Mandenkan, No. 64, 2021, pp. 81-104

Bamana tales recorded by Umaru Ɲanankɔrɔ Jara: A comparative study based on a Bamana–French parallel corpus

Andrij Rovenchak
Ivan Franko National University of Lviv
andrij.rovenchak@gmail.com

1. Introduction

Parallel text corpora have been firmly becoming an essential tool for natural language processing and linguistic studies in the domain of contrastive analysis, translation studies and lexicology (Borin 2002; Hansen-Schirra, Neumann & Čulo 2017; Doval & Sánchez Nieto 2019) Such corpora range from small-sized single-author or even single-text collections (Buk 2012; Sitchinava 2016) to large scale ones, like EUR-Lex Corpus[1] based on the European Union legislation and other documents (Baisa et al. 2016). Resources for African languages remain under-represented and mostly focused on Swahili (De Pauw, Wagacha & de Schryver 2011; Wójtowicz 2018), Amharic (Rychlý & Suchomel 2016; Woldeyohannis, Besacier & Meshesha 2018) or languages of South Africa (Wallmach 2000; Moropa 2007). The parallel Bamana–French corpus, which is a part of a larger project, the Bamana Reference Corpus (BRC, see Vydrin 2013; Vydrin et al. 2011–2019) is the only example of the Mande languages.

In the present work, six Bamana tales recorded by Umaru Ɲanankɔrɔ Jara (Oumar Nianankoro Diarra) are studied. Analysis is made using texts from the abovementioned Bamana–French parallel corpus. Distributions of parts of speech are obtained for both Bamana originals and French translations.

The following texts have been analyzed:

1. “Dununba kumata” (“Le tam-tam qui parle” = “The talking tom-tom”)
2. “Juguya sara” (“Le prix de la méchanceté” = “The price of wickedness”)
3. “Juman nɔrɔla farakolo la” (“Diouman s’est collée à une pierre” = “Diouman stuck to a stone”)
4. “Ntalen” (“Ntalen” [Parabole : araignée = Parable: spider])

[1] https://www.sketchengine.eu/eurlex-corpus/

5. “Sigidankelen ka labanko juguya” (“La fin tragique de Sigidankelen” = “The tragic end of Sigidankelen”)
6. “Warabilenkɔrɔ ka walijuya” (“La sainteté du vieux singe rouge” = “The holiness of the old red monkey”)

The first two texts were published in a children’s book entitled *Dununba kumata : Mali nsiirinw* (Diarra & Fenayon 2011a; Diarra & Fenayon 2011b), see Figure 1. Apart from this Bamana version, a French translation of the book (*Le tam-tam qui parle : contes du Mali*, translated by Umaru Ŋanankɔrɔ Jara and Antoine Fenayon) as well as a German one (*Die sprechende Trommel: Geschichten aus Mali*, translated by Tim Hentschel) also appeared in 2011. Four other tales were provided by Umaru Jara himself as handwritten notebooks.

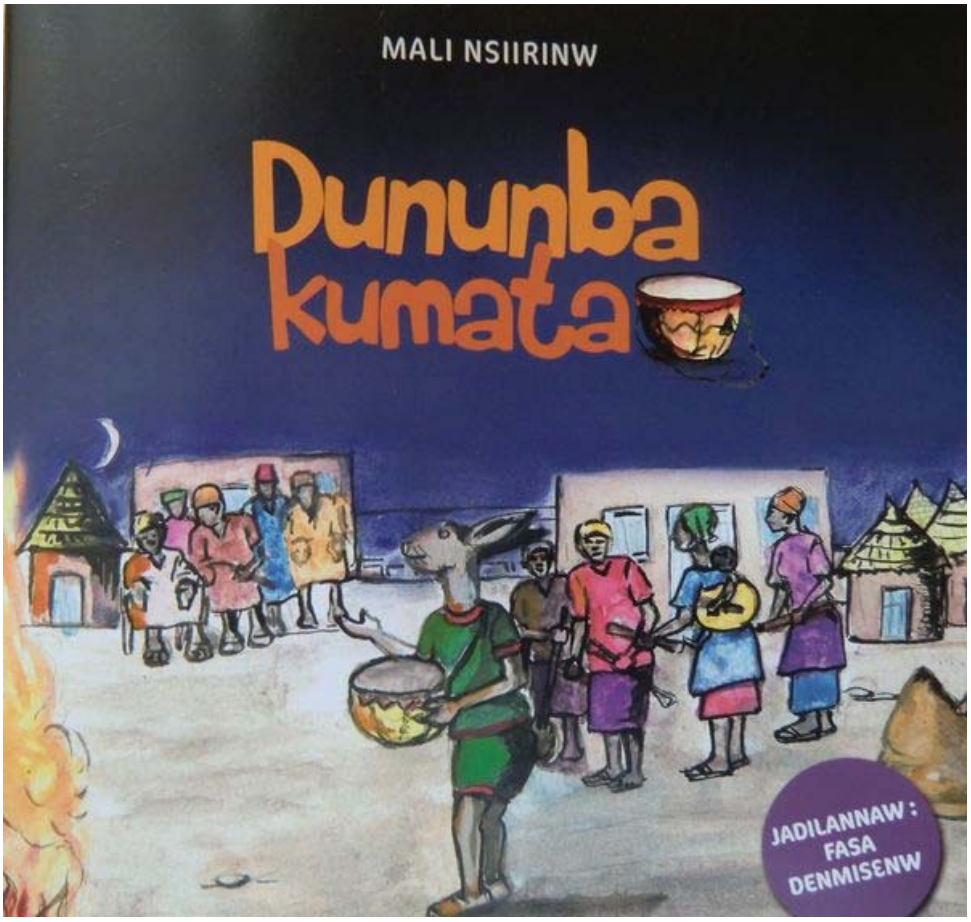

Figure 1. Book cover of *Dununba kumata : Mali nsiirinw* (Diarra & Fenayon 2011a). Image source: http://donniyakadi.over-blog.com.

The rest of the paper is organized as follows: Section 2 discusses details of autosemantic parts of speech (PoS) as well as PoS-tagging and lemmatization issues; Section 3 contains results about frequency data in the analyzed Bamana and French texts; Section 4 discusses some peculiarities of adjective functioning; Section 5 briefly describes an application of the theory of complex networks. Finally, conclusions are drawn in Section 6.

2. Autosemantic parts of speech and lemmatization

The paper analyzes the distribution of autosemantic parts of speech in the texts of the tales. The term 'autosemantic' refers to meaningful parts of speech (also known as content words), like nouns, verbs, adjectives, adverbs, etc. (Popescu, Altmann & Köhler 2010). These are contrasted with synsemantic (auxiliary) PoS (also known as function words), like particles, conjunctions, prepositions, etc. As there are no strict approaches to defining a particular PoS across languages, especially when dealing with languages of different families, it is worth discussing briefly which parts of speech are considered autosemantic in this work for the two languages, Bamana and French.

The problem of parts of speech in Bamana has been discussed in a number of works (see especially Vydrine 1999 and references therein). Using different approaches, the authors mostly agree on the core set of nominals, verbs, and adjectives (Creissels 1983; Kastenholz 1998; Dumestre 2003), even though their definitions and the respective PoS-attributions do not necessarily coincide. Bamana is also sometimes described as a language with flexible word classes (Rijkhoff & van Lier 2013). In the present work, I mostly adhere to the definitions of the parts of speech based on morphosyntactic criteria as described by Vydrin (2017a) and applied in the Bamana Reference Corpus.

For Bamana, texts from the tagged and disambiguated part of the BRC are used. The tools for building this and related corpora are described in detail by Maslinsky (2014). With PoS tags at hand, the following PoS are considered autosemantic: adjective, adverb (including preverbial), copula, determinative, noun, numeral, participle, pronoun (personal and non-personal), qualitative verb, and verb. Copulas are treated as autosemantic due to their syntactic role close to that of verbs. A similar syntactic criterion is applied to determinatives behaving like adverbs and to pronouns, which can substitute nouns (or adjectives in certain contexts).

French texts were lemmatized using the TreeTagger software (Schmid n. d.) yielding a set of tags (Stein 2003) corresponding to the following PoS treated as autosemantic: adjective, adverb, noun (including a separate NAM tag for proper names), numeral, pronoun (personal, possessive, etc.), and verb.

The autosemantic parts of speech in both languages and the respective tags are summarized in Table 1 for convenience.

Lemmatization in the Bamana texts was performed by cutting affixes corresponding to flexion[2], namely: verbal progressive suffix *-la/-na* (glossed as PROG), non-productive plural marker *-lu/-nu* (PL2), perfective intransitive marker

[2] http://cormand.huma-num.fr/gloses.html

-ra/-la/-na (PFV.INTR), optative marker *-ra/-la/-na* (OPT2), and plural marker *-w* (PL). Some examples are as follows:

sèginna ‘revenir.PROG’ is lemmatized as *sègin* ‘revenir’;

mínnu ‘REL.PL2’ is lemmatized as *mîn* ‘REL’;

táara ‘aller.PFV.INTR’ is lemmatized as *táa* ‘aller’;

mɔ̀gɔw ‘homme.PL’ is lemmatized as *mɔ̀gɔ* ‘homme’.

No optative morphemes have been attested in the analyzed texts.

Here and below, glosses are given in French as they appear in the BRC. This facilitates comparisons with the French translations in the parallel texts. The free French translations taken from the French part of the parallel corpus are followed by ther English equivalents.

Table 1. Autosemantic parts of speech and respective tags

Part of speech	Bamadaba tags	French TreeTagger tags
noun	n	NOM, NAM
verb	v, ptcp	VER
qualitative verb	vq	—
copula	cop	—
adjective	adj	ADJ
determinative	dtm	—
adverb	adv	ADV
numeral	num	NUM
pronoun	prn, pers	PRO (including PRO:DEM, PRO:IND, PRO:PER, etc.), DET:POS

Bamana lemmas obtained from the corpus underwent some normalization. First of all, contracted forms resulting from the vowel elision were lemmatized as full ones, e. g., copulas *y’* ‘être’ as *yé* ‘être’, *t’* ‘COP.NEG’ as *té* ‘COP.NEG’, *d’* ‘donner’ as *dí* ‘donner’, *f’* ‘dire’ as *fɔ́* ‘dire’, *k’* ‘faire’ as *kɛ́* ‘faire’, predicative markers *k’* ‘INF’ as *kà*: ‘INF’, *m’* ‘PFV.NEG’ as *ma* ‘PFV.NEG’, etc. Next, dialectal forms were replaced with primary ones according to the Bamadaba dictionary (Bailleul et al. 2011–2020), e. g., *búbagatoo* ‘termitière’ → *búbaganton* ‘termitière’, *dímin* ‘faire.souffrir’ → *dími* ‘faire.souffrir’, *tága* ‘aller’ → *táa* ‘aller’, *tágama* ‘voyage’ → *táama* ‘voyage’, etc. Finally, a few typos, mostly resulting from incorrect accent placement, were corrected.

The lemmatized French texts were manually post-processed to remove ambiguities and make some corrections. In particular, the two most frequent ambiguous lemmatizations were (variants are separated by a vertical line “|”)

suis	VER	suivre|être
fils	NOM	fil|fils

In both cases, only the second variant was found in the analyzed texts.

The most frequent incorrect lemmatization was

nouvelle(s)	ADJ	nouveau

instead of

nouvelle(s)	NOM	nouvelle

There is also another problem, which cannot be solved automatically, namely, the lemmatization of the French ‘un/une’. It is not always clear whether such a word should be considered an indefinite article or a numeral. This problem is known to occur in the tagging of texts in Romance languages. Sometimes, a portmanteau tag is used, e.g., \ARTi:NUMc in the Portuguese corpus (Bacelar do Nascimento et al. 2005). In the collocation ‘une fois’, which is very frequent in the text of tales, the tag corresponding to an indefinite article is used for ‘une’ in an example quoted by Salamanca (2019). In the present work, only those instances of ‘un/une’ are tagged as numerals where the cardinality is clear, for instance, ‘*Une* des femmes…’ or ‘*Un* mois passe**,** deux mois, trois mois, quatre mois, cinq mois…’.

To facilitate comparisons with the Bamana texts, all French personal pronoun lemmas were replaced with a person-number gloss, e.g.:

je	PRO:PER	1SG
me	PRO:PER	1SG
moi	PRO:PER	1SG

or

ils	PRO:PER	3PL
eux	PRO:PER	3PL
elles	PRO:PER	3PL

Table 2 shows statistics on parts of speech in the analyzed texts. To clarify the terms used below, consider the following examples. The number of tokens is the total number of running words, while the number of types is the number of different words (lemmas) in a given text. For instance, the sentence

(1a) ***À*** ***yɛ́lɛla*** *kà* ***yɛ́lɛ*** , *fɔ́* *kà* ***à*** ***ɲɛ́ji*** ***bɔ́*** .
pers **v** pm **v** conj pm **pers** **n** **v**
3SG **rire.PFV.INTR** INF **rire** jusqu'à INF **3SG** **larme** **sortir**
‘Il rit beaucoup, il rit tellement qu’il en pleura.’ = ‘He laughed a lot, he laughed so much that he cried.’ [“Juguya…”]

contains nine tokens, of which six are autosemantic PoS (given in boldface). The number of lemma types is six (*à, yɛ́lɛ*, *kà*, *fɔ́*, *ɲɛ́ji*, *bɔ́*), of those four are autosemantic

(*à, yélɛ*, *ɲéji*, *bɔ́*). Note that the verb *yélɛla* ‘rire.PFV.INTR’ was lemmatized as *yélɛ* ‘rire’.

Another example

(1b) ***“Jalakɔro”*** ***yé*** ***dùgu*** *yé* ***dùgu*** ***bèlebele*** .
n.prop **cop** **n** pp **n** **adj**
TOP **EQU** **terre** PP **terre** **gros**
‘Dialakoro est un village, un bien gros village.’ = ‘Dialakoro is a village, a pretty big village.’ [“Juman…”]

contains six tokens (including five autosemantic) and five types, of which four are autosemantic. Note that the two occurrences of *yé* are counted as different instances (copula and postposition).

Table 2. Data about the number of tokens and types in the Bamana and French versions of tales

	Dununba…		Juguya…		Juman…		Ntalen		Sigidan-kelen…		Warabi-lenkɔrɔ…		Whole collection	
	bam	fra	bam	fra	bam	fra	bam	fra	bam	fra	bam	fra	bam	fra
All lemma tokens	1491	1633	656	769	723	968	1723	2007	984	1202	1643	2017	7218	8596
Autosemantic lemma tokens	1055	1145	491	580	524	665	1264	1486	698	863	1180	1460	5212	6199
All lemma types	396	461	239	266	283	296	461	472	361	383	454	484	1178	1322
Autosemantic lemma types	335	426	195	233	235	268	396	437	307	351	386	448	1079	1270

From Table 2 one can see that autosemantic parts of speech account for 69–75% of all words used in a text while they are constitute 82–93% of the vocabulary (list of types). Details about their distribution are presented in the next section.

3. Frequency results for autosemantic PoS

Frequency lists of lemmas in the Bamana and French texts were compiled and are available from the author upon request. The numbers and percentages of autosemantic parts of speech in both text and vocabulary are shown in Tables 3–6.

Table 3. Distributions of autosemantic parts of speech in Bamana tales (types)

	Dununba…		Juguya…		Juman…		Ntalen		Sigidan-kelen…		Warabi-lenkɔrɔ…		Whole collection	
verb	110	34,0%	57	29,8%	60	26,8%	106	28,1%	93	32,0%	102	28,2%	304	28,2%
noun	157	48,5%	87	45,5%	116	51,8%	205	54,4%	150	51,5%	200	55,2%	658	61,0%
adj	7	2,2%	5	2,6%	10	4,5%	14	3,7%	7	2,4%	13	3,6%	32	3,0%
adv	17	5,2%	10	5,2%	11	4,9%	12	3,2%	10	3,4%	11	3,0%	37	3,4%
prn	17	5,2%	14	7,3%	14	6,3%	19	5,0%	16	5,5%	19	5,2%	23	2,1%
num	2	0,6%	5	2,6%	2	0,9%	7	1,9%	1	0,3%	4	1,1%	8	0,7%
dtm	8	2,5%	8	4,2%	6	2,7%	9	2,4%	9	3,1%	8	2,2%	11	1,0%
cop	6	1,9%	5	2,6%	5	2,2%	5	1,3%	5	1,7%	5	1,4%	6	0,6%
	324	**100%**	**191**	**100%**	**224**	**100%**	**377**	**100%**	**291**	**100%**	**362**	**100%**	**1079**	**100%**

On average, nouns account for around 50% of the vocabulary and verbs account for about another 30%. Note that the proportion of nouns increases to over 60% when the whole collection is analyzed. The reason is quite clear: in every new text, new objects and concepts are more likely to appear than words belonging to other parts of speech.

The category of verbs includes participles as well as qualitative verbs. The latter are not numerous: only 24 occurrences together in all six texts (seven in "Dununba…", two in "Juguya…", two in "Juman…", seven in "Ntalen", three in "Sigidankelen…", and three "Warabilenkɔrɔ…"). Of those, *dí* '[être] agréable' is found in all but one text (five occurrences), *ɲì* 'bon' is found in three texts (seven occurrences), and *kán* 'égal' is found in two texts (four occurrences). Depending on the approach used, qualitative verbs can be counted together with adjectives, so the above data allow for simple recalculations if required.

Table 4. Distributions of autosemantic parts of speech in Bamana tales (tokens)

	Dununba…		Juguya…		Juman…		Ntalen		Sigidan-kelen…		Warabi-lenkɔrɔ…		Whole collection	
verb	271	25,7%	108	22,0%	115	21,9%	276	21,8%	163	23,3%	276	23,4%	1208	23,2%
noun	383	36,3%	132	26,9%	218	41,6%	510	40,3%	252	36,0%	440	37,3%	1935	37,1%
adj	12	1,1%	6	1,2%	13	2,5%	17	1,3%	11	1,6%	15	1,3%	74	1,4%
adv	28	2,7%	16	3,3%	13	2,5%	21	1,7%	20	2,9%	18	1,5%	116	2,2%
prn	255	24,2%	160	32,6%	102	19,5%	283	22,4%	169	24,1%	299	25,4%	1268	24,3%
num	9	0,9%	9	1,8%	7	1,3%	28	2,2%	3	0,4%	10	0,8%	66	1,3%
dtm.	45	4,3%	20	4,1%	23	4,4%	62	4,9%	46	6,6%	55	4,7%	251	4,8%
cop	52	4,9%	40	8,1%	33	6,3%	67	5,3%	36	5,1%	66	5,6%	294	5,6%
	1055	**100%**	**491**	**100%**	**524**	**100%**	**1264**	**100%**	**700**	**100%**	**1179**	**100%**	**5212**	**100%**

When texts are analyzed, one counts tokens. Their proportion differs from the vocabulary (with types counted). On average, a half of a text is nearly equally split between verbs and pronouns, so words belonging to the respective parts of speech account for about 25% of autosemantic PoS in a text. Approximately another 40% of autosemantic words are nouns, see Table 4.

French translations demonstrate somewhat different proportions, both in the vocabulary and in the text (see Tables 5 and 6). Nouns constitute about 40% of the vocabulary, followed by verbs with nearly 30%. Adjectives and adverbs are represented in the vocabulary of individual texts in almost equal parts, about 10% each. In the French texts of the tales, nouns constitute about 30%. They are followed by verbs and pronouns, with the proportions similar to Bamana's (about 25% each).

The most pronounced difference between Bamana and French is the relative frequencies of adjectives and adverbs, especially in texts. With an average share of about 11% (of both vocabulary and text), adverbs in French are five times more frequent compared to the Bamana text (2.4%) and almost three times more frequent compared to the Bamana vocabulary (4.2%). Adjectives in French appear also nearly five times more frequently in texts (6.9% versus 1.5%) and more than three times in the vocabulary (10.4% versus 3.2%). Note that the respective numbers are the values averaged over individual texts, not those given in the last columns of Tables 3–6 corresponding to the whole collection of six tales.

Table 5. Distributions of autosemantic parts of speech in the French versions of the tales (types)

	Dununba…		Juguya…		Juman…		Ntalen		Sigidan-kelen…		Warabi-lenkɔrɔ…		Whole collection	
verb	140	32,9%	71	30,5%	73	27,2%	122	28,0%	112	31,9%	131	29,2%	372	29,3%
noun	167	39,2%	77	33,0%	104	38,8%	197	45,2%	140	39,9%	197	44,0%	603	47,5%
adj	45	10,6%	27	11,6%	34	12,7%	41	9,4%	29	8,3%	44	9,8%	151	11,9%
adv	45	10,6%	32	13,7%	35	13,1%	41	9,4%	41	11,7%	40	8,9%	96	7,6%
pron	26	6,1%	21	9,0%	21	7,8%	27	6,2%	29	8,3%	32	7,1%	39	3,1%
num	3	0,7%	5	2,1%	1	0,4%	8	1,8%	0	0,0%	4	0,9%	9	0,7%
	426	**100%**	**233**	**100%**	**268**	**100%**	**436**	**100%**	**351**	**100%**	**448**	**100%**	**1270**	**100%**

Table 6. Distributions of autosemantic parts of speech in French versions of the tales (tokens)

	Dununba…		Juguya…		Juman…		Ntalen		Sigidan-kelen…		Warabi-lenkɔrɔ…		Whole collection	
verb	309	27,0%	157	27,1%	163	24,5%	378	25,5%	230	26,7%	370	25,3%	1607	25,9%
noun	364	31,8%	122	21,0%	220	33,1%	497	33,5%	247	28,6%	450	30,8%	1900	30,7%
adj	70	6,1%	42	7,2%	61	9,2%	72	4,8%	58	6,7%	107	7,3%	410	6,6%
adv	125	10,9%	80	13,8%	72	10,8%	148	10,0%	111	12,9%	143	9,8%	679	11,0%
pron	273	23,8%	172	29,7%	148	22,3%	369	24,8%	217	25,1%	379	26,0%	1559	25,1%
num	4	0,3%	7	1,2%	1	0,2%	21	1,4%	0	0,0%	11	0,8%	44	0,7%
	1145	**100%**	**580**	**100%**	**665**	**100%**	**1485**	**100%**	**863**	**100%**	**1460**	**100%**	**6199**	**100%**

With the frequency data obtained for each text, it is easy to compile a frequency dictionary of the entire text collection. Table 7 contains a complete list of lemmas corresponding to autosemantic parts of speech common to all six tales. There are 46 lemmas in Bamana and also 46 in French. In accordance with the observations made above, the Bamana part contains only one adverb (*bì* ‘aujourd’hui’). While the lack of adverbs and adjectives is not unexpected (cf. Creissels 2003; Segerer 2008)[3] and can be partly compensated for by some other parts of speech, like determinatives or qualitative verbs, the absence of equivalents for French *grand* ‘big’ and *petit* ‘small’ in the Bamana list catches the eye immediately. Some reasons for this are discussed in the next Section.

The proportion of nouns is much higher in the Bamana list of common words (12) versus the French one (4). There are a number of reasons for such a relation. For instance, *dùgu* glossed as ‘terre’ can also denote ‘village’, which is reflected in the French side. The occurrences of ‘village’ in the French texts are also due to compound words, such as *dùgutigi* ‘chef du village’. On the other hand, *síra* ‘chemin’ mostly appears in the French texts not as a physical path but rather as more abstract concepts, such as ‘relation’ or ‘link’. A more detailed analysis can be made for all the instances, which is beyond the intended scope of the present study.

[3] Obviously, the approach to the definitions of parts of speech plays an important role here, as discussed in Section 2. In Bamana, similarly to some other Mande languages, adjectives and adverbs are rather heterogeneous word classes (Creissels 2009; Creissels & Sambou 2013; Dumestre 2011; Tröbs 2008; Tröbs 2014; Vydrin 2017a; Vydrin 2017b). The fractions in the texts of the disambiguated part of the Bamana Reference Corpus are, for instance, about 9 adjectives and 7 adverbs per 100 verbs. As for the list of types, one can refer to the latest version of the Bamadaba dictionary (Bailleul et al. 2011–2020), with about 24 adjectives and 16 adverbs per 100 verbs.

Table 7. Words common to all six texts in Bamana and French

	Bamana						French			
Rank	Lemma	PoS	Gloss	Freq	Cover	Rank	Lemma	PoS*	Freq	Cover
1	à	*pers*	3SG	442	8,5%	1	3SG	*pers*	507	8,2%
2	ò	*prn*	ce	205	12,4%	2	être	*v*	229	11,9%
3	kɛ́	*v*	faire	124	14,8%	3	ne	*adv*	166	14,6%
4	í	*pers*	2SG	112	16,9%	4	avoir	*v*	139	16,8%
5	ù	*pers*	3PL	107	19,0%	5	ce	*prn*	136	19,0%
6	nê	*pers*	1SG.EMPH	91	20,7%	6	son	*poss*	127	21,0%
7	kó	*cop*	QUOT	88	22,4%	7	pas	*adv*	113	22,9%
8	bɛ́	*cop*	être	87	24,1%	8	1SG	*pers*	96	24,4%
9	mîn	*dtm*	REL	65	25,3%	9	2SG	*pers*	82	25,7%
10	yé	*cop*	EQU	55	26,4%	10	dire	*v*	74	26,9%
11	í	*pers*	REFL	51	27,4%	11	3PL	*pers*	72	28,1%
12	fɔ́	*v*	dire	48	28,3%	12	tout	*prn*	62	29,1%
13	sé	*v*	arriver	48	29,2%	13	que	*prn*	56	30,0%
14	tùma	*n*	moment	48	30,1%	14	faire	*v*	55	30,9%
15	dɔ́	*dtm*	certain	44	31,0%	15	qui	*prn*	49	31,7%
16	dùgu	*n*	terre	44	31,8%	16	mon	*poss*	48	32,4%
17	mɔ̀gɔ	*n*	homme	44	32,7%	17	village	*n*	43	33,1%
18	ìn	*dtm*	DEF	43	33,5%	18	aller	*v*	41	33,8%
19	mîn	*prn*	REL	43	34,3%	19	jour	*n*	41	34,5%
20	sɔ̀rɔ	*v*	obtenir	42	35,1%	20	cela	*prn*	38	35,1%
21	táa	*v*	aller	42	35,9%	21	1PL	*pers*	36	35,7%
22	ń	*pers*	1SG	42	36,7%	22	prendre	*v*	34	36,2%
23	bɔ́	*v*	sortir	39	37,5%	23	grand	*adj*	32	36,7%
24	bɛ́ɛ	*dtm*	tout	39	38,2%	24	petit	*adj*	31	37,2%
25	ê	*pers*	2SG.EMPH	37	38,9%	25	ton	*poss*	31	37,7%
26	tɛ́	*cop*	COP.NEG	35	39,6%	26	arriver	*v*	29	38,2%
27	kélen	*num*	un	35	40,3%	27	en	*prn*	29	38,7%
28	yé	*v*	voir	34	40,9%	28	autre	*adj*	24	39,0%
29	tó	*v*	rester	33	41,6%	29	leur	*poss*	22	39,4%
30	dón	*n*	jour	32	42,2%	30	alors	*adv*	21	39,7%
31	kó	*n*	affaire	31	42,8%	31	aujourd'hui	*adv*	21	40,1%
32	nà	*v*	venir	29	43,3%	32	mettre	*v*	19	40,4%
33	dòn	*cop*	ID	28	43,9%	33	bien	*adv*	18	40,7%
34	yɔ́rɔ	*n*	lieu	27	44,4%	34	2PL	*pers*	17	40,9%
35	bìla	*v*	mettre	26	44,9%	35	celui	*prn*	17	41,2%
36	sú	*n*	nuit	25	45,4%	36	tout	*adv*	17	41,5%
37	kánto	*v*	s'adresser	22	45,8%	37	lever	*v*	16	41,7%
38	dɔ́n	*v*	connaître	20	46,2%	38	savoir	*v*	16	42,0%
39	dòn	*v*	entrer	20	46,6%	39	comme	*adv*	14	42,2%
40	bì	*adv*	aujourd'hui	19	46,9%	40	œil	*n*	13	42,4%
41	sí	*dtm*	aucun	13	47,2%	41	rester	*v*	13	42,7%
42	tɔ́gɔ	*n*	nom	12	47,4%	42	gens	*n*	12	42,8%
43	síra	*n*	chemin	11	47,6%	43	là	*adv*	11	43,0%
44	tɔ̀	*n*	le.reste	10	47,8%	44	devoir	*v*	10	43,2%
45	tìle	*n*	soleil	10	48,0%	45	voilà	*adv*	10	43,3%
46	fàn	*n*	côté	7	48,1%	46	beaucoup	*adv*	7	43,5%

* PoS tags for French are given according to the abbreviations accepted in the Bamana corpus, with an additional notation *poss* for possessive pronouns.

There is only one numeral, *kélen* ‘un’, in the Bamana list of common words but no numerals at all in the French one. The reason, at least partly, might be sought in the issue of the article/numeral ambiguity in tagging the French counterpart ‘un/une’ discussed in Section 2.

The “Cover” column in Table 7 shows the proportion of text covered by the respective lemmas relative to the total number of autosemantic tokens in all the texts (5212). So, the 46 lemmas common to all six texts in Bamana account for 48.1% of all words. Seventy lemmas are common to at least five texts and cover 54.7% of tokens. The lemmas common to at least three texts count 167 and cover already 66.4% of tokens. The numbers for the French versions are slightly different. There are 6199 autosemantic tokens in all six French texts, of which 46 are common to all six texts; they cover 43.5% of text. Eighty lemmas are common to at least five texts and cover 51.7% of text, while 219 lemmas are common to at least three texts and cover 66.9% of text.

4. Lack of size adjectives in the list of most frequent words in Bamana

Four reasons can be identified leading to a significantly smaller number of size adjectives in Bamana texts compared to their French translations. They are listed in subsections 4.1–4.4.

4.1. First of all, in Bamama diminutive suffix *-nin* and augmentative suffix *-ba* are used extensively in instances where one would expect ‘petit’ or ‘grand / gros’ in French. Several examples are shown in (2a–c).

(2a)	*Tòro*	*bɔ́ra*	*à*	*ka*	***wònin***	*fɛ̀*
	rat.voleur	sortir.PFV.INTR	3SG	POSS	**trou.DIM**	par

The respective French sentence reads ‘Toro sortit par son **petit trou**.’ = ‘Toro came out through his **little hole**.’ [“Dununba…”].

(2b)	*– **Cɛ̀nin***	*wó,*	*í*	*tɛ*	*sábali !*
	jeune.homme = **mâle.DIM**	hé	2SG	IPFV.NEG	être.patient

‘– **Petit garçon**, tu n’exagères pas ?’ = ‘– **Little boy**, aren't you exaggerating?’ [“Dununba…”].

(2c)	*Dúnuya*	***kó-ba***	*cáman*	*sún*	*bɛ́*	*mɔ̀ɔsow*	*lá :…*
	monde	**affaire-AUGM**	nombreux	tronc	être	homme.maison	à

‘L’origine de bien des **grandes œuvres** de la vie, c’est les femmes : …’ = ‘The origin of many **great works** of life is women: …’ [“Sigidankelen…”].

4.2. Single Bamana words can be translated into French by lexical equivalents containing two words, one of which is a size adjective (3a–c).

(3a)	*– Ń*	***kɔ̀rɔ***	*Kélennako*	*nê*	*séra*	*í*	*fɛ̀*
	1SG	**aîné**	NOM.M	1SG.EMPH	arriver.PFV.INTR	2SG	par

yàn bì
ici aujourd'hui
'– [Mon] **Grand frère** Kélénako, me voilà aujourd'hui devant toi.' = '[My] **Big brother** Kélénako, here I am today in front of you.' ["Juguya…"].

(3b) *Nê* ***dɔ́gɔmuso*** *nàna* *kó* *ń* *ka*
1SG.EMPH **cadette** venir.PFV.INTR QUOT 1SG SBJV

ɲɔ *dí* *à* *mà*
mil donner 3SG ADR
'Ma **petite sœur** est venue me demander du mil.' = 'My **little sister** came to ask me for millet.' ["Juguya…"].

(3c) *Ála* *y'* *à* *ládiya,* *kà* *nàfolo* ***cáman***
Dieu PFV.TR 3SG récompenser INF biens **nombreux**

dá *à* *yé :* *bà,* *mìsi,* *sàga,* *fàli.*
poser 3SG PP chèvre bovidé ovin âne
'Dieu avait fait de lui un homme riche : il possédait en **grand nombre** des ânes, des vaches, des moutons et des chèvres.' = 'God had made him a rich man: he owned donkeys, cows, sheep and goats in **large numbers**.' ["Juguya…"].

4.3. A descriptive synonymic translation can be used rather than a direct equivalent (4a–b).

(4a) *–* ***Wò*** *bɛ́* *dùnun* *lá, ...*
trou être tambour à
'– Il y a une **petite ouverture** au bas du tam-tam.' = '– There is a **small opening** at the bottom of the drum.' ["Dununba…"]. Note that in example (2a) *wònin* 'trou.DIM' was utilized.

(4b) *... ò* ***kámalen*** *y'* *í* *kánto : ...*
ce **jeune.homme** PFV.TR REFL s'adresser
'… son **petit ami** lui déclare : …' = '… her **boyfriend** tells her: …' ["Juman…"]

A few sentences further on, a diminutive is used instead:

(4c) *– Ɛ̀ɛ,* *ń* ***térinin****, ...*
pas.possible! 1SG **ami.DIM**
'– Eh ! Mon **petit ami**, …' = '– Hey ! My **little friend**, …' ["Juman…"]

4.4. Free translations can be too loose or contain idioms. These range from a single extra word, like in the following example (5a) from ["Dununba…"], to more sophisticated approaches as represented by (5b).

(5a) *Bámànan*$_1$ *kó*$_2$ *: « Jànfajuru*$_3$ *fyɛ́ku*$_4$ *kójugu*$_5$ *,* *à*$_6$ *bɛ*$_7$ *féreke*$_8$ *í*$_9$ *yèrɛ̂*$_{10}$ *kán*$_{11}$

lá$_{12}$ ».

‘{Les bambaras}$_1$ {disent}$_2$: {A force de}$_5$ {manier}$_4$ **{ta petite}$_0$** {corde de trahison}$_3$, {tu}$_6$ {finiras par}$_7$ {la nouer}$_8$ {autour de}$_{12}$ {ton}$_9$ {propre}$_{10}$ {cou}$_{11}$.’ = ‘'{The bamanas}$_1$ {say}$_2$: {By dint of}$_5$ {wielding}$_4$ **{your little}$_0$** {cord of betrayal}$_3$, {you}$_6$ {will end up}$_7$ {tying it}$_8$ {around}$_{12}$ {your}$_9$ {own}$_{10}$ {neck}$_{11}$.’

Here, *jànfajuru* ‘corde de trahison’ have neither an adjective nor the diminutive suffix in the Bamana text, while there is an extra description ‘ta petite’ in the French sentence.

(5b)	*Hálì*	*bì*	*bámànan*	*té*	*bálimamuso*
	même	aujourd’hui	bambara	COP.NEG	sœur

kó	*túlon*	*ná.*
affaire	jeu	à

‘Jusqu’à aujourd’hui, les bambaras ont une **grande considération** pour leurs sœurs.’ / ‘Until today, bamanas have had **great regard** for their sisters.’ [“Juguya…”].

4.5. Only in a few cases, size adjectives are used explicitly in Bamana. Two instances involving *bèlebele* ‘gros’ and *fítinin* ‘petit’ are shown in (6a,b):

(6a)	*Kùnɲɔgɔn*	*kélen*	*bɛ*	*dáfa*	*tùma*	*mîn*	*ná,*
	semaine	un	IPFV.AFF	compléter	moment	REL	à

ò	*y’*	*à*	*sɔ̀rɔ*	*ù*	*ye*	*fòrokɛnɛ*
ce	PFV.TR	3SG	obtenir	3PL	PFV.TR	champ. clarté

bèlebele	*yíriw*	*tìgɛ.*
gros	arbre.PL	couper

‘Une semaine après, ils avaient coupé les arbres sur une très **grande** surface.’ = ‘A week later, they had cut the trees over a very **large** area.’ [“Warabilen…”].

(6b)	*À*	*tóra*	*ò*	*cógo*	*lá*	*fɔ́*	*dón*	*dɔ́,*
	3SG	rester. PFV.INTR	ce	manière	à	jusqu'à	jour	certain

Ncí	*ye*	*búbaganton*	***fítinin***	*yé*	*tú*	*dɔ́*	*kɛ̀rɛfɛ̀*
NOM.M	PFV.TR	termitière	**petit**	voir	touffe	certain	côté.par

‘Tout resta comme ça jusqu'au jour où Nci vit une **petite** termitière près d'un bois.’ / ‘Everything remained like that until the day when Nci saw a **small** termite hill near a wood.’ [“Warabilen…”].

Such examples include also the pleonastic use of *fítinin* ‘petit’, e. g.:

(6c) *Jálakɔrɔka bέε, hálì dénmisɛn**nin*** ***fitininw, …***
TOP.GENT tout même petit.enfant.DIM **petit.PL**
‘Tout le monde à Dialakoro, même les **petits** enfants, …’ = ‘Everyone in Dialakoro, even **little** children, …’ [“Juman…”]. The word *dénmisɛn* is itself composed of *dén* ‘enfant’ and *mìsɛn* ‘petit’ and additionally gets here the diminutive suffix *-nin*.

5. Network analysis

Studies of languages using approaches from the theory of complex networks date back to early 2000s (Dorogovtsev & Mendes 2001; Ferrer i Cancho & Solé 2001) and remain an active field of research (Holovatch & Palchykov 2016; Markovič et al. 2019).

One of the approaches typically used to build a network from a text is as follows. Word types (in our case, autosemantic lemma types) are considered to be network vertices. Two vertices are connected by a link if the respective words are found in the same sentence. If there are several sentences where such two words occur, the links can be counted with multiplicity equal to the number of such sentences.

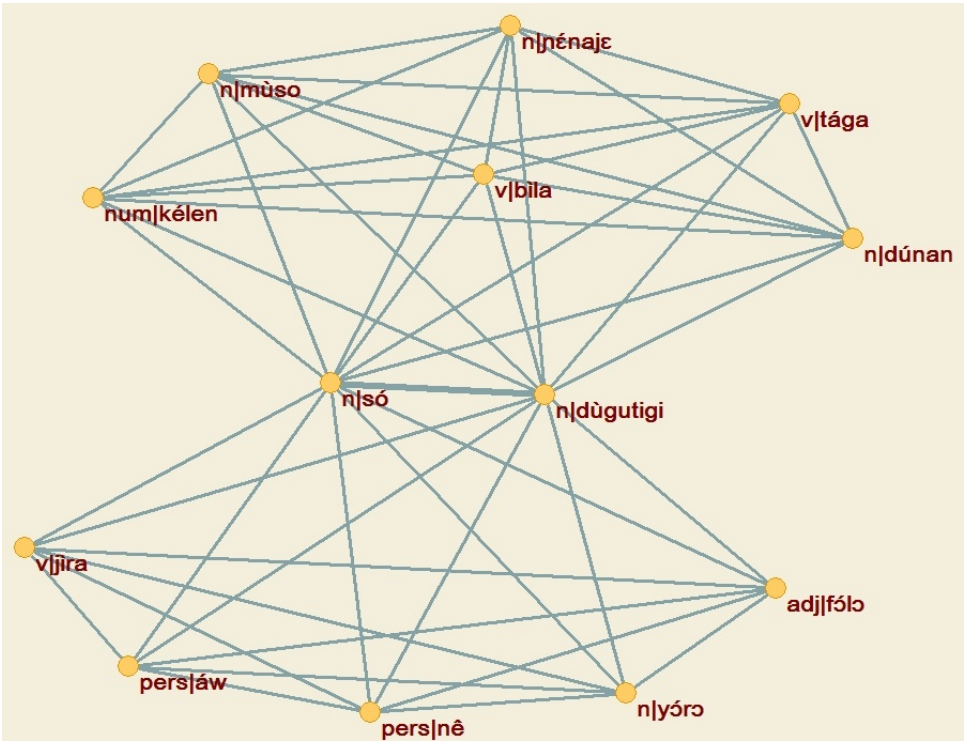

Figure 2. A sample network based on two sentences from the “Dununba…” tale. Note a thicker line between **n|só** and **n|dùgutigi**: this pair of vertices occurs twice in the sample sentences, so the link has multiplicity two.

The networks were built using own software (scripts in the Perl language). The visualization and calculation of the network parameters were made using the Pajek software (De Nooy, Mrvar & Batagelj 2011; Mrvar & Batagelj 1996–2018), which allows for the evaluation of many network properties, of which only those related to

distances between vertices are analyzed below in detail.

For illustration, consider a sample network in Figure 2, which is built using two consecutive sentences from the "Dununba…" tale. The sentences are (with autosemantic PoS given in boldface):

(7a)	***Áw***	*yé*	***dùgutigi***	*ka*	***só***	***yɔ́rɔ***	***jìra***
	pers	pm	**n**	pp	**n**	**n**	**v**
	2PL.EMPH	IMP	**chef.de.village**	POSS	**maison**	**lieu**	**montrer**

	nê	*lá*	***fɔ́lɔ***.
	pers	pp	**adj**
	1SG.EMPH	à	**premier**

'Mais **menez-moi d'abord** chez le **chef** du **village**.' = 'But **take me** to the **village chief first**.'

(7b)	***Mùso***	***kélen***	***bìlara***	***ɲɛ́najɛ***	***dúnan***	*ɲɛ́*	*kà*
	n	**num**	**v**	**n**	**n**	pp	pm
	femme	**un**	**mettre.PFV.INTR**	**réjouissance**	**étranger**	devant	INF

	tága	***dùgutigi***	*ka*	***só***.
	v	**n**	pp	**n**
	aller	**chef.de.village**	POSS	**maison**

'**Une** des **femmes accepta** de **l'accompagner** jusqu'à la **maison** du **chef**.' = '**One** of the **women agreed** to **accompany him** to the **chief's house**.'

A vertex can be isolated, that is, not linked to any other. Usually, such situations correspond to very short sentences often found in the direct speech. For instance, the vertex corresponding to *makɛ* 'maître' is isolated in the "Ntalen" tale. It appeared once in an exclamation translated as '– Eh, chef !':

(8)	<s>	*– Eɛ !*	</s>	<s>	***Mákɛ!***	</s>
		intj			**n**	
		pas.possible!			**maître**	

The simplified tags for the beginning of sentence <s> and for the end of sentence </s> are shown explicitly.

The distance between two non-isolated vertices is counted as the number of segments in the shortest path required to reach one vertex starting from the other. For instance, in Figure 2 the distance between **num|kélen** and **n|mùso** is $d = 1$, while the

distance between **num|kélen** and **v|jìra** is $d = 2$. Usually, even in large text networks mean values of the distance remain within 2.2–2.5 (Cong & Liu 2014; Caldeira et al. 2006; Buk, Krynytskyi & Rovenchak 2019). The short texts of tales have the mean distance values shifted towards $d = 2$, as expected, see Table 8.

The maximal distance between non-isolated vertices rarely exceeds 6, the language networks are thus regarded as "small worlds" (Ferrer i Cancho & Solé 2001) referring to the human society with "six handshakes rule" or "six degrees of separation" between people in the world (Watts 2004). Not surprisingly, in the analyzed case of short texts these values are smaller and most often equal to four, see Table 8.

Table 8. Some network properties of the Bamana and French versions of tales

	Dununba…		Juguya…		Juman…		Ntalen		Sigidankelen…		Warabilenkɔrɔ…	
	bam	fra	bam	fra	bam	fra	bam	fra	bam	fra	bam	fra
Sentences	127	146	63	64	66	68	139	118	75	57	136	150
Sent len (all)	11,7	11,2	10,4	12,0	11,0	14,2	12,4	17,0	13,1	21,1	12,1	13,4
Sent len (aut)	8,3	7,8	7,8	9,1	7,9	9,8	9,1	12,6	9,3	15,1	8,7	9,7
Mean distance	2,19	2,30	2,13	2,10	2,33	2,13	2,14	2,05	2,17	2,01	2,14	2,17
Max distance	4	4	4	4	5	4	4	3	4	3	4	5
Vertices	324	426	191	233	224	268	377	437	291	351	362	448
Links	6780	7336	2612	3768	3606	5078	8582	11704	4930	9570	7504	9700
Links per vertex	20,9	17,2	13,7	16,2	16,1	18,9	22,8	26,8	16,9	27,3	20,7	21,7

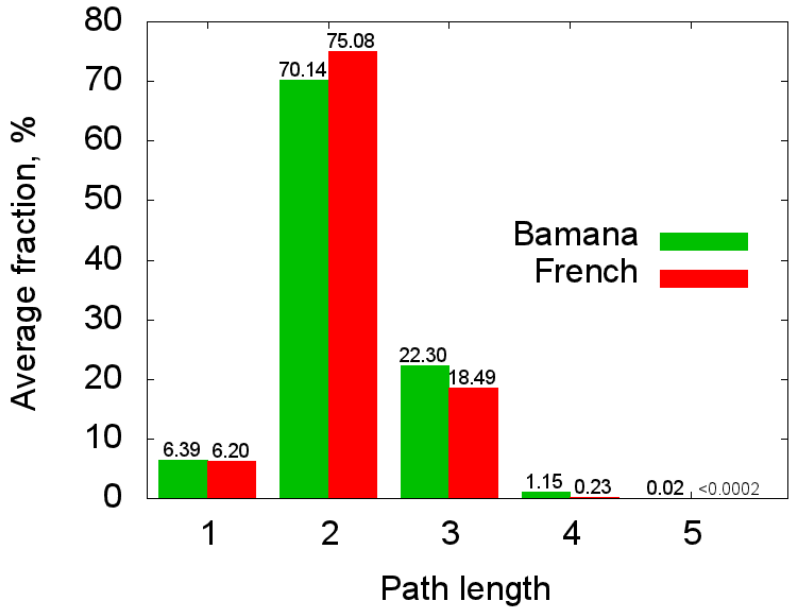

Figure 3. Distribution of distances between vertices in the networks of the Bamana (green) and French (red) versions of the tales. The average percentage is shown on the vertical axis.

It might be also interesting to look into the details of the path length distribution in the analyzed texts. A summary is shown in Figure 3. Most paths have length 2 (on average, over 70% in the Bamana texts and over 75% in the French texts). On the other hand, lengths 3 are slightly more frequent in the Bamana texts (22.3% versus 18.5%). The differences between the numbers, however, are not significant enough to draw any far-reaching conclusions.

Interestingly, in Bamana the correlation between sentence length and mean distance is not very significant, while in French the inverse correlation in very well pronounced, i. e., shorter sentences yield larger mean distances. The correlation coefficient in French is –0.84 versus –0.38 in Bamana. The reason is that mean sentence lengths are more evenly distributed in the Bamana texts (7.8 to 9.3) than in the French ones (7.8 to 15.1).

The number of vertices, as given in Table 8, is nothing but the number of autosemantic lemma types in Table 3. From the number of links per vertex one can conclude that, depending on the text, each lemma co-occurs in a sentence on average with 14–23 other lemmas in the Bamana texts and with 16–27 other lemmas in the French texts under study.

The highest number of links ranges from 118 in “Juman…” to 295 in “Ntalen”. Almost always it is associated with the pronoun *à* ‘3SG’ and only in “Juman…” it corresponds to the pronoun *ò* ‘ce’, with *à* ‘3SG’ on the second place having 104 links. A similar behavior is found in French.

My initial expectation was that mean distances in the networks for Bamana texts would be smaller compared to French ones. The reason is the smaller number of types covering a larger portion of texts in Bamana, see Table 7 and the frequency data by Rovenchak & Buk (2013). This was confirmed for the first analyzed text, “Dununba…”. However, an opposite relation was found for five other texts, see Table 8.

The observed data suggest, in particular, that mean distances in a text network are mostly defined by mean sentence lengths rather than some deeper properties of languages. Mean sentence lengths, on the other hand, are believed to be good author style markers (Yule 1939; Sichel 1974; Pande & Dhami 2015).

In the case of the texts under study it appears that differences in sentence lengths are often defined by the representation of the direct speech in the corpus. A proper treatment of the direct speech might require extending the end-of-sentence markers beyond the standard set of full-stop ‘.’, exclamation mark ‘!’, question mark ‘?’, and ellipsis ‘…’ (cf. Martin et al. 2003; Rovenchak & Buk 2013).

6. Conclusions

The results presented in the present work allow for conclusions in several domains: lemmatization and tagging of French texts in the Bamana–French parallel corpus, which has not been implemented yet, parts of speech distributions in Bamana and French with a special focus on adjectives, and network properties of texts.

From the preliminary preparation of the French texts for the analysis, namely, automated lemmatization and part-of-speech tagging, one can conclude that the TreeTagger software yields satisfactory results but requires additional manual tuning. The observations made in this work suggest how this tuning can be partly automated as well.

As a by-product of the network analysis of texts, the need to unify approaches to the treatment of sentence breaks in the direct speech comes out. This applies to both Bamana and French texts and should be taken into considered in the subsequent development of the Bamana Reference Corpus. Another conclusion to be drawn from the network analysis has a negative hue: it seems that some properties of text networks are just defined by the mean sentence length and might be of little use for in-depth language studies, especially for relatively short texts. In prospect, approaches not relying on sentence boundaries can be used to build text networks and study their properties.

The main body of the results concerns the distribution of autosemantic parts of speech in text and vocabulary. The analysis of the Bamana and French versions of the tales has revealed the similarities and differences between the languages. One of such peculiarities, the absence of adjectives among the most frequent words in Bamana, is discussed in detail through the analysis of size adjectives and several ways of their representation in Bamana compared to the French translations. Additional studies involving more texts are required to distinguish between language-related and genre-related features in detail.

Further research would include analysis of other pairs of texts and text collections from the Bamana–French parallel corpus, especially of different genres, as well as eventual expansions to the Maninka–French parallel corpus and inclusion of other language pairs (cf. Vydrine, Togo & Bulman 2017).

Acknowledgment

I am grateful to Valentin Vydrin for the discussions on some issues raised in the paper and for the help with the Bamana language and to Angela Kamianets for reading the manuscript.

Glosses

1,2,3	1st, 2nd,3rd person	NOM.M	male name
ADR	address postposition	OPT2	optative
AFF	affirmative	QUOT	quotation copula
AUGM	augmentative	PFV	perfective
COP	copula	PL	plural
DEF	"new definite article"	PL2	non-productive plural
DIM	diminutive	POSS	possessive
EMPH	emphatic	PP	polysemic postposition
EQU	equative copula	PROG	progressive
GENT	"genitive" suffix	REFL	reflexive
ID	identification copula	REL	relativization
IMP	imperative	SBJV	subjunctive
INF	infinitive	SG	singular
INTR	intransitive	TOP	toponym
IPFV	imperfective	TR	transitive
NEG	negative		

Parts of speech

adj	adjective	pers	personal pronoun
adv	adverb	pm	predicative marker
conj	conjunctive	poss	possessive pronoun
cop	copula	pp	postposition
dtm	determinative	prn	pronoun
n	noun	ptcp	participle
n.prop	proper noun	v	verb
num	numeral	vq	qualitative verb
ptcp	participle		

References

Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Rita Veloso, Sandra Antunes, Florbela Barreto & Raquel Amaro. 2005. The Portuguese corpus. In Emanuela Cresti & Massimo Moneglia (eds.), *C-ORAL-ROM : integrated reference corpora for spoken Romance languages*, 163–207. Amsterdam–Philadelphia: John Benjamins Publishing Company.

Bailleul, Charles, Artem Davydov, Anna Erman, Kirill Maslinksy, Jean Jacques Méric & Valentin Vydrin. 2011–2020. *Bamadaba : Dictionnaire électronique bambara-français, avec un index français-bambara*. http://cormand.huma-num.fr/bamadaba.html.

Baisa, Vít, Jan Michelfeit, Marek Medveď & Miloš Jakubíček. 2016. European Union Language Resources in Sketch Engine. In *The Proceedings of tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Borin, Lars (ed.). 2002. *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam–New York, N.Y.: Rodopi.

Buk, Solomija. 2012. Arkhitektura pol's'ko-ukrajins'koho ta ukrajins'ko-pol's'koho paralel'noho korpusu avtoperekladiv Ivana Franka [The architecture of Polish-Ukrainian and Ukrainian-Polish parallel corpus of Ivan Franko's self-translations]. *Slavia Orientalis* LXI(2). 213–230.

Buk, Solomija, Yuri Krynytskyi & Andrij Rovenchak. 2019. Properties of autosemantic word networks in Ukrainian texts. *Advances in Complex Systems* 22(6). 1950016. https://doi.org/10.1142/S0219525919500164.

Caldeira, S. M. G., T. C. Petit Lobão, R. F. S. Andrade, A. Neme & J. G. V. Miranda. 2006. The network of concepts in written texts. *European Physical Journal B* 49(4). 523–529. https://doi.org/10.1140/epjb/e2006-00091-3.

Cong, Jin & Haitao Liu. 2014. Approaching human language with complex networks. *Physics of Life Reviews* 11(4). 598–618. https://doi.org/10.1016/j.plrev.2014.04.004.

Creissels, Denis. 1983. Reflexions sur le système prédicatif du bambara. *Mandenkan* 6. 21–36.

Creissels, Denis. 2003. Adjectifs et adverbes dans les langues subsahariennes. In Patrick Sauzet & Anne Zribi-Hertz (eds.), *Typologie des langues d'Afrique & universaux de la grammaire. Volume 1: Approches transversales: domaine bantou*, 17–38. Paris: L'Harmattan.

Creissels, Denis. 2009. *Le malinké de Kita: Un parler mandingue de l'ouest du Mali*. Köln: Köppe.

Creissels, Denis & Pierre Sambou. 2013. *Le mandinka. Phonologie, grammaire, textes*. Paris: Karthala.

De Nooy, W., A. Mrvar & V. Batagelj. 2011. *Exploratory Social Network Analysis with Pajek* (Structural Analysis in the Social Sciences). 2nd edn. Cambridge University Press.

De Pauw, Guy, Peter Waiganjo Wagacha & Gilles-Maurice de Schryver. 2011. Exploring the SAWA corpus: collection and deployment of a parallel corpus English–Swahili. *Language Resources and Evaluation* 45(3). 331–344. https://doi.org/10.1007/s10579-011-9159-7.

Diarra, Oumar Niananko ro & Antoine Fenayon. 2011a. *Dununba kumata : Mali*

nsiirinw. Paris: Donniyakadi.

Diarra, Oumar Nianankoro & Antoine Fenayon. 2011b. *Le tam-tam qui parle : contes du Mali*. Paris: Donniyakadi.

Dorogovtsev, S. N. & J. F. F. Mendes. 2001. Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268(1485). 2603–2606. https://doi.org/10.1098/rspb.2001.1824.

Doval, Irene & M. Teresa Sánchez Nieto (eds.). 2019. *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.90.

Dumestre, Gérard. 2003. *Grammaire fondamentale du bambara*. Paris: Karthala.

Dumestre, Gérard. 2011. A propos des adverbes du bambara, ou de l'art d'accommoder les restes. *Mandenkan* 47. 3–11.

Ferrer i Cancho, Ramon & Richard V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268(1482). 2261–2265. https://doi.org/10.1098/rspb.2001.1800.

Hansen-Schirra, Silvia, Stella Neumann & Oliver Čulo (eds.). 2017. *Annotation, Exploitation and Evaluation of Parallel Corpora: Tc3 1*. Language Science Press.

Holovatch, Yurij & Vasyl Palchykov. 2016. Complex Networks of Words in Fables. In R. Kenna, M. MacCarron & P. MacCarron (eds.), *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*, 159–175. Springer International Publishing. https://doi.org/10.1007/978-3-319-39445-9_9.

Kastenholz, Raimund. 1998. *Grundkurs Bambara (Manding) mit Texten*. 2. überarbeitete Auflage. Köln: Köppe.

Markovič, Rene, Marko Gosak, Matjaž Perc, Marko Marhl & Vladimir Grubelnik. 2019. Applying network theory to fables: complexity in Slovene belles-lettres for different age groups. *Journal of Complex Networks* 7(1). 114–127. https://doi.org/10.1093/comnet/cny018.

Martin, Joel, Howard Johnson, Benoit Farley & Anna Maclachlan. 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*, 115–118.

Maslinsky, Kirill. 2014. Daba: a model and tools for Manding corpora. In *TALN-RECITAL 2014 Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014: African Language Processing)*, 114–122. Marseille, France: Association pour le Traitement Automatique des Langues. https://www.aclweb.org/anthology/W14-650.

Moropa, Koliswa. 2007. Analysing the English-Xhosa parallel corpus of technical texts with Paraconc: a case study of term formation processes. *Southern*

African Linguistics and Applied Language Studies 25(2). 183–205. https://doi.org/10.2989/16073610709486456.

Mrvar, A. & V. Batagelj. 1996–2018. *Pajek: analysis and visualization of large networks*. http://mrvar.fdv.uni-lj.si/pajek/.

Pande, Hemlata & Hoshiyar S. Dhami. 2015. Determination of the Distribution of Sentence Length Frequencies for Hindi Language Texts and Utilization of Sentence Length Frequency Profiles for Authorship Attribution. *Journal of Quantitative Linguistics* 22(4). 338–348. https://doi.org/10.1080/09296174.2015.1106269.

Popescu, Ioan-Iovitz, Gabriel Altmann & Reinhard Köhler. 2010. Zipf's law—another view. *Quality and Quantity* 44(4). 713–731. https://doi.org/10.1007/s11135-009-9234-y.

Rijkhoff, Jan & Eva van Lier (eds.). 2013. *Flexible Word Classes: Typological studies of underspecified parts of speech*. Oxford: Oxford University Press.

Rovenchak, Andrij & Solomija Buk. 2013. *Masadennin* (*The Little Prince* in Bamana): Experimental online concordance with parallel French and English texts. *Mandenkan* 50. 117–130.

Rychlý, Pavel & Vít Suchomel. 2016. Annotated Amharic Corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.), *Text, Speech, and Dialogue. TSD 2016. Lecture Notes in Computer Science, vol. 9924*, 295–302. Cham: Springer. https://doi.org/10.1007/978-3-319-45510-5_34.

Salamanca, Angel. 2019. A first look at StanfordNLP — A Python library ready to process 53 human languages. https://levelup.gitconnected.com/first-look-at-stanfordnlp-2b7d43190957.

Schmid, Helmut. n. d. TreeTagger - a part-of-speech tagger for many languages. https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/.

Segerer, Guillaume. 2008. Closed adjective classes and primary adjectives in African Languages. https://halshs.archives-ouvertes.fr/halshs-00255943.

Sichel, H. S. 1974. On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society. Series A (General)* 137(1). 25–34. https://doi.org/10.2307/2345142.

Sitchinava, Dmitri V. 2016. Evropejskij perfekt skvoz' prizmu parallel'nogo korpusa [European perfect through the prism of a parallel corpus]. *Acta Linguistica Petropolitana* XII(2). 85–114.

Stein, Achim. 2003. French TreeTagger Part-of-Speech Tags. https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/french-tagset.html.

Tröbs, Holger. 2008. Le Bambara. In Holger Tröbs, Eva Rothmaler & Kerstin Winkelmann (eds.), *La qualification dans les langues africaines = Qualification in African Languages*, 13–28. Köln: Köppe.

Tröbs, Holger. 2014. Some notes on the encoding of property concepts in Vai from a typological and comparative Central Mande perspective. *Mandenkan* 52. 111–130.

Vydrin, Valentin. 2013. Bamana Reference Corpus (BRC). *Procedia - Social and Behavioral Sciences* 95. 75–80. https://doi.org/10.1016/j.sbspro.2013.10.624.

Vydrin, Valentin. 2017a. Bamana jazyk [Bamana]. In Valentin Vydrin, Yulia Mazurova, Andrej Kibrik & Elena Markus (eds.), *Jazyki mira: Jazyki mande [Languages of the world: Mande languages]*, 46–143. St. Petersburg: Nestor-Historia.

Vydrin, Valentin. 2017b. Encoding of property concepts in Maninka of Guinea. In Valentin F. Vydrin & Anastasia V. Lyakhovich (eds.), *In the hot yellow Africa… In honor of Alexander Zheltov on the occasion of his 50th birthday*, 25–47. St. Petersburg: Nestor-Historia.

Vydrin, Valentin, Kirill Maslinsky, Jean Jacques Méric & Andrij Rovenchak. 2011–2019. Corpus Bambara de Référence (parallel Bambara-French subcorpus. http://cormand.huma-num.fr/sc_parallele.html.

Vydrine, Valentin. 1999. Les parties du discours en bambara : un essai de bilan. *Mandenkan* 35. 72–93.

Vydrine, Valentin F., Amadou Togo & Stephen P. D. Bulman. 2017. Bamana text and English translation of the epic of Sumanguru by Abdulaye Sako. *African Sources for African History* 15. 72–153.

Wallmach, Kim. 2000. Examining simultaneous interpreting norms and strategies in a South African legislative context: a pilot corpus analysis. *Language Matters* 31(1). 198–221. https://doi.org/10.1080/10228190008566165.

Watts, Duncan J. 2004. *Six Degrees: The Science of a Connected Age*. New York–London: W. W. Norton & Company.

Wójtowicz, Beata. 2018. Evaluating a 12-Million-Word Corpus as a Source of Dictionary Data. *International Journal of Lexicography* 31(3). 327–341. https://doi.org/10.1093/ijl/ecx011.

Woldeyohannis, Michael Melese, Laurent Besacier & Million Meshesha. 2018. A Corpus for Amharic-English Speech Translation: The Case of Tourism Domain. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 129–139. Springer International Publishing. https://doi.org/10.1007/978-3-319-95153-9_12.

Yule, G. Udny. 1939. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika* 30(3/4). 363–390. https://doi.org/10.2307/2332655.

Андрей Ровенчак

Баманские сказки, записанные Умару Ньянанкоро Джара: Сравнительное исследование на основе параллельного бамана-французского корпуса

В статье представлен анализ распределения автосемантических (знаменательных) частей речи в баманском и французском текстах баманских сказок, записанных Умару Ньянанкоро Джара. С этой целью использован параллельный бамана-французский корпус. Основное внимание уделено частотным соотношениям различных частей речи в двух языках. Составлен список слов, общих для всех текстов сказок. Проанализированы особенности функционирования в бамана прилагательных, обозначающих размер. Также кратко обсуждается применение теории сложных сетей.

Ключевые слова: параллельный корпус, бамана, французский язык, автосемантические части речи, частотный анализ, сложные сети.

Andrij Rovenchak

Bamana tales recorded by Umaru Ɲanankɔrɔ Jara: A comparative study based on a Bamana–French parallel corpus

The paper presents an analysis of the distribution of autosemantic (meaningful) parts of speech in Bamana and French texts of Bamana tales recorded by Umaru Ɲanankɔrɔ Jara. It is carried out using a Bamana–French parallel corpus. The focus is on part-of-speech frequencies in the two languages. List of words common to all the texts of the tales are compiled. Details of the representation of size adjectives in Bamana are analyzed. An application of the theory of complex networks is also briefly discussed.

Key words: parallel corpus, Bamana, French, autosemantic parts of speech, frequency analysis, complex networks.

Andrij Rovenchak

Contes bambara enregistrés par Umaru Ɲanankɔrɔ Jara: Une étude comparative basée sur un corpus parallèle bambara-français

L’article présente une analyse de la distribution des parties du discours autosémantiques (mots lexicaux) dans les textes bambara et français des contes bambara enregistrés par Umaru Ɲanankɔrɔ Jara. Cette analyse est réalisée à l'aide d'un corpus parallèle bambara-français. L'accent est mis sur les fréquences des parties du discours dans les deux langues. Une liste de mots communs à tous les textes des contes est compilée. Les détails de la représentation des adjectifs de taille en bambara sont analysés. Une application de la théorie des réseaux complexes est également brièvement discutée.

Mots clés: corpus parallèle, bambara, français, parties du discours autosémantiques, analyse fréquentielle, réseaux complexes.