemple des acronymes Procs of TALN-RECITAL 2013, Les Sables-d'Olonne, France, 17-21 june 2013, 6 p.

Zhang Y., Mangeot M., Bellynck V., Boitet C. (2014). Jibiki-LINKS: a tool between traditional dictionaries and lexical networks for modeling lexical resources. *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, Aug 2014, Dublin, Ireland.

[1] http://jibiki.fr/

[2] http://jibiki.univ-savoie.fr/motamot/

[3] http://www.dilaf.org/

[4] http://www.sil.org/computing/toolbox/

[5] http://fieldworks.sil.org/flex/

[6] http://tshwanedje.com/tshwanelex/

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

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

[9] http://www.lemon-model.net

[10] http://kaiko.getalp.org/about-dbnary/

[11] http://www.estfra.ee/

[12] http://jibiki.fr

[13] http://www.getalp.org/pivax/

[14] http://github.com/mangeot/jibiki

[15] https://hub.docker.com/r/mangeot/jibiki/