Masadennin (The Little Prince in Bamana)
Experimental online concordance with parallel French and English texts

Masadennin (Le Petit Prince) en bambara : une concordance expérimentale en ligne avec les textes parallèles français et anglais
Masadennin («МАЛЕНЬКИЙ ПРИНЦ» НА БАМАНА): ЭКСПЕРИМЕНТАЛЬНЫЙ ОНЛАЙНОВЫЙ КОНКОРДАНС С ПАРАЛЛЕЛЬНЫМИ ФРАНЦУЗСКИМ И АНГЛИЙСКИМ ТЕКСТАМИ

Andrij Rovenchak and Solomija Buk
Masadennin (The Little Prince in Bamana)

Experimental online concordance with parallel French and English texts

Masadennin (Le Petit Prince) en bambara : une concordance expérimentale en ligne avec les textes parallèles français et anglais

Masadennin («МАЛЕНЬКИЙ ПРИНЦ» НА БАМАНА): ЭКСПЕРИМЕНТАЛЬНЫЙ ОНЛАЙНОВЫЙ КОНКОРДАНС С ПАРАЛЛЕЛЬНЫМИ ФРАНЦУЗСКИМ И АНГЛИЙСКИМ ТЕКСТАМИ

Andrij Rovenchak and Solomija Buk

We are grateful to Valentin Vydrin for providing a copy of the Bamana translation of the novella and for his feedbacks during the compilation of the concordance.

0. Introduction

1. Text corpora being large collections of texts in the electronic form supplemented by linguistic annotation have become nearly a principal tool in linguistic studies as they provide solutions to a variety of tasks. A parallel text corpus is composed of source texts and their translations into one or more different languages (McEnery, Xiao, Tono 2006:47; Teubert & Čermáková 2007:73). Such corpora are a powerful tool for interlingual studies in the domain of contrastive linguistics, translatology, linguistic typology, stylistics, linguistics of text, dictionary compilation, etc.

2. The documentation of African languages in the form of electronic corpora remains poor. Minor exceptions are Swahili (Schryver 2002; Hurskainen 2004), Amharic (Asker et al. 2011), South African languages (Schryver 2002). References to running corpus projects, though with no details, can be found also for Hausa¹, Yorùbá², Somali³, Lingala and Cilubá⁴, and sporadically for some other languages. A recent project of Bamana text corpus⁵ (Vydrin 2008; Davydov 2010; Vydrin 2011; Davydov 2011) is the only one in the field of the Mande languages.
In the present paper we report the launch of an experimental online tool for the comparative analysis of Bamana texts with *The Little Prince (Le petit prince)*, a famous 1943 novella by Antoine de Saint-Exupéry, used as an example. The Bamana translation by Bukari Jara was published in 1989 under the title of *Masadennin* (De Sén Tęgiziperi 1989).

This novella is an excellent text for contrastive linguistic studies. Being translated into about two hundred languages and dialects from different language families, it is one of the most translated [secular] texts, if not the most translated indeed. Cross-linguistic comparisons based on *The Little Prince* were made by Bahrens (2005), Stolz (2007), Rovenchak & Buk (2011).

The original French text together with the English translation was used in this study alongside the Bamana text. This material should be sufficient to study the functionality of the concordance tool and learn about its possible improvements before similar techniques are applied to a wider set of texts.

### 1. Text alignment issues

Texts in parallel corpora can be aligned on different levels: words, phrases or clauses, sentences, paragraphs, chapters, etc. The word level is suitable only for closely related languages and must be done automatically for texts of any significant size, which requires well-elaborated translation-oriented software. On the other hand, the chapter alignment is of little help for any serious analysis. Of the remaining intermediate options, the sentence level is the most typical alignment style for parallel corpora.

A definition of the sentence suitable for an automated text processing is required. The most natural approach is to apply text splitting on the following punctuation marks: ‘.’ (period), ‘!’ (exclamation mark), ‘?’ (question mark), and ‘...’ (ellipsis) *if the following word starts with a capital letter* (Kelih & Grzybek 2005). Some drawbacks of such a definition are caused by the use of a period in abbreviations, including initials, but such situations can be handled quite easily (cf. Rudolf 2004:57-71). While capitalization is present in the analyzed text, the above definition should be simplified when working with Bamana texts in general as the use of capital letters is not ubiquitous there. Note that neither the use of capital letters nor punctuation rules were regulated by the 1982 orthography standard (cf. Skattum 2000). Bamana texts written entirely in lowercase are not rare. One should note that a wider set of sentence terminators can be also used when aligning parallel corpus. For instance, ‘:’ (colon) and ‘;' (semicolon) were added when aligning the English–Inuktut Parallel Corpus (Martin et al. 2003). Such approach can be useful in particular for texts with numerous occurrences of direct speech.

In the present study, we stay on the paragraph level so far, with the sentence level to be achieved further. It appears that the alignment of paragraphs cannot be fully automatized, but the required manual work is not very time-consuming.

We introduce two tags to control the text flow over paragraphs, namely `<parbreak>` and `<noparbreak>`. The first one is used if there is a paragraph break in the Bamana or English text but not in the French text. Compare the last paragraph(s) of Chapter IV in the three language versions given in Table 1. Note that the tag `</s>` is used to mark the end of a sentence.
<table>
<thead>
<tr>
<th>Bamana</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bari n’ta fe: magow ka n’ka kitabu manamana kalan. ››</td>
<td>Car je n’aime pas qu’on lise mon livre à la légère. ››</td>
<td>For I do not want any one to read my book carelessly. ››</td>
</tr>
<tr>
<td>Nin ko kargow maana bôlë be: gan n dusu la cogo min. ››</td>
<td>J’éprouve tant de chagrin à raconter ces souvenirs. ››</td>
<td>I have suffered too much grief in setting down these memories. ››</td>
</tr>
<tr>
<td>Ni n’bë n jiya k’a fo yan a tun be: cogo min, n’ta fe ka jîne: a ko de. ››</td>
<td>Si j’essaie ici de le décrire, c’est afin de ne pas l’oublier. ››</td>
<td>If I try to describe him here, it is to make sure that I shall not forget him. ››</td>
</tr>
<tr>
<td>O ni fana n be: se ka kei i ko maga koroba mënu mago te: fosi la tugun jatedew ko. ››</td>
<td>Et je puis devenir comme les grandes personnes qui ne s’intéressent plus qu’aux chiffres. ››</td>
<td>And if I forget him, I may become like the grown-ups who are no longer interested in anything but figures... ››</td>
</tr>
<tr>
<td>O fana de koson n ye kiriyan jëma buwati kelen san, a ni kiriyan gansaw. ››</td>
<td>C’est donc pour ça encore que j’ai acheté une boîte de couleurs et des crayons. ››</td>
<td>It is for that purpose, again, that I have bought a box of paints and some pencils. ››</td>
</tr>
<tr>
<td>N koñi be: n jiya ka jaw ke: minw bônë be: u tigi fe: kosebe. ››</td>
<td>J’essaierai, bien sûr, de faire des portraits le plus ressemblants possible. ››</td>
<td>I shall certainly try to make my portraits as true to life as possible. ››</td>
</tr>
</tbody>
</table>

**Table 1: A sample of aligned texts**
<table>
<thead>
<tr>
<th>Bamana</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
</table>
| ... N tun b’a fe k’a don ni m’go faamuyalen don.</s> Nka a tigilam’go tun be n jaabi tuma be: "Banfula don".</s> O la sa mini’tanka wo, turfinko wo, n tun t’o si kuma fo a ye tugun.</s> ... | ... Je voulais savoir si elle était vraiment compréhensive.</s> Mais toujours elle me répondait : "C’est un chapeau."</s> Alors je ne lui parlais ni de serpents boas, ni de forêts vierges, ni d’étoiles.</s> ... | ... I would try to find out, so, if this was a person of true understanding.</s> But, whoever it was, he, or she, would always say: “That is a hat.”</s> Then I would never talk to that person about boa constrictors, or primeval forests, or stars.</s> ...

A single paragraph in French corresponds to two paragraphs in Bamana and three paragraphs in English. While this situation is not quite typical for the text in general, such differences arise mostly when direct speech is involved, due to different approaches for its written representation, as in the example below (Table 2). Only in the English text does the direct speech trigger a separate paragraph.

Table 2: Aligned excerpts with different representations of direct speech
The <noparbreak> tag is placed at a point where the French text breaks into paragraphs but the Bamana or English text flows continuously. In the example given in Table 3 paragraphs are separated by blank lines and adjusted for convenience of the Reader. In the corpus, paragraph breaks are simple line feed symbols.

<table>
<thead>
<tr>
<th>Bamana</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>N ye n ka dakabanabaara jira mọọ korọbaw la, k’u niinja ni n ka ja b’u lasiran.</td>
<td>J’ai montré mon chef-d’oeuvre aux grandes personnes et je leur ai demandé si mon dessin leur faisait peur.</td>
<td>I showed my masterpiece to the grown-ups, and asked them whether the drawing frightened them. &lt;noparbreak&gt;</td>
</tr>
<tr>
<td>U ye n jaabi: &quot;Banfula dun bɛ se ka mọọ lasiran cogo di?&quot;</td>
<td>Elles m’ont répondu : &quot;Pourquoi un chapeau ferait-il peur ?&quot;</td>
<td>But they answered: &quot;Frighten? Why should any one be frightened by a hat?&quot;</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O la sa, ne dese ᵢa n timanadiya, bari n tun korọtọlen don ka n ka bolifẹn moto τi waraka; n ye ja in nkininkana.</td>
<td>Alors, faute de patience, comme j’avais hâte de commencer le démontage de mon moteur, je gribonnai ce dessin-ci.</td>
<td>By this time my patience was exhausted, because I was in a hurry to start taking my engine apart. So I tossed off this drawing.</td>
</tr>
<tr>
<td>N ko:</td>
<td>Et je lançai :</td>
<td>And I threw out an explanation with it.</td>
</tr>
<tr>
<td>-- Nin ye wagande ye. I bɛ saga min fe, a b’a kọọ.</td>
<td>-- Ça c’est la caisse. Le mouton que tu veux est dedans.</td>
<td>“This is only his box. The sheep you asked for is inside.”</td>
</tr>
</tbody>
</table>

Note that some English examples contain more sentences comparing to the two other language versions.

An XML-like markup is used in the corpus. Additionally to the three tags described above, some other markup is present in the texts. Except for self-evident meaning of tags <title>, <author>, <dedication>, the tag <c> marks chapters. Below, the beginning of the Bamana text with the markup is shown. Note that texts are coded using Unicode (UTF-8).
2. Concordance implementation

The concordance can be accessed online at

The screenshot of the main page is shown in Figure 1.

Fig. 1. The screenshot of the main page of the concordance

Technical realization of the online concordance is partly based on some previously
developed CGI-scripts in Perl, which were used for different purposes. In future, texts in
the corpus can be automatically converted for usage in some standard tool, e.g., the IMS
Open Corpus Workbench, as soon as a proper hosting is found for it.

One can work with the concordance in two modes: the parallel text output and the output
of an interlinearized Bamana text.

2.1. The search options for the parallel text output include:

• entire wordform (if this is selected, one can input several wordforms separated by a space,
such a query displays the results where at least one wordform is found);
- words starting with a given character sequence;
- words containing a given character sequence (this requires at least two characters in the query);
- words ending in a given character sequence;
- exact phrase search.

Presently, only the Bamana text can be searched. Expansion for the two other languages can be implemented in future without much effort.

A simple type assistant is provided to facilitate (via the copy-paste procedure) the input of the non-ASCII characters, namely: Ė / ɛ, ԋ / ɔ, N / ñ, Ʌ / ɲ, ņ / ŋ. The latter is a now-obsolete letter; in the novella text it appears only in the root ʃɛ (modern ɛ̀) ‘hen’ in the words ʃɛw ‘hen-PL’ and ʃɛɲinitɔ ‘≈ looking-for-hen’.

**Fig. 2. A sample output with three parallel text versions**

The search queries can also contain some standard regular expressions. Namely, wildcards are used: a period ‘.’ or question mark ‘?’ matches a single character, while an asterisk ‘*’ matches any (including zero) substring. For instance, the query ‘m.g.’ matches both magɔ and mago. To match the respective punctuation marks, an escape character ‘\’ is required: ‘\.’ for the period and ‘\?’ for the question mark.

A sample output for the query “wordforms starting with masa” is shown in Figure 2.

For the concordance, we use the form of output known as KWAL (key word and line), which can allow several lines of context either side of the key word (McEnery & Wilson 2001:198). The sentence containing the matched item is shown together with the sentence immediately preceding it and the sentence immediately following it (unless the item appears in the last sentence of a paragraph).

By ticking appropriate boxes one can choose the option of showing either both the French original and the English translation alongside the Bamana translation, one (French or English) text alongside the Bamana translation, or the sole Bamana text.

As the paragraphs can contain different number of sentences, we used a simple algorithm to define the position of a corresponding sentence in the French or English text based on
the position of the Bamana sentence. Assume the respective paragraph in Bamana contains \( N_{BM} \) sentences, the French one has \( N_{FR} \) sentences, and the English one has \( N_{EN} \) sentences. Let the matched item be found in the \( j \)th sentence of the Bamana paragraph, denote it as \( j_{BM} \). We then calculate the respective points \( j_{FR} \) and \( j_{EN} \) in the French and English text as follows:

\[
j_{FR} = j_{BM} / N_{BM} \times N_{FR}\;
\]

\[
j_{EN} = j_{BM} / N_{BM} \times N_{EN}\;
\]

and round them to the closest integer. The output of the concordance then consists of the sentences number \( j_{BM} - 1, j_{BM}, j_{BM} + 1 \) (Bamana text); \( j_{FR} - 1, j_{FR}, j_{FR} + 1 \) (French text); and \( j_{EN} - 1, j_{EN}, j_{EN} + 1 \) (English text). Note that the zeroth position corresponds to the last sentence of the preceding paragraph.

While such an approach results in some inconsistencies of the parallel texts, the output scope of three sentences significantly eliminates such a drawback. Another typical solution is the insertion of empty sentences to equal the paragraphs lengths (cf. Yang & Li 2004), but this must be done manually so far.

2.2. For the mode of interlinearized output the Bamana text was preprocessed using the Bamadaba lexical database derived from the electronic version of Bailleul’s dictionary (Bailleul 2000). Such a procedure is generally known as tagging as it results in each text item (typically word, number, and punctuation mark) being supplied with a tag, i.e. a set describing item’s part of speech, gloss, lemma, etc. depending on the need of the study. We will not go deep into technicalities of the tagging here as this goes beyond the scope of the paper. The following sentence is provided as an example: ‘And that is how I made the acquaintance of the little prince’.

The tags read \(<\text{PoS}&\text{gloss}&\text{lemma}>\), with variants separated by vertical stroke “|” and compound structure shown by means of “\(=\)”.

The search options in the interlinearized output mode slightly differ from the first mode of parallel output. Namely, the exact phrase search is substituted with the following three options:

- part of speech;
- gloss (presently, given in French);
- lemma.
The output for the query étoile ‘star’ in the gloss field is shown in Figure 3. Note how the compounds are listed in the respective fields. For simplicity, the search was run as .toile using a wildcard to substitute an accented letter.

Since the Bamana text has no tone marks, there would be typically several options to gloss a word. Taking into account also possible homonyms, this would increase the number of glossing variants enormously in some cases. A special parameter was set to control this number; its default value is 3 and can be changed before running a query if desired.

As one can see from the given sample, the tagging procedure was quite straightforward, for instance, the contracted forms (like b’a, n’a, etc.) were not recognized. The words missing from the dictionary are marked with DUMMY in the gloss field and undef in the part-of-speech field.

3. Prospects toward an automated sentence alignment.

The problem of an automated sentence alignment is the next task to be solved. As the analysis presented in Section 1 shows, the paragraph level requires manual preprocessing in some cases. This stage can be skipped by proceeding directly to the sentence level. Algorithms of sentence alignment are extensively discussed in the literature (cf. Gale & Church 1993; Melamed 1999; Fattah et al. 2007 and references therein). In this work, we would like to consider the alignment procedure based on the so-called anchor items. As suggested below, a hybrid approach is used to define anchor items, with statistical preprocessing of texts and further linguistic analysis of the obtained data.

Two types of anchor items can be identified (Gale & Church 1993; Martin et al. 2003). First, these are words containing numbers or pure numerical sequences of characters. Such items are not frequent in ordinary texts, and where the same item occurs in parallel
texts it is the sign to identify the respective sentences as corresponding to one another. Very frequent words are the second type of anchor items. They can be defined from frequency analysis of the languages involved in the alignment problem. We have compiled the data on word frequencies in the three language versions of *The Little Prince*. The list of the most frequent words is given in Table 4. The number of occurrences is very close in all three languages for the following three items of this list, which can be treated as anchor words, namely:

- *n* — je, j’, mon, moi, ... — I, me, my, ...;
- masadennin — [petit] prince — [little] prince;
- *nka* — mais — but.

Note that in some cases one needs to take inflection into account (i.e., pronominal declension in English and French or possessive marker in English) to obtain a proper mapping between languages. The presented data would be typical for fiction texts, in particular with respect to characters’ names, cf. (Buk & Rovenchak 2007). A preprocessing of texts is required to obtain relevant statistical data before proceeding to the alignment procedure.

### Table 4: Number of occurrences of top 25 most frequent words in three language versions of *The Little Prince*. Possible anchor words are in boldface.

<table>
<thead>
<tr>
<th>rank</th>
<th>Bamana</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>ka</td>
<td>le</td>
<td>the</td>
</tr>
<tr>
<td>2.</td>
<td>n</td>
<td>de</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>je, j’, mon, moi, ...</em> <strong>758</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>I, me, my, ...</em> <strong>883</strong></td>
</tr>
<tr>
<td>3.</td>
<td>a</td>
<td>je</td>
<td>to</td>
</tr>
<tr>
<td>4.</td>
<td>ye</td>
<td>et</td>
<td>a</td>
</tr>
<tr>
<td>5.</td>
<td>bɛ</td>
<td>il</td>
<td>and</td>
</tr>
<tr>
<td>6.</td>
<td>tun</td>
<td>les</td>
<td>of</td>
</tr>
<tr>
<td>7.</td>
<td>ko</td>
<td>un</td>
<td>that</td>
</tr>
<tr>
<td>8.</td>
<td>i</td>
<td>la</td>
<td>you</td>
</tr>
<tr>
<td>9.</td>
<td>ni</td>
<td>petit</td>
<td>is</td>
</tr>
<tr>
<td>10.</td>
<td>o</td>
<td>pas</td>
<td>he</td>
</tr>
<tr>
<td>11.</td>
<td>na</td>
<td>à</td>
<td>it</td>
</tr>
<tr>
<td>12.</td>
<td>la</td>
<td>ne</td>
<td>little</td>
</tr>
<tr>
<td>13.</td>
<td>tɛ</td>
<td>prince</td>
<td>said</td>
</tr>
<tr>
<td>14.</td>
<td>ten</td>
<td>que</td>
<td>was</td>
</tr>
</tbody>
</table>
In the present work, we have outlined the online concordance tool with the parallel texts of *The Little Prince* in Bamana, French, and English. To our knowledge, this is the first project of this sort involving the Bamana language and has a large research potential. The concordance tool is presently freely available in the Internet.

One can work with the concordance in two modes: the parallel text output and the output of an interlinearized Bamana text. In the parallel text output, different search options are accessible, namely: entire wordform, words starting with/ ending in/ containing a given character sequence, and exact phrase search. The search options in the interlinearized output mode also include part of speech, gloss, and lemma, but not the exact phrase search.

Presently, the parallel texts in the corpus are aligned by paragraphs, and a simple algorithm to define the position of a corresponding sentence in the French or English text based on the position of the Bamana sentence is suggested. Automated sentence alignment based on the so-called anchor items (most frequent words) is briefly discussed for future applications. Further plans include extending search possibilities to all the language versions as well as adding more texts to the corpus.

In prospect, the parallel texts in the form of online concordances can be used as a supplementary tool for compiling Bamana–French, Bamana–English and French–Bamana, English–Bamana dictionaries. In particular, such a tool provides word translation equivalents in the context as well as illustration of a word usage. Also, the strategy and tactics of different translators can be discovered.
BIBLIOGRAPHY


NOTES
1. http://www.sfb632.uni-potsdam.de/hausa/
http://ktf.franko.lviv.ua/~andrij/vai-concord.html
8. This electronic version was kindly provided by the Saint-Petersburg group working on the Mande corpus (Valentin Vydrin, Artem Davydov, Anna Erman, Kirill Maslinskiy).

ABSTRACTS
The launch of the online concordance with the parallel texts of *The Little Prince* in Bamana, French, and English is reported. Two working modes are available: the parallel text output and the output of an interlinearized Bamana text. In the parallel text output, the following search options are accessible: entire wordform, words starting with/ ending in/ containing a given character sequence, and exact phrase search. The search options in the interlinearized output mode also include part of speech, gloss, and lemma, but not the exact phrase search.
The parallel texts in the corpus are aligned by paragraphs, and a simple algorithm to define the position of a corresponding sentence in the French or English text based on the position of the Bamana sentence is suggested. Automated sentence alignment based on the so-called anchor items (most frequent words) is briefly discussed for future applications.

L'article présente une concordance mise en ligne des textes parallèles bambara, français et anglais du Petit Prince. Deux régimes de travail sont disponibles : la sortie d'un texte parallèle et la sortie d'un texte bambara interlinéarisé. Dans la mode du texte parallèle, des possibilités de recherche accessibles sont les suivantes : le mot-forme entier ; des mots commençant / se terminant par/ contenant des séquences données des caractères ; des phrases entières. Les possibilités de recherche dans la mode interlinéarisée sont : partie de discours, glose, lemme, mais non pas une phrase entière.

Les textes parallèles du corpus sont alignés par paragraphes ; un simple algorithme est fourni pour définir la position de la phrase dans le texte français ou anglais en partant de la position de la phrase bambara. L'alignement automatique des phrases se basant sur les point d'ancrage (les mots les plus fréquents) est brièvement discuté ; cela peut être utile pour des applications ultérieures.

INDEX

Keywords: Bamana, Concordance, Parallel Corpus, Text Alignment, Anchor Item

topics bambara

motsclés Бамана, конкорданс, параллельный корпус, выравнивание текста, опорный элемент

Mots-clés: concordance, corpus parallèle, alignement de texte, point d’ancrage

AUTHORS

ANDRIJ ROVENCHAK
Lviv University

SOLOMIJA BUK
Lviv University