Glossed electronic corpora of Mande languages:
A perspective that we cannot avoid

1. Since the last two decades, linguistics has entered the era of electronic corpora of languages, which has considerably altered the methods and the possibilities and the accuracy of language studies. In this field, African linguistics is lagging far behind: it can boast a Swahili 12 million-word corpus created by Arvi Hurskainen (Hurskainen 2004; unfortunately, this corpus is not open for the broad public); corpora of the languages of the Republic of South Africa are being elaborated. So far, that seems to be all. To my knowledge, not a single Mande language benefits from the advantages of the corpus linguistics.

What are those advantages? In fact, any linguist working with an African language has a language corpus of his or her own; no language analysis or description is possible without it. And if this corpus is in electronic form, the linguist’s job becomes incredibly easier: instead of perusing dozens and hundreds of pages line by line in search of a form or construction needed, the desired data can be found during a couple of minutes (or even seconds) through the operation “Search”.

However, if we need our data to be representative, the search operation should be performed on a more or less considerable sample of the texts. How big must be the sample? In 1987, a very first corpus-based Cobuild English Dictionary was published which made a revolution in English lexicography. That dictionary was based on an electronic corpus of about 7 million words (the state of 1982). This figure may seem ridiculous in comparison with the 120 million words of the British National Corpus or the 150 million words of today’s Russian National Corpus, but it proves that even a relatively modest corpus may become crucial.

Apart from the size, another aspect of a language corpus is its organization. In other words, which options are offered by the corpus for searching? If it is just a collection of texts in Word format, only a search by a word form (or by a simple combination of words) is possible. For many studies, it may be insufficient, and besides, such a search often produces a great deal of “noise” (i.e. homonymous forms, which may in some cases outnumber by far the desired form). Especially for an African language, it is highly desirable to have a GLOSSED CORPUS, where each word (or, better, each morpheme) would be accompanied with its equivalent in a meta-language (English or French, depending on the official language of the country where the African language in question is spoken). And the corpus must be provided with a powerful search engine which should allow a search by form, by lexical and grammatical meaning, by their combinations…

A third point: a corpus should be available in the Internet and open to any user. Otherwise, it will remain just an enlarged version of an “individual corpus” and will never produce a desired effect.1

We can say that any language needs a public electronic corpus, but it would be more realistic to start with big languages, spoken by millions, used in a large scale in the education programs, having more or less considerable literature, taught in the universities outside Africa. In the Mande family, the major candidates may be Bamana

1 Cf. the effect produced by the creation of the Russian National Corpus, open for the broad public since April 2004, on the development of Russian linguistics: today, a majority of dissertations about Russian language are corpus-based.
and Maninka of Guinea, maybe Jula and Mandinka. Second grade candidates are Mende, Soninke, Soso, maybe Kpelle and Dan. However, to launch such a project is a question of a will: if there is a group of qualified people ready to work, for example, on a Vai corpus, this language may outdo Maninka and Bamana.

Let us consider the types of problems which will certainly arise once we decide to launch a project of a language corpus.

2. SOFTWARE

The software for a language corpus consists of two parts: a morphemic analyzer and a search engine.

For the morphemic analysis, the easiest solution (which does not necessarily means the most effective…) would be to use the Toolbox, a free software elaborated by the Summer Institute of Linguistics and used by thousands of field linguists all over the world. This software can be downloaded from the site www.sil.org, it is relatively easy to master (a one- or two-day training may be sufficient), and it can be tuned to the specifics of a language. Many people use it only for dictionary making, but its other function is to produce glossed texts correlated with the dictionary.

For this purpose, two files are created in the Toolbox: a text file and a dictionary file. In the dictionary file, all the morphemes (both roots and affixes) are inputted, and their equivalents in the meta-language are provided in special fields; in yet another field, a part-of-speech marker is given. When a phrase is introduced into the text file, a command “Interlinearize” makes the program divide each word into morphemes and find an equivalent to each morpheme in the dictionary file. As a result, the analyzed sentence is represented in a parsing field divided into morphemes, and in yet another field, each morpheme is given an equivalent in the meta-language. If a word or a morpheme is not in the dictionary, the program suggests to create a new entry for it in the dictionary.

An important option of the Toolbox, especially for a language whose literary norm is still unsteady, is introduction of variants of morphemes; those variants are also taken into account when a search is done for parsing. As a result, if a word is represented in a text with its secondary (non-standard, reduced, etc.) variant, it can still be identified and glossed.

The Toolbox works best of all with isolating and agglutinative languages, where morpheme boundaries are easy to establish. However, there are some procedures for handling deviations from the additive model of morphology as well: a fusional form can be given in the dictionary file with indication of its underlying form (in the latter, all the morphemes can be represented by their non-fusional variants, real or artificial). This option allows correct glossing of tonal morphology and “morpheme-operations”, such as consonant alternations.

If there are homonymic forms, Toolbox offers all the variants for a manual choice. It creates a problem for the languages rich in allomorphs and characterized with a complicated word structures (which is very typical of polysynthetic languages, but not only them). To some extent, this problem can be resolved through the introduction of “word formulae”: in this case, the program will not suggest combinations of morphemes contradicting the predefined patterns (for example, it will not try to add a tense or a voice suffix to a noun).

The Toolbox can be used also as a search engine, more powerful than, for example, a Word: a search can be done by fields; a concordance can be created. However, its principal shortcoming is lack of possibility to combine several parameters of a search. Therefore, its potential as a software for a language corpus is limited; it can be used for a pilot project of a language corpus, but for a corpus containing millions of words it
would be more expedient to look for another software. In this case, a question of convertibility of the Toolbox-parsed data to the other format will rise.

3. LINGUISTIC PROBLEMS

3.1. For many Mande languages, the problem of graphics remains unsolved. Firstly, a language may have more than one graphic basis. For example, Maninka and (to some extent) Bamana are written in Roman script and in Nko. Mandinka is written in Adjami (Arabic characters) and in Roman script, marginally in Nko. Which one to take for the corpus?

Even when only Roman script is used for a language, difficulties do not stop there, because there may be more than one variant of orthography: the Ivoirean and the Liberian orthographies for Dan; the old orthography (with letters ç, ö, ng, ny) and the new orthography (with e, æ, y, ĩ) for Bamana, etc.

Which of the graphic systems (and their variants) to choose? Paradoxically, there is no need to make a choice. Toolbox is able to establish equivalence of forms across the boundaries of graphic systems, and, certainly, other engines for morphological analysis can do it too. The logic of this process is very simple. In each entry of the dictionary file, the “main” graphic form of a word or a morpheme (for example, in Roman script) is inputted into the “lemma” field, and all its alternative graphic forms (in Roman script, in Nko, in Adjami…) are given in the field “variant”. So, all the alternative forms (sòbo, sòbò, sòbó, صبّو, صبو etc.) are regarded by the program as equivalents of the main form (sòbo ‘meat’), and it is the main form that will be given by the Toolbox in the line of parsing.

It is true that the abundance of alternative forms may increase the amount of the homography to be solved manually. However, this homography is not created by the Toolbox, it just reflects the homonymy which exists in the language. Important is to accept as the “main form” the one which reflects the relevant phonological characteristics and ignores predictable combinatory alternations.

So, for the words of two major tonal classes (High and Low/Rising) in Bamana and Maninka, it suffices to indicate only the lexical tones of their first syllables (high or low, respectively). Even in the context where a word looses its lexical tone (e.g., the tone of bilen ‘red’ in the attributive syntagma: só bilen ‘red house’, or the tone of sèn ‘foot/leg’ in the determinative compact syntagma: misi sèn ‘a cow leg’), we can still mark it (só bilen, misi sèn), as far as this tonal neutralization is predictable – provided that the tonal article is marked, the type of syntagma can be established without fault. As for the lexemes with “minor tonal patterns”, their tones should be marked on each syllable.

3.2. Another problem is the handling of the dialectal variants. One solution may be the same as for graphic and orthographic variants: dialectal forms can be put into the field “variant”, the program establishes their equivalence with the standard forms, and the latter appear in the parsing line. However, a large amount of dialectal forms may increase greatly the homonymy, for which reason it may be expedient to create a separate “dialectal corpus” (at least for the dialects which are considerably convergent from the standard form in what concerns the phonetic form of words).

The problem of errors in the texts (which may be very numerous in Mande languages, provided that those languages are, as a rule, not properly taught in schools) has its theoretical solution, but, unfortunately, it cannot be solved automatically; it increases the amount of manual work.
3.3. It is evident that a tagged and glossed electronic corpus cannot be created without a detailed and sufficiently formal description of the morphology of the language in question. In relation to the Mande languages, situation is uneven.

Bamana is an advanced language thanks to Gérard Dumestre’s grammar and numerous publications by many other authors; there are good bilingual dictionaries where tones are marked.

The position of Guinean Maninka is less favorable: there is a Nko dictionary with tonal markings, but it is monolingual, and its definitions often remain inaccessible to the non-native speakers of Maninka; there are linguistic studies written in Nko, a book by Claire Grégoire, and some other publications, but, generally speaking, the degree of formal analysis is rather insufficient.

Soninke has a good description of grammar (Yacouba Diagana’s dissertation; Ousmane Diagana’s book, etc.) and several dictionaries; unfortunately, there is no tonal dictionary.

There are many publications on Mandinka grammar, there is a small, but very good tonal dictionary by Denis Creissels et al., and several dictionaries without tones…

Whatever may be the situation of a language, the problem can be put in a different way: it is the work on an electronic corpus that will be a tremendous impetus for creation of a detailed formal description of the language, and it will provide abundant material for such a description.

4. AVAILABILITY OF TEXTS

To create an electronic corpus of texts in a language, one needs to have those texts, preferably in an electronic format. How can we get them?

A self-evident answer is: from the Internet. Schryver (2002) suggests to make searches by words of the language which are specific enough (so that they may not be confused with homonymous words of other languages), and to download texts from the sites found by this way. His experience is that in October 1999, ten days of work brought him texts for one million words in Swahili. In August 2002, he downloaded texts in 11 African languages of RSA for two million words during five days.

Undoubtedly, this source should be exploited to the maximum. However, my guess is that for African languages, except for Swahili and the languages of the Republic of South Africa, results will be much less encouraging. My impression is that by now, the number of sites where one can found texts in Bamana (let alone entirely Bamanophone sites) is rather scanty. At the end of August 2008, there were only 219 entries in the Bamana Wikipedia, and most of those pages were not verbose at all. So, the entry Buteli ‘bottle’ looks as follows: Fenw min be feere buteli kɔnɔ: ji, nɔnɔ, dolo (‘what is sold in bottles: water, milk, alcohol). Out of the 10 words of this tiny text, 3 are written with errors (the correct forms would be: fenw, feere, dolo). In many other entries of the Bamana Wikipedia, situation is even worse. It is easier to keyboard a text from paper than to correct such electronic texts. Another problem is that one cannot be sure that all the texts of the Bamana Wikipedia are written by native speakers of Bamana.

Another option is to address publishers and editors of periodicals asking them to provide electronic versions of their publications. Creators of the Russian National Corpus used this option very fruitfully: any sensible businessman understands that including his publication into the National Corpus is the best publicity. It is not clear however what will be the reaction of the not very numerous publishers dealing with books and newspapers in African languages: it is highly probably that they may ask for huge sums of money, and any argument concerning non-commercial character of the
project will be of no help. Probably, assistance of African colleagues may help the situation.

Finally, there is a very simple, but labor-intensive way: manual keyboarding of the texts available on paper. A corpus of 100,000 words makes 200 to 500 pages (depending on the size of characters, intervals, etc.). So, to create a one million corpus of a Mande language in one year is a feasible task of a small team of workers, and a glossed text of this size would be a great progress.

5. THE MANPOWER

In any case, creation of an electronic corpus of a Mande language is a very time-intensive job; it is a question of many thousands of hours of work. It is evident that one person can hardly produce even a small glossed and tagged corpus; it is a team job. And requirements to each member of such a team are high. He or she should:

a) have a good command of the language in question;

b) have a good command of the meta-language (English or French);

c) have a linguistic training sufficient for mastering the morphological tagging.

Even if we do not take into account the financial aspect, it is evident that to find such workers is not an easy task. As far as I can see the situation, it can be senior students of African languages in Russian or Western universities, or native speakers of the languages in question, literate in their mother tongue, with University diploma in philology. The team must be headed by a supervisor, a trained linguist, specialist in the language; he must solve all the difficulties which will certainly arise every day.

The “good side of the story” is that every member of the team will undoubtedly improve his or her practical and theoretical knowledge of the language.

6. FINANCES

To create a corpus of an African language is a task which is also money-consuming. In the situation where those languages are paid lots of lip-services, but a real interest in them is rather low, it will hardly be easy to find a sponsor in the sphere of business. However, it is not impossible to get an academic grant which will cover the costs of the initial stage of such a project (in fact, at this stage one could work even without funds!). Once a pilot model is created, a quest for sponsors will be easier. Finally, this kind of intellectual production may interest computer companies which are planning to create search engines and software environment in new languages; ministries of education of African countries, objectively interested in standardization of the languages of their countries...

Difficulties are enormous, but hope is there. It is for us to decide: are we ready to get engaged in this hard (but very promising) job, or shall we leave it for tomorrow?

References

