Towards The Manding Corpus: Texts Selection Principles and Metatext Markup

Artem Davydov
St. Petersburg State University, Faculty of Oriental and African Studies
199 034, Universitetskaya nab. 9/11, St. Petersburg, Russia
artdavydov@mail.ru

Abstract

In the present paper I discuss some issues related to the development of the Manding (Mande family, Niger-Congo) corpus. In the first part of the paper I give a brief survey of the potential text sources for the corpus, emphasize the problems related to their usage and make an attempt to formulate some basic principles of texts selection. In the second part I propose a typology of texts to be used for the metatext markup in the corpus.

1. Introduction

Manding is the branch of closely related languages inside the Mande family, Niger-Congo. In the present paper I will discuss some basic principles of texts selection for the electronic corpus of Manding, which is now being built, and give some preliminary remarks on the metatext markup in this corpus. The major language in question is Bamana (spoken in Mali), although the parsing tool we develop will be potentially able to work with some other Manding languages (Maninika, Dyula). The potential range of users of the corpus, as we see it, is relatively wide and includes Mandeists (both grammarians and lexicographers), typologists, specialists in language development and native speakers of Manding who are interested in their own language (school teachers, publishers, etc.).

The development of the corpus is only in its initial stage, therefore the present paper is rather a material for discussion, than an account of how our corpus is already arranged.

2. Texts selection principles

In (Sinclair 2005) two types of criteria of text selection are discussed: the *external* criteria, which take into consideration the communicative function of a text, and the *internal* criteria, which reflect details of the language. Sinclair emphasizes that corpora should be constructed exclusively on external criteria, i.e. the contents of a corpus should be selected disregarding the language they contain. I share Sinclair's opinion, as the scholar's goal is to describe the linguistic reality rather than to alter it; but accepting the external criterion as the only one for building the Manding corpus we face certain difficulties.

Below I give a brief survey of the available text sources for the corpus. As I will attempt to show in this paragraph, building a corpus of a language recently put into writing, we cannot but move towards the language standardization. We have no other choice than to use the internal criteria along with the external criteria, at least at the initial stage of the corpus building.

2.1. Oral texts

The Malian society is not monolingual. Bamana-French code-switching is a common practice, especially among the intellectual elite; unadapted French loanwords are frequent in Bamana oral texts. Thus, the Bamana corpus, if we want it to reflect the real everyday languaging, requires the parsing tool not only for Bamana, but also for French. Still, there are some sources which represent the oral form of Bamana in its "pure" form, such as transcriptions of speeches and interviews made by linguists and ethnologists in villages, where the French language is not in use. Transcriptions of TV- and radio programs should only be used with great prudence for building the oral speech sub-corpus. Two groups of texts must be distinguished here:

- spontaneously generated texts;
- texts which had been first written and then pronounced.

The texts of the second group (movie scripts, pre-written radio programs, etc.) can not be considered as oral, or, at least, they should be attributed to a separate category of oral texts. This point may seem self-evident, but, surprisingly, in many linguistic corpora (in the Russian National Corpus, for example) these two groups of texts are undistinguished.

2.2. Written texts

As concerns written texts, the following types are relevant:

2.1.1. Published folklore texts

Such texts as popular tales, anecdotes, epic poetry, etc. are of great importance, as far as they represent, in fact, oral speech. Fairy-tale publications are numerous, anecdotes are regularly published in the *Kibaru* newspaper (Bamako) and there are a lot of scientific publications of epic Manding texts, so there are no problems with the availability of texts.

2.1.2. Fiction books

Like transcriptions of TV- and radio programs, fiction books must be used with great caution, for the language it may contain is often highly unnatural, as it represents a translation of a mediocre quality from French. For instance, a Bamana novel "Kanuya Wale" was first written by Samba Niaré's in French (under the title "Acte d'amour") and then translated by the author into Bamana. The author himself mentioned that he first wrote the novel in French, because he could not express his ideas directly in his mother tongue. Thus, the novel contains syntactic structures, word usages and even morphological features untypical for the spoken Bamana. As a result, the novel is sometimes almost incomprehensible without the support of the French original text.

Writing first in French, then translating the texts into Bamana is rather common for Malian writers. The question is, should we disregard translated texts at all (it was the decision of the developers of the Russian National Corpus) or should we consider French calques to be a characteristic trait of the modern Bamana literature and include texts like "Kanuya Wale" into the corpus? When a corpus for a language with a long written tradition is created, it is easy to avoid translations from other languages, since original texts are readily available. But can we afford the same working with Bamana?

2.1.3. The press

The press is another potential source of corpus materials, though it raises the same questions as the fiction literature.

2.1.4. Educational and religious literature

The other possible sources are: school textbooks, educational booklets, religious literature, etc. Judging by the bibliography compiled by G. Dumestre (1993), the diversity of genres of the Bamana literature is rather rich. Of course, this diversity must be adequately reflected in the corpus.

2.1.5. Nko publications

Finally, there is another type of texts we should not forget about – the Nko publications. The original N'ko script was invented in 1940s for the Maninka language by a Guinean self-educated scholar Suleymane Kanté. Since then N'ko has evolved to a rich written tradition and a strong cultural movement which has many followers in Manding-speaking communities all over the West Africa, especially in Guinea. The followers of the N'ko movement believe that various Manding languages (Bamana, Maninka, Mandinka, etc.) are the dialects of the same language called "N'ko" (*ń kó* means 'I speak' in the majority of the Manding languages). Some steps towards the official recognition of N'ko have been made in Guinea, where it is used for formal schooling.

Using N'ko texts for building a corpus has both advantages and disadvantages. The most advantageous point is that in N'ko publications all tones are indicated. Unfortunately, disadvantages are even more numerous. First of all, using the N'ko script in the corpus will diminish the range of its potential users. Of course, this problem is solvable as it is possible to automatically convert N'ko into Roman letters. Secondly, N'ko is not just an alphabet, but a whole written tradition. The N'ko written language has significantly evolved from the spoken Maninka. The texts in N'ko contain numerous neologisms created by the followers of the N'ko movement, and even syntactical constructions untypical of the spoken Maninka. If N'ko texts constitute the main bulk of the corpus, we run the risk to have a corpus of a variety which has only a limited use among the followers of the N'ko movement. Finally, despite the pan-Manding ideology of the N'ko movement, the absolute majority of N'ko publications are in Maninka. However, the popularity of N'ko in Mali is growing; some field work needs to be done in order to find out if any N'ko texts in Bamana (published or unpublished) are already disposable.

2.3. Conclusion

As it has been shown above, certain types of texts raise doubts. The question is, should they be included into the corpus or not. In my opinion, it must be answered positively, for, under existing conditions of lack of resources, it would be too wasteful to reject any text data. On the other hand, every text included into the corpus must be provided with detailed metadata in order to allow the user to single out the sub-corpora he is interested in.

At the same time, at the initial stage of the corpus building the preference must be given to the texts which represent the "pure" Bamana language (folklore, village recordings, etc.) in order to develop an operational parser/tokenizer. At the next stage, more complex and "artificial" texts can be added.

3. Metatext Markup

The Manding texts classification proposed below takes into account the EAGLES recommendations on text typology (Sinclair 1996) and the metatext markup of the Russian National Corpus, as described by Savchuk (2005), but is accommodated to the Manding data specifics. To describe each text in the corpus, I propose various parameters which can be united into several groups:

- text metadata;
- data about the author;
- linguistic metadata;
- subject matters;
- technical metadata.

In the subsequent paragraphs these groups of parameters are discussed in more detail.

3.1. Text metadata

1. The *title* of the text (if any).
2. The *date* of creation of the text (exact or approximate), which is not to be confused with the date of publication.
3. The *size* of the text (in words).
4. *Original text* or *translation*. If the text is a

	Bamana		**Maninka**	
TAM	**Affirmative**	**Negative**	**Affirmative**	**Negative**
Perfective (transitive)	*-ra*	*má*	*-da, -ra*	*má*
Perfective (intransitive)	*yé*		*kà*	
Present	*bɛ́*	*tɛ́*	*yé…-la*	*tɛ́…-la*
Future	*bɛ́nà*	*tɛ́nà*	*dí, dínà*	*tɛ́*
Optative	*ká*	*kána*	*yé*	*ká, kána*

Table 1: Bamana and Maninka TAM-systems

translation, then the language of the original must be indicated.

5. *Channel: written* or *oral*.
6. Bibliographical data (for published texts): the place and the year of publishing; the name of the publishing house; the number of pages. For periodicals: issue number.
7. For manuscripts: the place and the date of creation (if known); the form of the text (handwritten, typewritten, electronic).
8. For oral speech events: the place and the date of recording.

3.2. Data about the author

This point includes the following parameters:

1. *Author's full name* (if known). If the author is using an alias, his real name must also be indicated (if known). If the text has several authors, all of them must be listed. Note: In the Manding corpus, information concerning the author's name is more useful, than, for example, in the corpora of European languages, as it indicates ethnic and caste origin of the author.
2. *Author's age* at the moment of creation of the text (exact or approximate).
3. *Author's gender* (male, female or unknown).
4. *Author's fluency in language* (native- or non-native speaker). Texts written in Bamana (or in any other Manding language) by non-native speakers are few, but the quantity of texts generated by non-native speakers will significantly increase when it comes to the sub-corpus of oral speech, the Manding languages being broadly used as lingua franca in West Africa.

3.3. Linguistic metadata

3.3.1. Manding variety

The Manding corpus is intended to be multilingual, although on the initial state, only Bamana data will be included. It is potentially possible to use the same morphological parser for three out of four major Manding varieties (Bamana, Maninka and Dyula), but not Mandinka, as its morphology differs significantly from the other varieties. Thus, the main options for the *Manding variety* parameter are: Bamana, Maninka, Dyula. Here, we face certain difficulties connected to the fact that borders between Manding varieties are only vague. In this point, only a conventional decision is possible. The most convenient criterion for distinguishing the major Manding languages is the system of verbal TAM (tense, aspect, modality) markers. Table 1 shows the TAM markers in Bamana and Maninka of Guinea.

In addition, the linguistic diversity inside the major Manding varieties is considerable, although "normative", or "standard" variants are available ("standard Bamana", based on the dialect of Bamako; Maninka-Mori, based on the dialect of Kankan). Dialect tags must be assigned to the texts to the extent possible. At the same time, texts written or pronounced in standard Bamana may have some characteristics of other dialects, depending on the origin of the writer/speaker. Two decisions are possible here: to assign only one dialect tag to each text depending on a quantitative criterion, or to indicate both dialects. The second decision seems preferable, as the presence of two dialect tags itself will tell the user that the text represents a "mixed" idiolect.

Dialectal tags can be assigned according to the external criterion (if the origin of the speaker/writer is known) or to the internal criterion (if the dialectal origin of the data can be identified reliably).

3.3.2. Writing system

Manding uses various writing systems. The main options for the *Writing system* parameter are:

1. *Malian Roman-based orthography*, which exists in two variants, old and new. The absolute majority of texts published in Mali use this system of orthography.
2. *French-based spontaneous orthography* used by people literate in French, but illiterate in their mother tongue. This system is not standardized, but texts written in this "orthography" have much in common: digraphs *ou* and *gn* for /u/ and /ɲ/, respectively, acute accents on the final vowel, etc.
3. The *Adjami* writing system based on the Arabic alphabet. Like the French-based spontaneous orthography, it is not standardized and exists in many local and individual variants.
4. The *N'ko* writing system, which has already been mentioned in p. 2.1.5.

5. Other Roman-based orthographies used for various Manding idioms in different countries of West Africa.

As shown by Vydrine (2008, 19-20), it is technically possible to establish correlations between different graphical forms of the same word in the Manding corpus. Each text in the corpus is stored in two graphical forms: in its original form and in “standard” transcription. The latter is very much alike the new Malian Roman-based orthography, but as opposed to it, has tonal markers.

3.4. Subject matters

The *subject matters* parameter is used for the metatext markup in every major European language corpus I know. The majority of these corpora are based on the EAGLES text typology recommendations. For the Manding corpus I propose a more concise list of subject matters for two reasons: firstly, the variety of texts in Manding is inferior in comparison with any European language; secondly, a comparatively small corpus that we are building would not need such a detailed subject matters classification as a large one. For the same reasons I do not propose to introduce the second dimension to the subject matters classification, although some corpus developers do that. For instance, in the Russian National Corpus, two dimensions are used: *subject matters* and so-called *functional spheres* (journalism, technical, academic, official and business, day-to-day life, advertising, theological, electronic communication). On the other hand, in many large corpora (such as British National Corpus), only one dimension is successfully used.

The subject matters classification I propose (which does not pretend to be exhaustive) reflects the real functioning of the Manding languages. So far, it consists of twenty six tags, and some of them are united into groups:

1. *Folklore: Fairy-tales, Anecdotes, Epics, Proverbs, Traditional Songs Lyrics.*
2. *Fiction: Prose, Movie Scripts, Theatre, Poetry, Popular Songs Lyrics.*
3. *Religious: Christian, Islamic, Other.*
4. *Educational: School Textbooks, Language and Literacy, Health, Agriculture, Science.*
5. *Personal: Personal Records, Correspondence.*
6. The other tags are: *Journalism, Sports, History, Advertising, Business, Everyday Life.*

Tags can be combined without any limits. For example, a school textbook in chemistry will acquire the following tags: *Educational: School Textbooks* and *Educational: Science*. For a merchant’s trade log the tags will be: *Personal Records* and *Business*, etc.

3.5. Technical metadata

Finally, to trace the corpus updating each text is to be provided with the following information:

- the name of the project member who added the text to the corpus;
- the date of the adding the text to the corpus.

4. Conclusion

The suggested metatext markup system will provide a user with the ability to create sub-corpora with the specified parameters. It will also help to control the process of filling the corpus with new text data and to estimate the balance of the corpus.

5. Acknowledgements

The Manding Corpus project is supported by Russian Foundation for Basic Research (RFBR) under Grant № 10-06-00219-a: Elaboration of the model of electronic corpus of texts in Manding languages (Maninka, Bamana).

6. References

Dumestre, G. (1993). *Bibliographie des ouvrages parus en bambara*. In: Mandenkan nº 26, pp. 67–90. [Online]. Available: http://llacan.vjf.cnrs.fr/PDF/Mandenkan26/26Dumestre.pdf (accessed March 2010)

Савчук С.О. (2005). Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции. In: Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., с. 62–88. [Savchuk, S.O. Metatext *Markup in the National Corpus of the Russian Language: Basic Principles and Functions.* In The National Corpus of the Russian Language: 2003-2005. Results and perspectives. Moscow, pp. 62–88.] [Online]. Available: http://ruscorpora.ru/sbornik2005/05savchuk.pdf

Sinclair, J. (2005). *Corpus and Text - Basic Principles.* In Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 1–16. [Online]. Available: http://ahds.ac.uk/linguistic-corpora/

Sinclair, J. (1996). *Preliminary recommendations on text typology*. EAGLES Document EAG-TCWG-TTYP/P. [Online]. Available: http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Vydrine, V. (2008). Glossed electronic corpora of Mande languages : A perspective that we cannot avoid. In *Mande languages and linguistics. 2nd International Conference: Abstracts and Papers*. St. Petersburg, pp. 16–23.