The Mwan language : dictionary and corpus of texts

Elena Perekhvalskaya
Institute for Linguistic Studies, Russian Academy of Sciences, 9 Tuchkov p. 196054 St. Petersburg, Russia
elenap96@gmail.com

Résumé. Le projet d'un dictionnaire et un corpus de textes glosés en langue mwan a démarré en 2004. Auparavant, aucun dictionnaire de cette langue n'avait existé, et seuls quelque textes avaient été publiés. Le système d'écriture utilisé dans ces publications a été non-systématique car elle n'assurait pas la représentation exacte du contour tonal de mots. Actuellement le dictionnaire mwan a 2247 entrées, il est également utilisé pour interlinearization automatique de textes mwan. 48 textes sont glosés à ce moment (38000 mots). Ce corpus est prêt à la conversion en corpus numerique en ligne (à la base de NoSketchEngine software), et publiée en Internet ; ils seront donc disponibles à la communauté linguistique.

Abstract. The project of making a dictionary and a corpus of interlinearized texts of the Mwan languages started in 2004. Previously there were no dictionary of this language, and only a few text were published. The writing system used in these publications was controversial as it did not made the accurate fixation of the tonal contour of words. At present the dictionary of Mwan has 2247 entries, the dictionary is also used for automatic interlinearization of Mwan texts. The number of the glossed texts is actually 48 (38000 words). These text are ready to be converted into the on-line Corpus (with the help of the NoSketchEngine software), and be published in the Internet, therefore they will be available to the linguistic community.

Mots-clés : Corpus Mwan, dictionnaire, Mwan, Mandé Sud

Keywords: Mwan Corpus, dictionary, Mwan, South Mande.

1 Introduction

Mwan is a small language of the Southern Mande group spoken in the Kongasso subprefecture in central Côte d'Ivoire. According to Ethnologue-14, there were about 20000 ethnic Mwan (Ethnologue code: moa ISO 639-3). Typologically, Mwan can be characterized by its complex tonology. There are three level tones and two contour tones. A significant part of the inflectional morphology is tonal. Derivational morphology is based on compounding. One of the most interesting, from typological viewpoint, features of the language is a great number of pronominal series marked for polarity, focus and grammatical relations. The first serious work on Mwan was the article of M. Bolli and E. Flick that presentes a description of the Mwan phonology (Bolli, Flick, 1978). Later the work on Mwan was carried out by C. Fleming (Fleming, 1995) and A. Yegbé (Yegbé, 2002).

Mwan was never an officially written language, it was never used in mass media. Only three books were ever published in Mwan; a Syllabaire (Zogbé Djè 1998), a book of folk tales containing 20 texts (Gogbé 2001) and a recently accomplished translation of the New Testament (Bible 2006). My work on the Mwan language was carried out in the frames of the project of creating dictionaries, grammars and text corpora for the South Mande languages (Perekhvalskaya, 2004, 2007, 2008, 2011, 2013). In this presentation I will limit to the dictionary and the Mwan text corpus. I used the Toolbox software, and I will discuss the problems which I faced in the course of my work.

2 Writing system

Computer orientated linguistic work needs a consistent and single-valued writing system. It can probably be achieved only in the case when the writing system of a language is strictly codified as, for instance, the French or English orthography or if it is used only for linguistic purposes by one linguists.

Languages with a resent writing tradition as a rule suffer from:

- the lack of coordination between those who write; in extreme cases every one may use his/her own system of writing;

- the lack of consistency: one and the same word may be written down differently in the limits of the same text, sometimes on the same page;

- over-distinction or under-distinction of segmental or suprasegmental relevant units;

- dialect or idiolect variation.

It is necessary to elaborate an appropriate writing system which would be basic for the automatic text processing.

This system must be consistent and, probably (but not obligatory), based on the existing orthography sometimes with some modification.

The conversion of the existing texts into texts suitable for automatic processing may be automatic only if the correlation between their writing systems is consistent, otherwise it has to be done manually.

In my Mwan corpus, I use the writing system that was created on the basis of the existing alphabet based on the Latin script proposed by Margrit Bolli and Eva Flick in 2000 in the frames of the SIL International activities. The main defect of the Bolli and Flick's representation of the Mwan sound system is the tone marking. They denoted tones with punctuation marks (apostrophe, hyphen, equal sign). For one vowel words High tone was marked by the apostrophe, Low tone by the hyphen, Middle tone by the absence of a sign, the equal sign marked the modulated tone: e.g. *'fɛ* /fɛ́/ 'house'; *ye* /yē/ 'to see'; *-yi* /yì/ 'water'. For two-vowel words, if the two vowels bear the same tone, only the tone of the first vowel was marked: *'peni* /pénı̰/ 'sting'; *bie* /bīē/ 'elephant'; *-vakɔ* /vàkɔ̀/ 'sugar cane'. If the first vowel bears Low tone, and the second is "higher" (High or Middle), the end of the word is labeled with the apostrophe: *-gbaan'* /gbə̀ə̰̄/ 'dog'; *-soo'* /sɔ̀ɔ́/ 'horse'. If the tone of the first vowel is High, and the tone of second vowel is "lower" (Low or Middle), the end of the word is marked by the hyphen: *'pubɔ-* /púbɔ̄/ 'to greet'; *'kpata-* /kpátà/ 'rack'. The Middle tone on the first vowel is not marked, the High tone of the second vowel is denoted with the apostrophe, the Low tone of the second vowel being denoted with the hyphen: *kɔnɛ'* /kɔ̄nɛ́/ 'bug'; *nina-* /nı̰̄nə̀/ 'to return'. For three-vowel and more complex words only the tone of the initial vowel is denoted: *-amasrɔyi* /àmə̰̄srɔ̀yí/ 'because'; *laanima* /lāānı̰́mə̰̄/ 'upwards'; *'ŋkɛna* [ŋ́kɛ̀nə̀] 'good morning'.

This system of tone marking makes it impossible to record accurately the tone contour of the word and therefore can not be used for the language documentation. In my project the tones are consistently marked: the Low tone is indicated by "gravis", the High tone by the sign "acute accent", the Middle tone is denoted by the macron, the sign "circumflex" is used for rare cases when the modulated HL (v̀v) tone is beard by a short vowel. Nasalized syllables are marked by the tilde under the vowel: *bīē* [bīē] 'elephant', *gbə̀ə̰̄* /gbə̰̄ə̰̄/ 'dog', *kɔ̄nɛ́* /kɔ̄nɛ́/ 'bug', *púbɔ̄* /púbɔ̄/ 'to greet'.

3 Dictionary

The dictionary for the automatic language processing may be just auxiliary, containing only necessary fields : lemma, alternative variants, tags and glosses. However, it must contain not only lexemes of the language but also bound morphemes with all their free or context depending variants.

In the frames of my project, the Mwan dictionary created on the basis of the Toolbox software is at the same time a full dictionary of the language and the auxiliary dictionary for interlinearization. It contains fields of a full Mwan-French-English-Russian dictionary and also fields for glosses (in three languages). An export made from the Toolbox (Dictionnaire mwan-français) is available on line http://mandelang.kunstkamera.ru/files/mandelang/introd_mwan.pdf; http://mandelang.kunstkamera.ru/files/mandelang/mwan_dic.pdf

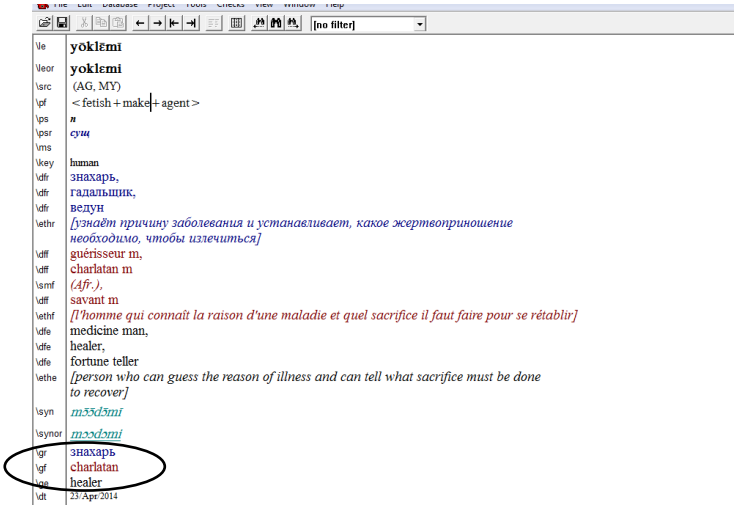

FIGURE 1 : Page of the Mwan-French-English-Russian Dictionary

At present, the Mwan dictionary contains 2247 entries : words and bound morphemes. The lemma (the field \le) is given in scientific orthography; the field \leor copies the lemma in practical writing. No transcription is given as the scientific notation makes it possible to establish unambiguously the phonemic structure of the word. Verbs are given as roots in the field \le, which is used for interlinearization and in nominalized from (with the suffix -le) in the field \leor which is intended for the native speakers of Mwan.

The dictionary contains also bound morphemes (inflectional or derivational), which are needed for the morphological analyzer. The field \a contains all variants, segmental and tonal, free and contextual. When the variant form is an indissoluble unit, it is unscrambled in the field u\. Example : there is a tonal morpheme in Mwan. The Middle tone marks the Habitual aspect in verb. So the dictionary contains the morpheme coded as -= and glossed HAB. Habitual verbs forms are unscrambled (the field \u) as having the morpheme -= (see Figure 2.)

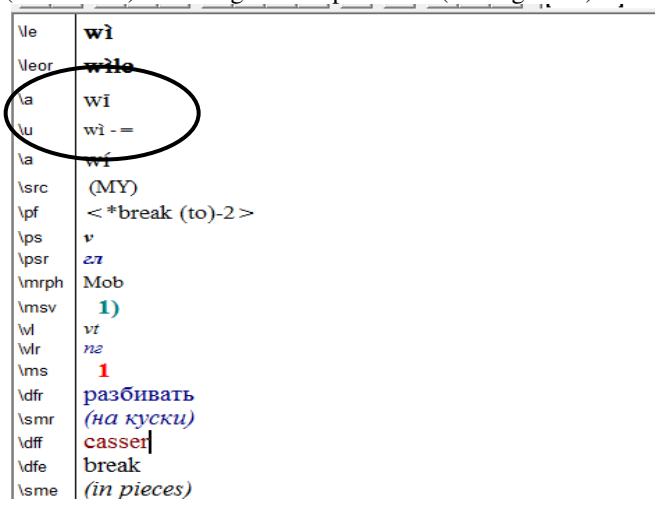

FIGURE 2 : Presentation of a tonal morpheme

Every entry contains all the grammatical information concerning the corresponding lexeme or morpheme : 1) word class (\ps); 2) morphonological, free or dialect variants (\a); 3) information on the inflectional type (\mrph); 4) irregular forms

(\gre); 5) stylistic usage (\use). The idiomatic usages are given inside the corresponding entry. Many entries are provided with illustrative examples. For polysemous words definitions, explanations, examples and idiomatic expressions are given for each value.

3 Interlinearization

The interlinearization of texts is made in the frames of the Toolbox software. The line \mb is produced automatically by the morphological analyzer with the help of the morphemic dictionary. Each morpheme is glossed (in three languages). If there are two or more possibilities of the morpheme analysis, the homonymy is removed manually (see the box in Fig. 3.)

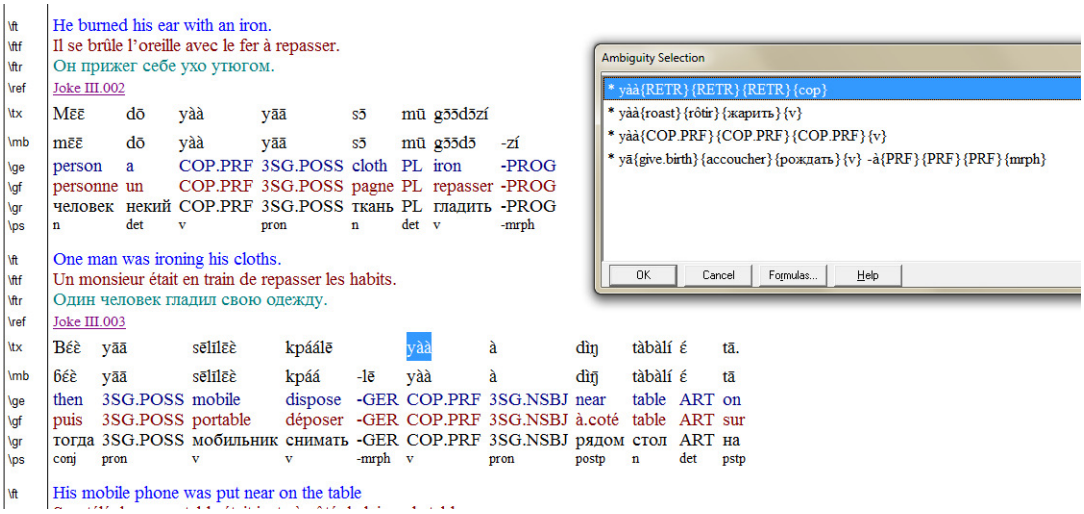

FIGURE 3 : P

In some cases the word frequency is taken into account. Homonyms of frequent lexemes or morphemes which are much rarer, are derived from the morphemic search. They are listed under the field \lx. If necessary, they are moved back to the field \le (manually).

So, the dictionary, in fact, contains a part which is NOT used for interlinearization (Passive words). However, it is possible to move them to the Active part when necessary.

Examples of these "removed" homonyms: *dɔ̄ɔ̄* 'winged termite', homonym for *dɔ̄ɔ̄* 'that' (marker of indirect speech); *yàà* 'prickly yam', homonym for *yàà* 'was, were (Perfective of the copula *ò*), *yàà* 'to roast, to grill'.

Another example is taken from my Udihe Corpus project[1]. The dictionary of this language contains three homonyms: 'ai 'ai 'elder brother', 'vodka, strong spirit' and 'ai 'buttocks'. As the majority of the available Udihe texts are folk tales, so 'elder brother' is a very frequent lexeme, 'vodka' is much less frequent, and 'buttocks' is extremely rare. Therefore, the last two units are placed to the Passive part of the dictionary.

Sometimes such homonymy of unequal value can be solved by taking in consideration the context. E.g. in Udihe Corpus there two homonym morphemes: -ni 'marker of the 3 Sg' and -ni 'a rare variant the Dative case marker'. The variant of the Dative marker appears only in postpositions and only before the markers of person/number (postpositions are a specific class of nouns). I put the sequences like *-nini* as a separate entry in the dictionary without a gloss but with the field u\ which "explains" the form as *-du* 'DAT' and *-ni* '3SG' (Fig. 4).

FIGURE 4 : P

[1] I am also engaged into the corpus project of Udihe which is not an African language. However, the experience obtained while making the corpus of a small mainly unwritten language may also be useful for preparing corpora of an African language.

Words belonging to different word classes (which have different tags in the field \ps) are given as different entries, even if the etymological link between them is obvious, e.g. : Mwan *kɔ̀ɔ̀* 'hand' (noun) and *kɔ̀ɔ̀* 'with', 'at', 'in the hands of' (postposition); Udihe: *tuəʒə* 'winter house', *tuəʒə* 'to spend the winter, to hibernate'.

Glosses. From the semantic point of view glosses do not provide the true translation, as they are conventional. One and the same gloss is ascribed to all the values of a polysemous word, including, for instance, valence changing p-labile verbs; e.g. Mwan: the verb wlā has the following meanings: 1) 'to arrive', 'to come in'; 2) to enter (school), to join (organization); to convert oneself (to religion); 3) 'to make come', 'to make enter'; 4) to put on (of hats). The unique gloss is enter. The noun wī denotes: 1) animal (general word); 2) meat; the gloss is meat. Obviously, a gloss may be more or less opportune in different contexts.

In Udihe, the word *mafa* has the following main meanings: 1) old man; 2) husband; 3) bear. All the three words are extremely frequent in folk tales and sometimes appear in the same text, like "The bear said to the girl: I have found you a husband". It would be really misleading to use the same gloss for all the three meanings which seem nevertheless the values of the same word (at least etymologically). They can be given as separate entries in the dictionary. It would multiply variants in the ambiguity selection box. It is possible to ascribe more than one gloss to the same entry (Figure 5)

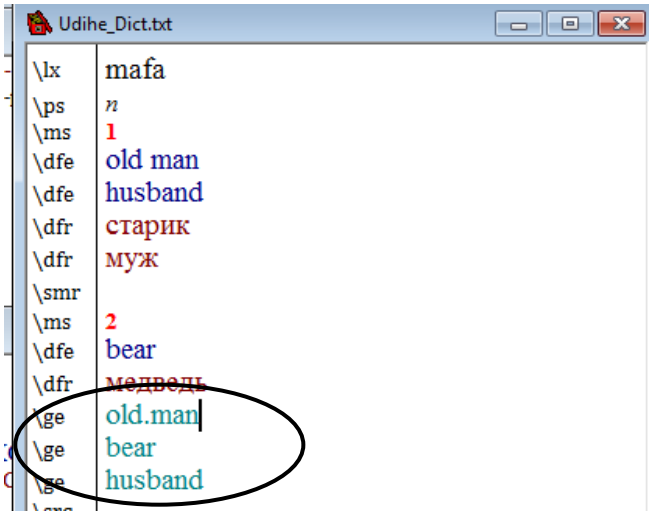

FIGURE 5 : Doubling (trebling) of a gloss

In this case the Ambiguity selection box will give only one possibility: *mafa*, which will significantly reduce the number variants; the selection between meanings will be maid in the next step.

The free translation is also given in three languages.

4 Corpus of the glossified texts

4.1. Composition of the Corpus

At present, the number of interlinearized Mwan texts is 48, the total number of words is about 38000. The texts belong to the following genres :

- oral transcribed texts :
  - folk tales;
  - tales (including tales of witchcraft);
  - oral history;
  - dialogues;
- texts taken from published sources:
  - folk tales;
  - tales (including funny stories);
  - proverbs;

- translations from French.

Oral texts, especially dialogues, contain a large amount of incomplete sentences, hesitation pauses, discursive markers and loans from other languages, mainly, from French and Jula. All these elements have to be present in the dictionary (or in a supplementary dictionary) otherwise the automatic processing would be impossible.

Oral texts present the following problems:

- unfinished or illegibly pronounced words;

- discursive markers and expletive words;

- words from other languages, especially when recording dialogues with code-switching.

Written texts, as a rule, use the inconsistent orthography:

- compositions of two or more roots may be written as one solid word and as two or three words in the frames of one and the same text;

- suffixes are written together with the root or separately;

- imprecise tone marking.

Both types of texts contain a lot of non generally known toponyms and anthroponyms.

As Toolbox makes it possible to use more than one dictionary for interlinearization, all anomalous segmental units may be grouped in a Supplementary dictionary. It is possible to make several supplementary dictionaries: for anthroponyms, for non adopted words of the dominant language, for some sort of rubbish (incomplete words, pauses of hesitation etc).

As for the inconsistency in written texts, the original version should be presented in a separate field (\txor) which is not used in automatic processing. The field \tx has to be filled with the uniformly written words.

4.2. Corpus search

The interlinearised, annotated and translated text can be easily transferred into an Internet accessible interface with the possibility of searching (for instance, using the NoSketchEngine software platform).

At present it was done for the Udihe project. As an illustration I present two concordances for the discussed above word *mafa* ‘bear, husband, old man’ 1) as a single word *mafa* (Figure 6) and 2) *mafa-ni* with the marker of 3SG(Figure 6):

Shneider_Anuj20.338	ñauxe mafalaha biixiu waamuhini . Ɲeneje .	Mafani husband-3SG	gulinkini . Joxowe , zaktawa xebusini Goc
Shneider_Anuj20.357	Dogbo ŋuhagiheti . Timadula teegiheti .	Mafani husband-3SG	iseisi-ni sul'aima mam'asani jemi alagdig'a
Shneider_Anuj9.063	zeuwe zeptei o-si , sama ali-da ehi emegi .	Mafani old.man-3SG	budehi-jaza , mam'asani inigi bihi-jaza
Shneider_Anuj9.063	bihi mam'asani mafanami b'a-giini , tei	mafani old.man-3SG	mam'asanami bunige buadini b'a-giini (
Baskakova III.07.042	, meisiheni , tagdahani maŋga bejezini .	Mafani husband-3SG	amăzaŋazi b'oŋihani , mam'asai tuxi doolor
Baskakova III.14.044	Azigama sitewe-tene amini dielani ambugiheni .	Mafani husband-3SG	meisiheni , esini ŋua . Dogbo dul'aŋkini

FIGURE 6 : Results of the search for *mafa-ni*

Shneider_Anuj14.072	xauntasi-ga-i " ogbö xoktoni bise-jeu ,	mafa bear	xoktoni bise-jeu ? " Anci , guŋke , j'eu
Baskakova III.03.001	bagdiheti , bimie , bimie omo amba , omo	mafa bear	emeheni . Zuu aziga bihileni emeheni .
Baskakova III.03.011	Digalahani amba digahani . Neŋuni neehani	mafa bear	digalahani mene moxozi . Mafa digahani
Baskakova III.03.012	neehani *mafa* digalahani mene moxozi .	Mafa bear	digahani . Zuu aziga mafalahani . Exini-tene
Baskakova III.03.050	b'ahanzifei ŋenieti amintigifei . Omo	mafa bear	, omo amba sitetigi digaŋkini : - Jele
Baskakova III.03.054	iigiheti . Zugdifei bisiti . Omo amba , omo	mafa bear	bisiti amiti . Mafa mene buatigi ŋenieni
Baskakova III.03.055	bisiti . Omo amba , omo *mafa* bisiti amiti .	Mafa bear	mene buatigi ŋenieni . Amba mene buatigi

FIGURE 7 : Results of the search for *mafa*

The obtained results show that these words are not complete homonyms: *mafa* 'bear' does not attach the personal marker; *mafa* 'husband, old man' is always used with the personal marker > *mafa-ni* which is logical, as *mafa-ni* 'husband' is a kinship term.

This small demonstration shows great possibilities of language corpora for linguistic research. In theory, any Udihe noun can attach the 3 Sg suffix, so the distinction between these two words was not clear. The research based on the massive data showed that the words *mafa* 'bear' and *mafa-ni* 'old man', 'husband' differ by their grammatical behaviour, thus, they probably should be regarded not as different meanings of the same word but as different words.

Conclusion

The demonstrated above algorithm of creating language coropora may be represented as follows:

transcription of an oral text digitizing of a printed text

|

unifirm written representation

|

creating a Toolbox project

|

interlinearisation with Toolbox

|

concerting the interlinearized texts into a linguistic corpus using the NoSketchEngine software platform.

This is a rather strait and relatively easy way to create a coprus of a rarer used language.

References

BIBLE (2006). -Jan 'Nranle- 'Sɛwɛ. Le nouveau Testamenten mwan de Cote d'Ivoire. Bienne, Suisse : Wycliffe Bible Translators.

BOLLI M., FLICK E. (1978). La phonologie du Muan. *Annales de l'Université d'Abidjan. Sér. H.*, T. XI, Fasc. 1.

FLEMING C.B. (1995). *An introduction to Mona grammar*. Thesis (M.A.). Arlington : University of Texas.

GOGBÉ A. (2001). *Mwa ta can mu-le –gɛ. Contes Mwan*. Abidjan : CIL.

PEREKHVALSKAYA E. (2004). La morphologie verbale du mwan (Côte-d'Ivoire). *Mandenkan* 39, 69-85.

PEREKHVALSKAYA E. (2007). Les propositions relatives en mwan. *Mandenkan* 43, 47-59.

PEREKHVALSKAYA E. (2008). Body parts and their metaphoric meanings in Mwan and other South-Mande languages. *Mandenkan* 44, 53-62.

PEREKHVALSKAYA E. (2011). Nominalization in Mwan. *Mandenkan* 47, 57-75.

PEREKHVALSKAYA E. (2013). L'espace déictique dans la langue mwan. *Mandenkan* 50, 103-116.

YEGBÉ K.A. (2002). *Processes of nominalization in Mwan*. Nairobi : Nairobi Evangelical Graduate School of Theology.

ZOGBÉ DJÈ P. (1998). *Mwa mu 'an 'sɛwɛ-kɔɔ' a daan*. Abidjan : CIL.