Corpora Mandeica: text corpora for Mande languages (West Africa)

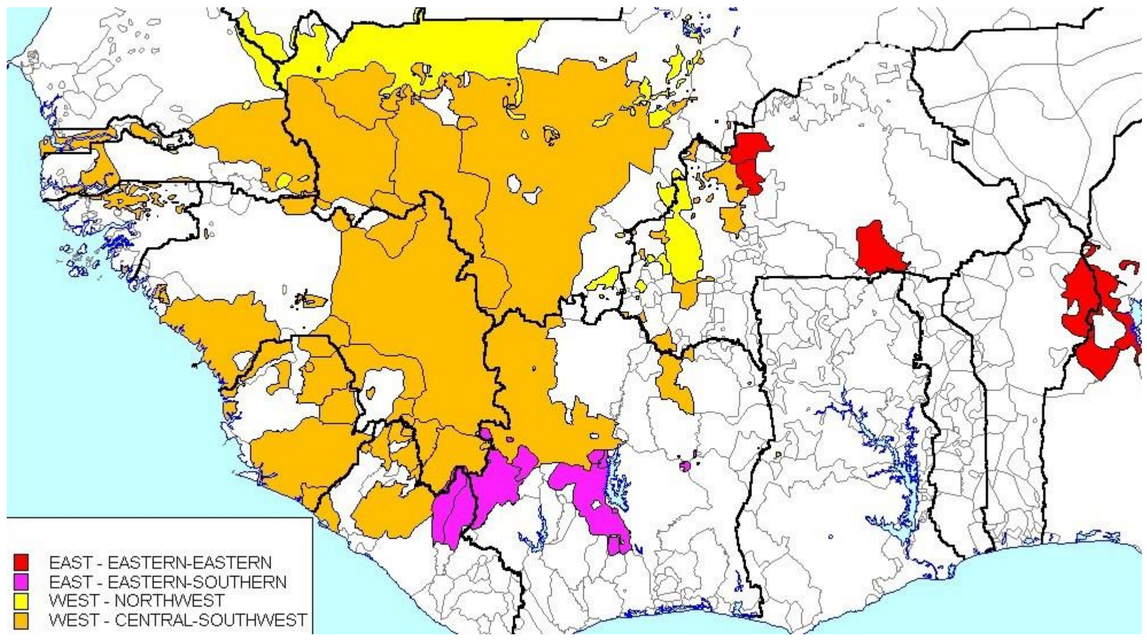
http://cormand.huma-num.fr/mandeica/

Valentin Vydrin

INALCO – LLACAN (CNRS UMR 8135) – St. Petersburg State University

Mande language family

More than 60 languages in West Africa. The best-known are Bambara, Maninka, Jula, Soninke, Susu, Mende.



Corpora Mandeica project: an overview

Corpora Mandeica is meant to provide Mande languages with electronic resources, openly accessible on line.

A full set of the tools for each language includes:

1. An annotated corpus of written texts, with subcorpora:
- disambiguated;
- non-disambiguated;
- parallel (a Mande language – French);
- syntactically annotated (in the Universal Dependencies format)

The annotation is performed through the means of the **Daba** software package developed by Kirill Maslinsky. The annotation concerns: *POS-tags, French/English/Russian glosses; eventually, tones.*

The search engine in the Internet: **NoSketchEngine**.

2. An electronic dictionary on line.
3. An electronic library.
4. A spellchecker (for Open Office, NeoOffice, Firefox).
5. Keyboards for smartphones and computers.

Corpus Bambara de Référence

http://cormand.huma-num.fr/

The project started in 2009, available on line since 2011.

In 2019, the quasi-totality of texts published in Bambara has been included into the Corpus.

- The entire size of the Corpus is over 11 million words.
- The disambiguated subcorpus size is about 1,140,000 words.
- The parallel corpus (Bambara – French): about 225,000 words.
- A small syntactically annotated corpus, about 12,000 words.

Other tools:

- An electronic library (about 500 books and about 1000 issues of periodicals), http://cormand.huma-num.fr/biblio/
- A spellchecker for OpenOffice and Firefox (developed by Andrij Rovenchak and Jean Jacques Méric).
- A Bambara-French electronic dictionary *Bamadaba,* > 13,000 lexemes, http://cormand.huma-num.fr/bamadaba.html

A project of a corpus-based Bambara Orthographic Dictionary is in the process of elaboration.

An audio-corpus, an automatic translator, a corpus-driven Bambara dictionary are visible in the perspective.

Corpus Maninka de Référence

http://cormand.huma-num.fr/cormani/

Available on line since 2016. Two subcorpora:

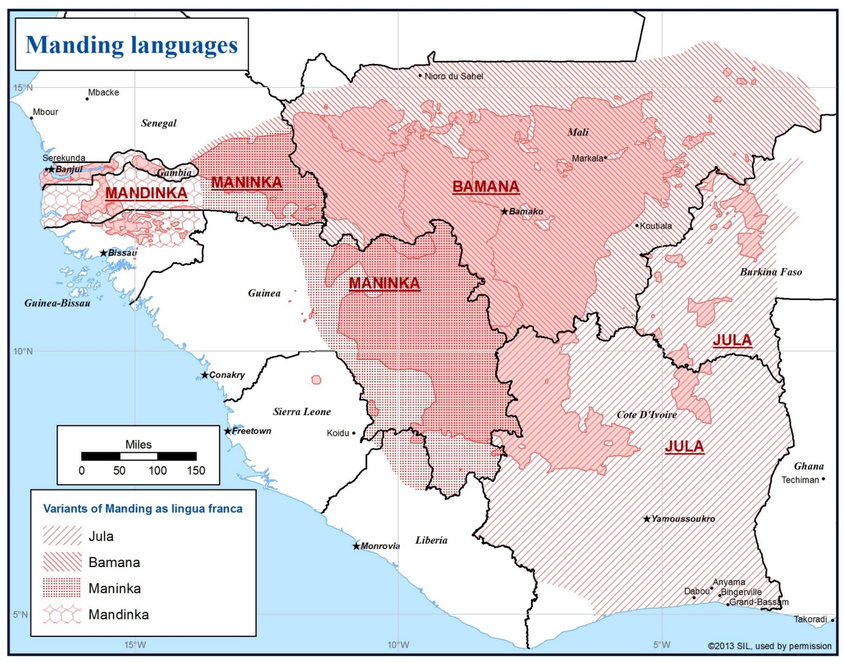
- Maninka Corpus (in Latin graphic), about 385,000 words;
- Nko Corpus, more than 3,220,000 words.

Both graphics (Nko and Latin) are mutually convertible. Search in both corpora is available in Nko and in Latin graphics.

Other tools:

- An electronic library, http://cormand.huma-num.fr/maninkabiblio/
- A spellchecker for OpenOffice and Firefox (developed by Jean Jacques Méric).
- A Maninka-French-English-Russian electronic dictionary *Malidaba,* almost 7000 entries,

http://cormand.huma-num.fr/cormani/dictionnaire.html



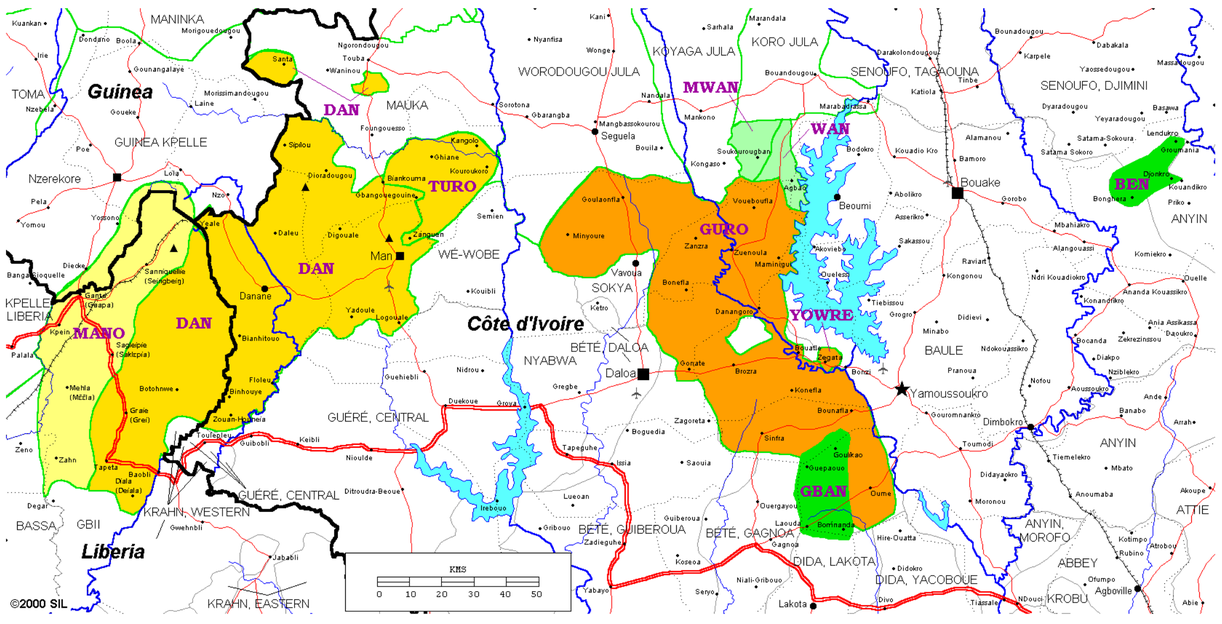
Corpus dan de l'Est, http://cormand.huma-num.fr/dan/

On line since April 2018.

Size: about 464,000 words.

Other tools: Dictionnaire dan de l'Est – français – anglais – russe, > 3600 entrées.

Bibliothèque électronique dan, http://cormand.huma-num.fr/danbiblio/



Corpus mwan, http://cormand.huma-num.fr/mwan/index.html

On line since March 2019.

Size: about 46,500 words.

Other tools: Dictionnaire mwan – français – anglais – russe, > 2300 entries.

Further perspectives:
Corpora for Guinean Kpelle, Mano, Guro, Soninke, Kakabe...