Mandenkan

Bulletin semestriel d'études linguistiques mandé

50 | 2013

Numéro 50

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

Electronic version

URL: <https://journals.openedition.org/mandenkan/268>

DOI: 10.4000/mandenkan.268

ISSN: 2104-371X

Publisher

Llacan UMR 8135 CNRS/Inalco

Printed version

Date of publication: 1 December 2013

Number of pages: 117-130

ISSN: 0752-5443

Electronic reference

Andrij Rovenchak and Solomija Buk, "*Masadennin (The Little Prince in Bamana)*", *Mandenkan* [Online], 50 | 2013, Online since 01 December 2013, connection on 08 July 2021. URL: <http://journals.openedition.org/mandenkan/268> ; DOI: <https://doi.org/10.4000/mandenkan.268>

This text was automatically generated on 8 July 2021.

Les contenus de *Mandenkan* sont mis à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International.

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 Cilubà⁴, 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.

- 3 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 Sɛn Tɛgiziperi 1989).
- 4 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).
- 5 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

- 6 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.
- 7 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-Inuktitut Parallel Corpus (Martin et al. 2003). Such approach can be useful in particular for texts with numerous occurrences of direct speech.
- 8 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.
- 9 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 1: A sample of aligned texts

Bamana	French	English
Bari n'ta fε mɔgɔw ka n ka kitabu manamana kalan. </s> <parbreak>	Car je n'aime pas qu'on lise mon livre à la légère. </s>	For I do not want any one to read my book carelessly. </s>
Nin ko kɔrɔw maana bɔli bε gan n dusu la cogo min.</s> A san wɔɔrɔ ye nin ye ka ban, kabini n terikε taara n'a ka saga ye.</s>	J'éprouve tant de chagrin à raconter ces souvenirs.</s> Il y a six ans déjà que mon ami s'en est allé avec son mouton.</s>	I have suffered too much grief in setting down these memories.</s> Six years have already passed since my friend went away from me, with his sheep.</s>
Ni n bε n jija k'a fɔ yan a tun bε cogo min, n t'a fε ka jine a kɔ de.</s> Mɔgɔ ka jine i tεri kɔ, dusukasiko don.</s> Bεε ma deli ka teri mara.</s>	Si j'essaie ici de le décrire, c'est afin de ne pas l'oublier.</s> C'est triste d'oublier un ami.</s> Tout le monde n'a pas eu un ami.</s>	If I try to describe him here, it is to make sure that I shall not forget him.</s> To forget a friend is sad.</s> Not every one has had a friend.</s>
O ni fana n bε se ka kε i ko mɔgɔ kɔrɔba minw mago tε fosi la tugun jatedenw kɔ.</s>	Et je puis devenir comme les grandes personnes qui ne s'intéressent plus qu'aux chiffres.</s>	And if I forget him, I may become like the grown-ups who are no longer interested in anything but figures...</s> <parbreak>
O fana de kosɔn n ye kiriyɔn jεma buwati kelen san, a ni kiriyɔn gansanw.</s> N si bε yɔrɔ min na, ka segin japegeɛn ma, a ka gɛlɛn, sanko n ma n yeɛ sifile fosi la ni minijan bosobali ni minijan bosolen tε, ka n si to sanji wɔɔrɔ la!</s>	C'est donc pour ça encore que j'ai acheté une boîte de couleurs et des crayons.</s> C'est dur de se remettre au dessin, à mon âge, quand on n'a jamais fait d'autres tentatives que celle d'un boa fermé et celle d'un boa ouvert, à l'âge de six ans !</s>	It is for that purpose, again, that I have bought a box of paints and some pencils.</s> It is hard to take up drawing again at my age, when I have never made any pictures except those of the boa constrictor from the outside and the boa constrictor from the inside, since I was six.</s>
N kɔni bε n jija ka jaw kε minw bɔlen bε u tigi fε kosebε.</s> Nka n dalen tε a la k'a fɔ n bεna se.</s> Ja dɔ ka fisa, tɔ kelen bɔlen tε a fε tugun.</s>	J'essaierai, bien sûr, de faire des portraits le plus ressemblants possible.</s> Mais je ne suis pas tout à fait certain de réussir.</s> Un dessin va, et l'autre ne ressemble plus.</s>	I shall certainly try to make my portraits as true to life as possible.</s> But I am not at all sure of success.</s> One drawing goes along all right, and another has no resemblance to its subject.</s>
N bε fili dɔɔnin fana kundama na.</s> Yan, masadennin ka jan kojugu.</s> Yan, a ka dɔgɔ kojugu.</s> N bε sigasiga fana a ka finiw jε na.</s>	Je me trompe un peu aussi sur la taille.</s> Ici le petit prince est trop grand.</s> Là il est trop petit.</s> J'hésite aussi sur la couleur de son costume.</s>	I make some errors, too, in the little prince's height: in one place he is too tall and in another too short.</s> And I feel some doubts about the color of his costume.</s>

O de kosɔn b bɛ mɔmɔli kɛ fan ni fan, jɛ ni jɛbaliya cɛ./s>	Alors je tâtonne comme ci et comme ça, tant bien que mal./s>	So I fumble along as best I can, now good, now bad, and I hope generally fair-to-middling. </s> <parbreak>
N na fili fana yɔɔ misɛnnin dɔw la minw jɛci ka bon kosebɛ./s> Nka a' kana jigi n na./s> N tɛrikɛ tun tɛ fosi jɛfɔ./s>	Je me tromperai enfin sur certains détails plus importants./s> Mais ça, il faudra me le pardonner./s> Mon ami ne donnait jamais d'explications./s>	In certain more important details I shall make mistakes, also./s> But that is something that will not be my fault./s> My friend never explained anything to me./s>
I b'a sɔɔ a jɛ na ne tun ye a jɔɔɔn dɔ ye./s> Nka ne kunna ma diya, n tɛ se ka sagaw ye wagande datugulenw kɔnɔ./s>	Il me croyait peut-être semblable à lui./s> Mais moi, malheureusement, je ne sais pas voir les moutons à travers les caisses./s>	He thought, perhaps, that I was like himself./s> But I, alas, do not know how to see sheep through the walls of boxes./s>
I b'a sɔɔ n ni mɔɔɔ kɔɔɔbaw ka surun jɔɔɔn na./s> N kɔɔɔlen na kɛ./s>	Je suis peut-être un peu comme les grandes personnes./s> J'ai dû vieillir./s>	Perhaps I am a little like the grown-ups./s> I have had to grow old./s>

- 10 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

Bamana	French	English
... N tun b'a fɛ k'a dɔn ni mɔɔɔ faamuyalen don./s> Nka a tigilamɔɔɔ tun bɛ n jaabi tuma bɛɛ: "Banfula don"./s> O la sa minijanko wo, tufinko wo, n tun t'o si kuma fɔ 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: <parbreak> "That is a hat."/s> <parbreak> Then I would never talk to that person about boa constrictors, or primeval forests, or stars./s> ...

- 11 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 3: Aligned excerpts with different text flow over paragraphs

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

- 12 Note that some English examples contain more sentences comparing to the two other language versions.
- 13 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).

```

<title>Masadennin</title>
<author>Antuwani de Sɛn TɛGIZIPERI</author>
<subtitle>(Bukari JARA y'a bayɛɛma)</subtitle>
<edition>EditionJamana</edition>
<dedication>
Lewɔn Wɛrɛti tɔɔɔ la
...
</dedication>
<c>I
Ka ne tɔ n sanji wɔɔɔɔnan na, don dɔ, n ye ja dakabana dɔ ye kitabu
dɔ kɔnɔ, min tun bɛ tu fin kun kan, n'a tɔɔɔ tun ye "Ko yelenw ni
kɔ kelenw.</s> Minipan dɔ tun don, a bɛ ka wara dɔ kunun. Ja ladege
file.</s>

```

2. Concordance implementation

- 14 The concordance can be accessed online at <http://www.ktf.franko.lviv.ua/~andrij/science/LP-bam.html>. The screenshot of the main page is shown in Figure 1.

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

- 15 Technical realization of the online concordance is partly based on some previously developed CGI-scripts in Perl, which were used for different purposes⁶. In future, texts in the corpus can be automatically converted for usage in some standard tool, e. g., the IMS Open Corpus Workbench⁷, as soon as a proper hosting is found for it.
- 16 One can work with the concordance in two modes: the parallel text output and the output of an interlinearized Bamana text.
- 17 **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.
- 18 Presently, only the Bamana text can be searched. Expansion for the two other languages can be implemented in future without much effort.
- 19 A simple type assistant is provided to facilitate (via the copy-paste procedure) the input of the non-ASCII characters, namely: ε / ε, џ / џ, ђ / ђ, џ / џ, Σ / Σ. The latter is a now-obsolete letter; in the novella text it appears only in the root џε (modern sɛ) 'hen' in the words џεw 'hen-PL' and џεɲinitɔ '≈ looking-for-hen'.

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

	A sunɔgɔra	-- Pas si petit que ça...	"Not so small that — Look!
1	Wa n ye masadennin dɔn nin cogo in de la .	Et c'est ainsi que je fis la connaissance du petit prince.	And that is how I made the acquaintance of the little prince.
2	Wa n ye masadennin dɔn nin cogo in de la .	Et c'est ainsi que je fis la connaissance du petit prince.	And that is how I made the acquaintance of the little prince.
3			
4	A meenna sanni n ka a boyɔɔɔ dɔn . Masadennin tun te ban ne pininka la , nka a tun be i ko ale te ne ka pininkaliw men . A tun be fili ka kuma minw fɔ olu de ye kow lajeya ne ye dɔnnin dɔnnin .	Il me fallut longtemps pour comprendre d'où il venait. Le petit prince, qui me posait beaucoup de questions, ne semblait jamais entendre les miennes. Ce sont des mots prononcés par hasard qui, peu à peu, m'ont tout révélé.	It took me a long time to learn where he came from. The little prince, who asked me so many questions, never seemed to hear the ones I asked him. It was from words dropped by chance that, little by little, everything was revealed to me.
5	-- Aa , nin ye yeleko ye Masadennin ye yeleba ce pi dɔ ke min ye n dusu bɔ kosebe . N b'a fɛ mɔgɔw ka n ka bone ke sebeko ye .	ça c'est drôle... Et le petit prince eut un très joli éclat de rire qui m'irrita beaucoup. Je désire que l'on prenne mes malheurs au sérieux.	"That is funny!" And the little prince broke into a lovely peal of laughter, which irritated me very much. I like my misfortunes to be taken seriously.
6	Ani bɔlɔ . A be i ko o kuma balala masadennin na :	-- Bien sûr. La proposition parut choquer le petit prince	"That is so. But the little prince seemed shocked by this offer.
7	A kuntilenna na fururu Masadennin y'i kan to , a sebe la :	-- N'importe où. Alors le petit prince remarqua gravement	"Anywhere. Then the little prince said, earnestly:

20 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 *mɔgɔ* and *mago*. To match the respective punctuation marks, an escape character '\' is required: '\.' for the period and '\?' for the question mark.

21 A sample output for the query "wordforms starting with **masa**" is shown in Figure 2.

22 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).

23 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.

24 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:

$$j_{FR} = j_{BM} / N_{BM} \times N_{FR} ; j_{EN} = j_{BM} / N_{BM} \times N_{EN}$$

25 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.

26 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

27 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 dictionary⁸ (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: *Wa n ye masadennin dɔn nin cogo in de la*. 'And that is how I made the acquaintance of the little prince'.

28 The tags read <PoS&gloss&lemma>, with variants separated by vertical stroke “|” and
compound structure shown by means of “=”.

29 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.

Fig. 3. A sample output of the interlinearized text

Email from Google ktf.franko.lviv.ua/cgi-bin... ktf.franko.lviv.ua/cgi-bin/LP-wordforms-i.cgi?wf=.toile&by=6&varmax=3

Concordance of the gloss **.toile**:

1	A	b'a	wele	i	n'a	f3	"	dolobisigi	3251	.	"		
	3SG	DUMMY	appeler	REFL	DUMMY	action.de.direrater	DQUOT	astéroïde	3251	FSTOP	DQUOT		
	ah!		appel	2.SG		dire		étoile					
						image							
	pers	undef	v	pers	undef	n	punct	n	num	punct	punct		
	intj		n	pers		v		n					
						v		n					
						n							
	à	b'a	wéle	i	n'a	f3	"	dolobisigi	3251	.	"		
	à		wéle	i		f3		dôlo					
						f3		bisigi					
2	Gansan	m'a	to	n	dalen	b'a	la	ko	masadenin	be	b3	tile	noʃedol
	inutilement	DUMMY	"to"	je	DUMMY	DUMMY	POSS	dire	petit prince	IPFV	bambou	soleil	DUMMY
	gratuit		donner.du."to"				ce.n'est.pas.vrai	laver	roi	IPFV.AFF	arracher		
			rester				dans	arriver	enfant		sortie		
	adv	undef	n	pers	undef	undef	conj	cop	n	pm	n	n	undef
	adj		v				intj	v	n	pm	v		
			v				pp	v	n		n		
	g3nsan	m'a	t3	n	dalen	b'a	l3	k3	masadenin	b3	b3	tile	noʃedol
	g3nsan		t3				l3	k3	m3sa	b3	b3		
			t3				l3	k3	d3n	d3n	b3		
3	O	dolobisigi	in	yera	sije	kelen	d3rn	y3r3jan	filelan	na	san	1909	na

- 30 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.
- 31 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.
- 32 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.

- 33 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.
- 34 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*.
- 35 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.

rank	Bamana		French		English	
1.	ka	870	le	455	the	977
2.	n	807	de	424	I	544
			<i>je, j', mon, moi, ...</i>	758	<i>I, me, my, ...</i>	883
3.	a	710	je	322	to	467
4.	ye	632	et	295	a	411
5.	bɛ	518	il	267	and	351
6.	tun	403	les	254	of	336
7.	ko	375	un	240	that	322
8.	i	324	la	221	you	310
9.	ni	267	petit	201	is	301
10.	o	267	pas	186	he	297
11.	na	230	à	174	it	264
12.	la	227	ne	171	little	261
13.	tɛ	222	prince	171	said	195
14.	ten	219	que	164	was	193
15.	kɛ	193	mais	138	prince['s]	187
16.	ma	188	c'est	136	in	176
17.	masadennin	180	tu	135	my	164
18.	min	177	des	131	not	160
19.	kan	176	dit	125	me	153
20.	don	171	une	122	but	150
21.	dɔ	167	qui	104	have	141
22.	fɛ	162	me	103	are	136
23.	nka	145	pour	100	for	135
24.	mɔgɔ	130	ce	98	one	132

25.	u	130	bien	96	his	119
-----	---	-----	------	----	-----	-----

4. Conclusions

- 36 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.
- 37 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.
- 38 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.
- 39 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

- ASKER Lars, ARGAW Atelach Alemu, GAMBÄCK Björn, EYASSU ASFEHA Samuel & NIGUSSIE HABTE Lemma, 2009, "Classifying Amharic webnews", *Information Retrieval* 12(3), pp. 416-435.
- BAILLEUL Charles, 2000, *Dictionnaire Bambara-Français*. 2e éd., Bamako, Donniya.
- BEHRENS Leila, 2005, "Genericity from a Cross-linguistic Perspective", *Linguistics* 43(2), pp. 275-344.
- BUK Solomija, ROVENCHAK Andriy, 2007, "Statistical parameters of Ivan Franko's novel *Perekhresni stežky* (The Cross-Paths)", in GRZYBEK, Peter and KÖHLER Reinhard (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday* (Quantitative Linguistics; 62), Berlin-New York, Mouton de Gruyter, pp. 39-48.
- DAVYDOV Artem, 2010, "Towards the Manding corpus: Texts selection principles and metatext markup", in: DE PAUW Guy, GROENEWALD H. J. & de SCHRYVER Gilles-Maurice (eds.), *Proceedings of the*

- Second Workshop on African Language Technology (AfLaT 2010)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 59-62.
- DAVYDOV Artem, 2011, "Electronic corpus of Bamana: acquisition and principles of metatextual mark-up", *Mandeica Petropolitana II (Acta Linguistica Petropolitana 7, Part 2)*, pp. 381-393.
- DE SEN TEGIZIPERI Antuwani, 1989, *Masadennin* (Bukari Jara y'a bayeɓɓe), Bamako, Edition Jamana.
- FATTAH Mohamed Abdel, BRACEWELL David B., REN Fuji & KUROIWA Shingo, 2007, "Sentence alignment using P-NNT and GMM", *Computer Speech and Language* 21, pp. 594-608.
- GALE, William A. & WARD CHURCH Kenneth, 1993, "A program for aligning sentences in bilingual corpora", *Computational Linguistics* 19(1), pp. 75-102.
- HURSKAINEN Arvi, 2004, "Swahili language manager: A storehouse for developing multiple computational applications", *Nordic Journal of African Studies* 13(3), pp. 363-397.
- KELIH, Emerich & GRZYBEK Peter, 2005, "Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispiel slowenischer Prosatexte)", *LDV-Forum* 20(2), pp. 31-51.
- MARTIN Joel, JOHNSON Howard, FARLEY Benoit & MACLACHLAN Anna, 2003, "Aligning and Using an English-Inuktitut Parallel Corpus", in *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts. Data Driven Machine Translation and Beyond*. Edmonton, May-June 2003, pp. 115-118.
- MCENERY Tony, WILSON Andrew, 2001, *Corpus linguistics: an introduction*. 2nd ed., Edinburgh University Press.
- MCENERY Tony, XIAO Richard, TONO Yukio, 2006, *Corpus-based language studies: an advanced resource book* (Routledge Applied Linguistics Series), Routledge.
- MELAMED I. Dan, 1999, "Bitext maps and alignment via pattern recognition", *Computational Linguistics* 25(1), pp. 107-130.
- ROVENCHAK Andrij, BUK Solomija, 2011, "Defining thermodynamic parameters for texts from word rank-frequency distributions", *Journal of Physical Studies* 15(1), 1005 (6 pages).
- RUDOLF Michał, 2004, *Metody automatycznej analizy korpusu tekstów polskich [Methods of the automatic analysis of the Polish text corpus]*, Warszawa Uniwersytet Warszawski.
- SCHRYVER Gilles-Maurice de, 2002, "Web for/as corpus: A perspective for the African languages", *Nordic Journal of African Studies* 11(2), pp. 266-282.
- SKATTUM Ingse, 2000, « Le bambara écrit à l'école fondamentale ». *Nordic Journal of African Studies* 9(3), pp. 108-132.
- STOLZ Thomas, 2007, "Harry Potter meets *Le petit prince*: On the usefulness of parallel corpora in crosslinguistic investigations", *Sprachtypologie und Universalienforschung* 60(2), pp. 100-117.
- TEUBERT Wolfgang, ČERMÁKOVÁ Anna, 2007, *Corpus linguistics: A short introduction*, London, Continuum.
- VYDRIN Valentin, 2008, "Glossed electronic corpora of Mande languages: A perspective that we cannot avoid", in *Mande languages and linguistics. 2nd International Conference, St. Petersburg (Russia), September 15-17, 2008. Abstracts and Papers*, pp. 15-22.
- VYDRIN Valentin, 2011, "Electronic glossed corpus of Bamana texts: first stage", *Mandeica Petropolitana II (Acta Linguistica Petropolitana 7, Part 2)*, pp. 343-380.

YANG Christopher C., WING LI Kar, 2004, "Building parallel corpora by automatic title alignment using length-based and text-based approaches", *Information Processing and Management* 40, pp. 939-955.

NOTES

1. <http://www.sfb632.uni-potsdam.de/hausa/>
<http://nlp.fi.muni.cz/projekty/deb2/emasters/www/tshwanelex/members/gmds/cv.html>
 2. http://www.world-lang.osaka-u.ac.jp/user/liccosec/eng/africa/language_02.html
 3. <http://www.ling.helsinki.fi/uhlcs/readme-all/README-afro-as-nig-con-lgs.html>
<http://nlp.fi.muni.cz/projekty/deb2/emasters/www/tshwanelex/members/gmds/cv.html>
 4. <http://nlp.fi.muni.cz/projekty/deb2/emasters/www/tshwanelex/members/gmds/cv.html>
 5. <http://cormand.tge-adonis.fr/>
 6. <http://ktf.franko.lviv.ua/~andrij/science/Franko/concordance.html>
<http://ktf.franko.lviv.ua/~andrij/vai-concord.html>
 7. <http://cwb.sourceforge.net/>
 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.

В СТАТЬЕ ГОВОРИТСЯ О ПУБЛИКАЦИИ В ИНТЕРНЕТЕ ОНЛАЙНОВОГО КОНКОРДАНСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ «МАЛЕНЬКОГО ПРИНЦА» НА БАМАНА, АНГЛИЙСКОМ И ФРАНЦУЗСКОМ ЯЗЫКАХ. ПРЕДСТАВЛЕНЫ ДВА РЕЖИМА РАБОТЫ: С ВЫВОДОМ В ВИДЕ ПАРАЛЛЕЛЬНОГО КОРПУСА И С ВЫВОДОМ В ВИДЕ СИНХРОНИЗИРОВАННОГО БАМАНСКОГО ТЕКСТА. В ФОРМАТЕ ПАРАЛЛЕЛЬНОГО КОРПУСА ДОСТУПНЫ СЛЕДУЮЩИЕ ПОИСКОВЫЕ ОПЦИИ: ПО ЦЕЛОЙ СЛОВОФОРМЕ, ПО ПОСЛЕДОВАТЕЛЬНОСТИ БУКВ В НАЧАЛЕ, КОНЦЕ ИЛИ СЕРЕДИНЕ СЛОВА, ПО ФРАЗЕ. ПОИСКОВЫЕ ОПЦИИ В ИНТЕРЛИНЕАРИЗИРОВАННОЙ МОДЕЛИ ТАКЖЕ ВКЛЮЧАЮТ ПОИСК ПО ЧАСТИ РЕЧИ, ПО ГЛОССЕ, ПО ЛЕММЕ, НО НЕ ПО ФРАЗЕ.

ПАРАЛЛЕЛЬНЫЕ ТЕКСТЫ В КОРПУСЕ ВЫРАВНЕНЫ ПО АБЗАЦАМ. ПРЕДЛАГАЕТСЯ ПРОСТОЙ АЛГОРИТМ ОПРЕДЕЛЕНИЯ ПОЗИЦИИ СООТВЕТСТВУЮЩИХ ПРЕДЛОЖЕНИЙ ФРАНЦУЗСКОГО И АНГЛИЙСКОГО ТЕКСТОВ. ИСХОДЯ ИЗ ПОЗИЦИИ В ТЕКСТЕ БАМАНСКОГО ПРЕДЛОЖЕНИЯ. КРАТКО ОБСУЖДАЕТСЯ ВОЗМОЖНОСТЬ ПРИМЕНЕНИЯ В ПОСЛЕДУЮЩИХ ВЕРСИЯХ АВТОМАТИЧЕСКОГО ВЫРАВНИВАНИЯ ПРЕДЛОЖЕНИЙ НА ОСНОВЕ ТАК НАЗЫВАЕМЫХ ОПОРНЫХ ЭЛЕМЕНТОВ (НАИБОЛЕЕ ЧАСТОТНЫХ СЛОВ).

INDEX

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

Subjects: bambara

motsclesru 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