Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

CORPUS-DRIVEN BAMBARA SPELLING DICTIONARY

Vydrin V. F. (vydrine@gmail.com)
INALCO—LLACAN (CNRS, UMR-8135)—IUF,
St. Petersburg State University

Méric J. J. (jjmeric@gmail.com)
Paris

A model for the development of a corpus-driven spelling dictionary for the Bambara language is described. First, a list of about 4,000 lexemes characterized by spelling variability is extracted from an electronic Bambara-French dictionary. At the next stage, a script is applied to determine the number of occurrences of each spelling variant in the Bambara Reference Corpus, separately for the entire Corpus (more than 11 million words) and for its disambiguated subcorpus (about 1.5 million words). Statistics on the diversity of sources and authors are also obtained automatically. The statistical data are then sorted manually into two lists of lexemes: those whose standard spelling can be established statistically, and those requiring evaluation by expert linguists. Some difficult cases are discussed in the paper. At the final stage, a representative expert commission will discuss all those lexemes for which statistical data alone do not suffice to define a standard spelling variant, before taking a final decision on each. The resulting Bambara spelling dictionary will be published electronically and on paper.

Key words: Bambara language, spelling dictionary, spelling norm

DOI: 10.28995/2075-7182-2020-19-1180-1187

КОРПУСНОЙ ОРФОГРАФИЧЕСКИЙ СЛОВАРЬ ЯЗЫКА БАМАНА

1. Introduction[1]

Bambara is a Mande language (Manding < Western Mande < Mande < Niger-Congo) spoken by some 15 million people (taking into account both L1 and L2 speakers) in Mali, where it is understood by the great majority of the population. The language is taught in a number of primary and secondary schools, and some Bambara classes are also taught in universities; there is a broad network of literacy courses for adults; there is a Bambara-language written press. However, French, the language of the former metropolitan country (which has the status of "official language", while Bambara is one of 13 "national languages" of Mali), retains its leading position in the administration and education, and it unquestionably dominates throughout the written sphere.

The first Bambara orthography was created in 1967; it was reformed in 1982. An orthography guide was published [Anonyme 1979] which formulated some basic rules.[2] However, Bambara written practice shows that orthographic variability remains very high even in published texts, and this variability usually stems from the fact that numerous cases are simply not covered by the rules.

In 2011, a Bambara Reference Corpus was developed and put online in open access [Vydrin 2013]; [Vydrin, Maslinsky & Méric 2011]. By February 2020, the great majority of published Bambara texts were included in the Corpus, and its size now exceeds 11 million words, of which more than 1.5 million belong to the manually disambiguated subcorpus. For the annotation of the Corpus, the electronic Bambara-French dictionary Bamadaba is used [Bailleul et al. 2011], which is also available on line. Most texts are accompanied by metadata (the name of the source, the name of the author of the document, etc.).

Tone marking is absent in the Bambara orthography. However, in the Bambara Reference Corpus, tonal diacritics are provided: these are added in the process of the automatic annotation of tokens.

The availability of the Bambara Reference Corpus has radically changed the situation in Bambara language studies, for it considerably facilitates the checking of hypotheses on representative data.

In this paper, we present the project of a Bambara spelling dictionary that we are currently developing in collaboration with Malian colleagues on the basis of the Bambara Reference Corpus.

---

[1] This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris—ANR-18-IDEX-0001.

[2] Bambara belongs to the Manding language group, along with Maninka, Jula, Mandinka and many other varieties spoken in neighbouring countries by tens of millions of speakers. Since the 1960s, numerous expert meetings on the standardization of orthography across Manding varieties have taken place, but they have had no tangible effect on Bambara orthographic practice, and this aspect will not be taken into account in the present paper.

2. The work model

The idea underlying the project is very simple: most of the time, if more than one spelling variant of a word appears in written texts, it is the most frequently used variant that should be retained (except in hypothetical cases where the most frequent variant violates an established orthographic rule).

The Bamadaba dictionary is taken as the starting point for this project: spelling variants are included more or less systematically in this dictionary, which is a prerequisite for the automatic morphological parsing of texts. First of all, all lexemes which have orthographic variants[3] are automatically extracted from the dictionary and represented as a spreadsheet. It turns out that among the 13,365 lexemes in the Dictionary, almost 4,000 show some kind of variability in their spelling.

A script has been written to obtain statistical data for these variable items, i.e., the number of occurrences of each spelling variant. For each lemma and its derived forms, the script effectuates automatic queries to the NoSketch-Engine Corbama corpus: total number; number of identified (unambiguous) forms; for each form, the number of sources and the top 6 sources (with respect to number of occurrences), the number of authors and the top 6 authors using this form.

As was mentioned in the Introduction, the Bambara Reference Corpus contains texts comprising more than 11 million tokens; of these, texts totalling more than 1.5 million tokens have been manually disambiguated. The remaining texts (about 9.5 million tokens) undergo automatic morphological analysis (based on the Bamadaba dictionary and on a set of morphological rules), producing about 70% of tokens with two or more variants of analysis. Then a program carrying out automatic disambiguation (developed by J. J. Méric) is applied, based on a collection of more than 30,000 Bambara collocations and formal rules of sentence structure; this brings the ambiguity rate down to some 28%. This program also undoubtedly produces some mistaken disambiguations: their prevalence can be tentatively evaluated at 1% (though this may vary depending on the genre of the text). Automatic disambiguation is used as a supplement to a process of manual (human-operated) disambiguation; errors are then spotted and counts compared to the numbers for pre-disambiguated words. A more detailed analysis of the automatic disambiguation process and the error rate is beyond the scope of the present paper.

The statistics of occurrences for each spelling variant are represented by two figures: first, the total number (including ambiguity), and second, the number of "disambiguated occurrences", which reflects not only the occurrences in the manually disambiguated subcorpus, but also the automatically disambiguated occurrences in the rest of the Corpus. These data are integrated into the same spreadsheet.

As a means of preventing the "whelks problem" [Kilgariff 1997: 138–139], the same script retrieves the number of sources (and the names of the sources) where the spelling variants occur, and the names of the authors using those variants: this

[3] Bambara is a predominantly isolating language, with some elements of agglutination. Therefore, irregular word-forms not automatically derivable from the lemma are almost absent. When we speak of spelling variants, we usually mean variability in the basic forms of lexemes (lemmata).

is to check for cases where a variant used by one prolific author might numerically outweigh other variants used by numerous authors.

A sample of the statistical data is represented in Table 1.

Table 1

translation	variant	Non-disamb.	Disamb.	Sources	Authors
finger	bólokɔni	135	71	26	9
	bólonkɔni	245	105	63	23

In the column labelled "Non-disamb.", the number of occurrences in the entire Corpus is indicated; the column labelled "Disamb." indicates the number of occurrences which have been disambiguated (whether manually or automatically). In the column "Sources", we find the number of sources where the variant occurs in the disambiguated subcorpus, and in "Authors", the number of authors who have used the variant. In the spreadsheet we also have columns where the names of the sources and the names of the authors are listed for the disambiguated occurrences (up to 6 sources; if there are more sources available, the top 6 sources by number of occurrences are mentioned).

In the rather simple case represented in Table 1, it is the second form (*bólonkɔni*) that will be given preference, as it shows a higher number of occurrences, sources, and authors.

After the automatic extraction of the statistical data, we pass to its preliminary manual evaluation. This is intended to sort the variable lexemes into two sets: those whose statistical data proves to be sufficient to identify a standard variant, and those requiring further discussion among experts.

Each spelling variant will be provided with one of the following markers:

- r (recommended form);
- t (tolerated form);
- e (form to be avoided);
- d (debatable variant);
- n (irrelevant, e.g. forms with elision).

If at least one variant of a lexeme is marked with the index *d,* all other variants are also marked with *d*.

The lexemes will be subdivided into two sets: those whose variants are indexed with *r, t, e* (and can therefore be regarded as needing no further consideration) and those whose variants are indexed with *d* or *0* (and will thus be placed in the set "to be discussed").

Another option would be to further subdivide the former set: first, the lexemes whose variants bear the indexes *r* and *e*; second, those for which *r* and *t* appear. In this case, the latter group could be also submitted to the consideration of the expert commission.

3. Anticipated questionable cases

Even at the present stage, when we are beginning to sort the lexemes and their spelling variants, we already envisage certain types of cases where decisions will be difficult to take. Let us consider some typical cases (many others will certainly emerge in the course of the work).

3.1. Non-decisive statistical data

When one orthographic variant outdoes the other(s) at least tenfold with regard to number of attestations, as in the cases represented in Table 2, decisions are easy to take. So, for these particular lexemes, we can select as standard spelling variants the following: *dánkan* 'bank', *bɔ̀nɛ* 'misfortune', *búran* 'in-law', *cɛ́mancɛ* 'middle'.

Table 2

translation	variant	Non-disamb.	Disamb.	Sources
bank, shore	dánkan	1329	65	49
	dángan	22	3	2
misfortune	bɔ̀nɛ	1296	81	44
	bùnɛ	3	1	1
	bɔ̀nɔ	59	0	0
in-law	búran	666	177	93
	bíran	60	10	5
middle	cɛ́mancɛ	961	595	387
	cámancɛ	41	40	22

However, there are also lexemes whose spelling variants do not differ greatly in their statistics: for examples see Table 3.

Table 3

translation	variant	Non-disamb.	Disamb.	Sources
bread	búru	291	1	1
	nbúru	35	1	1
	nbúuru	69	69	37
	búuru	338	96	29
to bite	cín	485	402	182
	kín	1,151	129	51
life	díŋɛnatigɛ	901	151	121
	díŋɛlatigɛ	847	96	74
	díyɛnnatigɛ	62		
	jyénlatigɛ	29	8	3

Let us consider each lexeme in this small sample.

- 'bread': the elevated number of occurrences of *búru* in the non-disambiguated corpus might be due to homonymy (or quasi-homonymy) with *búru* 'government', *búru* 'trumpet', and *bùru* 'sediment'. *nbúru* is relatively rare (however, the plausible hypothesis that among the occurrences of this form one would find numerous instances of *nbúru* 'government' proves to be wrong; in fact there are none). The main candidates for the standard form are *nbúuru* (which prevails in terms of number of sources) and *búuru* (which wins out in number of occurrences).
- 'bite': the discrepancy between the numbers for the disambiguated and non-disambiguated subcorpora might be explained by the quasi-homonymy of the variant *kín* with the word *kìn* 'area, district'. In any case, both variants, *kín* and *cín,* are well represented in the Corpus.
- 'life': of the four variants, two (*díɲɛnatigɛ* and *díɲɛlatigɛ*) substantially outdo the other two (*díyɛnnatigɛ, jyénlatigɛ*).

The variants *nbúuru* and *búuru, kín* and *cín, díɲɛnatigɛ* and *díɲɛlatigɛ* can be suggested as more or less equally acceptable (or one of the variants can be marked as "recommended" and the other as "tolerated"). It is yet to be decided whether all those cases where more than one variant is retained will be included in the set assigned for evaluation by the commission of experts.

There are also false positive cases, where the statistical data seem unequivocal but other considerations may prevent us from taking final decisions at the preliminary selection stage. Let us consider two typical cases (among many others).

3.2. Long vowels

In Bambara, long and short vowels are phonologically contrastive only in the non-final syllable of a foot, e.g. *bára* 'dancing ground' : *báara* 'work'. Even in this position, vowel length may be unstable among speakers in some words. In Table 4, statistics are given for two words of this type.

Table 4

translation	variant	Non-disamb.	Disamb.	Sources
Bozo (ethnic group)	bòso	437	153	84
	bòoso	4	4	4
to call	wéle	10,099	4,868	2,595
	wéele	295	142	79

On the basis of the statistics alone we should reject both forms with long vowels (*bòoso, wéele*). However, it seems that written practice here diverges from the oral norm: in a pronunciation experiment [Vydrin 2020], both these words were pronounced with long vowels in the initial syllable by about half of the speakers. It seems that the forms with long vowels reflect the original pronunciation, and their retention (as optional variants) is worthy consideration.

3.3. Word-initial prenasalization

In Bambara, nouns with a prenasalized initial consonant are relatively numerous, e.g.: *nkàsa* ~ *nkàsan* [ŋkàsa ~ ŋkàsã] 'herbaceous plant Ipomoea muricata', *nfírinfirin* [mfírimfirĩ] 'butterfly', *nsàna* [nsàna ~ nzàna ~ zàna] 'proverb'. The nasal element probably goes back to an archaic prefix *N- whose reflexes can be found in various languages of the Manding group, but also outside this group [Vydrine 1994].[4] Prenasalization in words of this type is often unstable across Bambara dialects, and even among speakers of Standard Bambara. In table 3, corpus statistics are given for variants of four words of this type.

Table 5

translation	variant	Non-disamb.	Disamb.	Sources
Guinea worm	nsɛ̀gɛlɛn	3	3	2
	sɛ̀gɛlɛn	104	13	6
	sègelen	76	20	4
thief	nsòn	707	116	24
	sòn	938	649	201
date palm	ntámàro	67	56	33
	támàro	49	48	24
fishing hook, fishing rod	dóolen	3	3	2
	ndóolen	0	0	0
	dólen	55	55	16
	ndólen	1	1	1
	dɔ́len	15	0	0

If we follow the statistics, the following variants should be considered "winners" and therefore selected as standard variants: *sɛ̀gɛlɛn* 'Guinea worm', *sòn* 'thief', *ntámàro* 'date palm', *dólen* 'fishing hook'. The only variant with a prenasalized initial consonant, *ntámàro,* is an Arabic loan. In Arabic, the word for 'date palm' is *tamr,* with no initial nasal. Therefore, the prenasalization in the word *ntámàro* most probably results from the regressive nasal spread.

Paradoxically, in this set, the forms retaining archaic features (which a linguist may wish to select as standard) display low scores. It seems appropriate to include at least some of these lexemes in the "discussion list" which will be submitted for the consideration of the expert commission.

4. Spelling variability beyond the dictionaries

There are some other types of spelling variability that are difficult to detect from the dictionaries. These are mainly cases where an expression can be written as a single word or separately. Among these, one can mention numerals divisible by 10 from 30 to 90 (e.g.

[4] There are also some verbs and adjectives with prenasalization; however, these are not numerous. Prenasalization in these items is not due to archaic morphology; it can be explained by a process of regressive nasality spread.

bisaba vs. *bi saba* ‘30’); compound postpositions (e.g., *à ɲémà* ‘before it’ vs. *à ɲɛ mà* ‘as it should be’); preverbal adverbs and noun groups converted to preverbal adverbs, etc.

Such cases can be enumerated through examination of published Bambara texts. In fact, many of them have been detected in the course of work on the disambiguation of texts for the Bambara Reference Corpus. These variants can be also checked for their frequency in the Corpus and treated along the same lines as the other cases of variability.

5. The final stage: an expert commission

After the preliminary selection (carried out by the authors of the present paper), a commission of Bambara language experts will give their judgement on those variants whose statistical data prove indecisive. This commission will be composed, besides the members of our working group, of Malian linguists representing research institutions (such as Académie Malienne des Langues, AMALAN, and Université des Lettres et des Sciences Humaines de Bamako, ULSHB) and official bodies related to national language education and language policy.

After the approval of the standard variants of lexemes, we intend to publish the results in the format of a spelling dictionary of Bambara, both in electronic form (where all statistical data can be displayed and dynamically searched) and on paper, as a reference book for Bambara language teachers and students.

If the work on the Bambara spelling dictionary proves successful, the question may arise of transforming the mechanism described in this paper into a permanent one.

References

1. *Anonyme* (1979) Guide de transcription et de lecture du bambara. Bamako: DNAFLA.
2. *Bailleul, Charles, Artem Davydov, Anna Erman, Kirill Maslinsky, Jean-Jacques Méric & Valentin Vydrin* (2011) Bamadaba : Dictionnaire électronique bambara-français, avec un index français-bambara. http://cormand.huma-num.fr/bamadaba.html.
3. *Kilgariff, Adam* (1997) Putting frequencies in the dictionary. International journal of lexicography 10(2). 135–155.
4. *Vydrin, Valentin* (2013) Bamana Reference Corpus (BRC). In Chelo Vargas-Sierra (ed.), Procedia—Social and Behavioral Sciences: Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013), vol. 95, 75–80. Alicante: Elsevier. http://www.sciencedirect.com/science/journal/18770428.
5. *Vydrin, Valentin* (2020) Vowel elision and reduction in Bambara. Italian Journal of Linguistics 32(1).
6. *Vydrin, Valentin, Kirill Maslinsky & Jean-Jacques Méric* (2011) Corpus Bambara de Référence. http://cormand.huma-num.fr/index.html.
7. *Vydrine, Valentin* (1994) Traces of Nominal Classification in the Mande Languages: the Soninke Evidence. St. Petersburg Journal of African Studies 3. 63–93.