Mandenkan
Bulletin semestriel d'études linguistiques mandé

62 | 2019
Numéro 62

Positional skipgrams for Bambara: a resource for corpus-based studies

Les skipgrams *positionnels pour le bambara : une ressource pour la recherche linguistique orientée corpus*

ПОЗИЦИОННЫЕ СКИП-ГРАММЫ ДЛЯ БАМАНА: РЕСУРС ДЛЯ КОРПУСНЫХ ИССЛЕДОВАНИЙ

Kirill Maslinsky

Electronic version
URL: http://journals.openedition.org/mandenkan/2119
ISSN: 2104-371X

Publisher
INALCO

Electronic reference
Kirill Maslinsky, « Positional skipgrams for Bambara: a resource for corpus-based studies », *Mandenkan* [Online], 62 | 2019, Online since 07 May 2020, connection on 14 May 2020. URL : http://journals.openedition.org/mandenkan/2119

Mandenkan, No 62, 2019, pp. 165-183

Positional skipgrams for Bambara: a resource for corpus-based studies

Kirill Maslinsky
Institute of Russian Literature (Pushkinskij Dom) RAS
Saint Petersburg, Russia
maslinych@gmail.com

1. Introduction

N-grams — fixed-length sequences of adjacent tokens collected from textual data — have been widely used in computational linguistics and natural language processing for decades (Rosenfeld 2000). A frequency list of n-grams obtained from a corpus has proven to be a simple yet powerful tool to represent contextual information and sequential phenomena in natural language. Publishing n-gram frequencies is also a way to share statistics on word distribution in a corpus, the most notable example being the Google Ngrams corpus (Brants & Franz 2006).

The idea that is key to the practical success of n-grams in a wide variety of language modeling tasks (from spelling correction to speech recognition) is to extract the information encoded in the relative positioning of linguistic units into a list of easily quantified atomic "co-occurrence events". When used in a general sense, the approach leaves room for flexibility in choosing how to build n-grams, and what to include in them. Adjacency constraints can be relaxed to include items occurring anywhere within a fixed-width context window, thus producing *skipgrams*. There is also no need to limit the scope to the lexical or graphical level, as in the traditional word n-grams and letter n-grams, or even to the surface level in general. In cases when linguistic annotation is available for the text, it may be used for building n-grams. The most common example of the latter is to make n-grams from part of speech tags of subsequent words that reveal recurrent word order patterns. The part-of-speech n-grams are used in diverse fields, for instance, in text-to-speech generation (Taylor & Black 1998) or in sentiment analysis (Jaggi, Uzdilli & Cieliebak 2014). Thus n-grams can represent phenomena other than plain lexical co-occurrence.

This article presents a new online dataset of linguistically rich n-gram frequency data for Bambara based on the disambiguated part of the Corpus bambara de référence[1] (Vydrin 2013). N-grams in this dataset were constructed with the aim to capture those types of information that are available in the morphologically annotated corpus of Bambara. Beyond the usual lexical focus, n-grams were supplemented with paradigmatic grammatical information and positional features that should allow for inferences to be made about various aspects of morphosyntax.

Making this dataset publicly available is a way to provide access to the linguistic data derived from the full annotated corpus for a wider audience of students and researchers without disclosing copyright-protected texts. The data format has to be general enough to allow open-ended exploration and use of the data in broad areas of linguistic research, language learning, and downstream NLP tasks. In my view, the n-grams list format matches this objective and has the additional benefit of retaining readability by a human as well as a machine. While simple tabular format makes data easily quantifiable for research and engineering tasks, for a human reader, a frequency-ordered n-grams list preserves meaningful linguistic categories such as lexemes, grammatical tags, and relative word positions in a sentence.

The article is structured as follows. Sections 2–4 explain the methodology and data used for constructing n-grams for Bambara, followed by section 5 with brief illustrations of how the n-gram data presented here may be employed in corpus-based linguistic research. A discussion of the advantages that positional skipgrams provide in the low-resourced setting is presented in section 6.

2. Positional Skipgrams

The approach used in this article to combine lexical, grammatical, and positional information in a single n-gram framework is tentatively labeled here *positional skipgrams*. To make sense of this framework, consider a sentence in Bambara that has part of speech tags defined for each token.

(1)	*í*	*k'*	*à*	*dɔ́n*	*kó*	*nàta*	***mùso***	*tɛ*	*ná*	*ɲùman*	*tóbi*	.
	pers	pm	pers	v	cop	n	n	pm	n	adj	v	c
		pm:-5	*pers:-4*	*v:-3*	*cop:-2*	*n:-1*	***n:0***	*pm:1*	*n:2*	*adj:3*	*v:4*	*c:5*

'You should know that a greedy woman won't cook a good sauce'

[1] The corpus search interface as well as general info about the corpus are available online at: *http://cormand.huma-num.fr/*. The new dataset is available at *http://cormand.huma-num.fr/ngrams*.

For this sentence, the list of regular word n-grams (*bigrams*)[2] would include the pairs of consecutive words: *í – k'*, *k' – à*, *à – dɔ́n*, etc. This is the most common ("default") reading of the term *word n-grams* in the literature. In case of *skipgrams*, pairs of all words that fall within the fixed-width context window (five words on each side in our example) are considered a co-occurrence. The list of skipgrams would include pairs that are up to five words apart in the sentence: *í – à*, *í – dɔ̀n*, *í – kó*, *k' – mùso*, etc. In contrast to regular word skipgrams, in *positional skipgrams* a numeric index is appended to the second item in the pair that reflects its relative position in respect to the first item: 1 indicates the next word to the right, -3 indicates the third word on the left, and 0 is the word itself. Besides that, in the variant of positional skipgrams presented here part-of-speech tag is used instead of the co-occurring word as the second item of the skipgram. In our example, for the word *mùso* the following *positional skipgrams* will be generated: *mùso – n:0*, *mùso – pm:1*, *mùso – n:-1*, etc.[3] Essentially, this list may be read as a set of statements equivalent to: "in this sentence, *mùso* co-occurs with a noun in a previous position, with an auxiliary (predicative marker) in the following position, and is itself tagged as noun".

Instead of tracking word co-occurrence events, positional skipgrams record the information on the occurrence of the word in a certain position in the surface syntactic structure, to the extent that syntactic information is reflected in the sequence of part of speech tags. As usual with n-grams, this positional occurrence is represented as a series of atomic "co-occurrence events". In this representation, the structure of the context is lost, but the disparate events (words and sentences) thus become comparable. For example, two occurrences of a word can share a significant part of their positional skipgrams while not sharing that many context words. The same principle makes it possible to compare different words by the similarity of their syntactic contexts (in terms of the relative frequencies of their positional collocates).

While the idea of appending the positional index to the collocate is all that is needed to define positional skipgrams in general, several other constraints should be observed to make them more relevant as linguistic data and to make sure that they are tractable in downstream computational tasks.

1. Note that in the examples above words are never included as positional collocates to other words. While technically nothing prevents us from doing so, the focus of the method is to relate words to the underlying linguistic categories,

[2] The term n-gram presupposes a variable number of co-occurring words, but in the data and in the examples discussed in this article the *n* is always limited to two.

[3] Depending on the task at hand, it may be convenient to record reverse co-occurrence events (*pm:1 – mùso*, etc.) simultaneously to simplify further processing.

and more generally, to recurrent phenomena at the non-lexical level. Essentially, what we are interested in is the type of contexts that words are likely to *share*. Moreover, in a less-resourced setting where lexical data are already sparse, multiplying the lexicon size by the positional dimension would be clearly detrimental for statistical inference of any kind.

2. Since the positional part of speech tags are included as a proxy for syntactic structure, it is reasonable to require that n-grams do not cross sentence boundaries. At the same time, punctuation tokens could be recorded as collocates to keep track of the relative positions of the word in respect to sentence and clause boundaries (for instance, the final stop in the example *(1)* that would produce *mùso – c:5*).
3. To further compensate for lexical sparsity, it makes sense to include n-grams consisting of two positional tags alongside positional skipgrams with words. For instance, accumulating frequency counts for *n:0 – pm:1* would help track the fact that nouns tend to fill the position before predicative markers as an integral part of the data.

3. Related work

In such a long and rich tradition as application of n-grams in natural language processing hardly anything can be truly novel. But to summarize, compared to other n-gram building methods positional skipgrams are characterized by the two distinctive features: they combine information from different levels of annotation, and they incorporate positional information into the n-gram in the form of a positional index.

Positional skipgrams as implemented in this article combine features from two different levels of annotation in the form of n-grams, namely lexical items and grammatical categories. This simple cross-level setup seems to be uncommon in practical n-gram applications in recent literature on natural language processing. This could be due to the fact that in the history of language modeling with n-grams grammatical categories (part-of-speech tags and the like) were primarily seen as a desired result to be produced by the model or at least as a latent variable for better word prediction, but definitely not as input data (see, for example, (Brown et al. 1992)).

In contrast, in more basic research, where the goal is language description rather than solving applied tasks, there is a rich tradition of looking for patterns that combine lexical items with syntactically defined slots. In corpus linguistics, the constructs that encompass both lexical and grammatical components in a single pattern were used to identify idiomatic constructions, and to make inferences about lexical meaning (e.g. polysemy) based on usage. These include *behavioral profiles* suggested by Hanks

(1996) as a generalization of verb complimentation patterns; *collostructions* (Stefanowitsch & Gries 2003) that track co-occurrence between words and constructions; *colligations* as "co-occurrence of word forms with grammatical phenomena" (Gries & Divjak 2009); and more ad-hoc instruments, like gapped patterns used to identify grammatical constraints in multi-word expressions (Kopotev et al. 2013). A common methodological feature of all the above approaches is that to collect data, researchers have to pre-define a specific construction or pattern they are looking for. The positional skipgrams approach is different from all the above constructs in that it does not specify a particular construction, but rather captures any constructions that can be reduced to the set of lexical items and grammatical categories positioned in text at some fixed interval in respect to each other.

Positional skipgrams explicitly record the position of a collocate relative to the current word. Common skipgram-based models may incorporate positional information implicitly. In particular, it has been shown that word2vec actually benefit from taking distances between words into account by using the decreasing weight coefficient for more distant words (Levy, Goldberg & Dagan 2015). The closest to our approach is the work by Ling et al. (2015) that included "what words go where" type of information in addition to "what words go together" in word2vec by creating separate models for each position of a context word relative to the current word. The idea to have positional information as a part of term in n-gram itself was motivated by the example *rythmical n-grams* in the work of Petr Plecháč in quantitative analysis of poetry. He uses n-grams to represent the structural position of sounds in the verse line (Plecháč 2019: 38).

4. Dataset description

The dataset presented in this article was built using the manually disambiguated part of the Bambara Reference Corpus (corbama-net). As of December 2019, the disambiguated subcorpus contains 1.3M words in 1650 documents. The corpus provides token-level morphological annotation as well as document-level metadata on the author, the source of the text, and several tags categorizing the medium, genre, and theme of the text (on metadata, see for details: (Davydov 2010)). The annotation provided in the corpus was obtained using the morphological processor Daba based on a dictionary and a set of rules (Maslinsky 2014), followed by manual disambiguation by Bambara-proficient operators.

The annotation layers available in this subcorpus include the orthographically normalized token (part of the corpus is in the old Bambara orthography), lemma, part of speech tag, and a gloss (lexical equivalent) in French. For multi-morpheme words there is also a recursive structure that annotates each morpheme with the similar

attributes of a form, a part of speech tag, and a gloss. Grammatical morphemes, as well as standalone function words are assigned a Leipzig-style formal gloss from a standard list of glosses for Bambara[4] instead of the French equivalent.

The main objective of publishing this dataset is to present quantitative data on morphosyntactic regularities and variation in the corpus. Hence other types of variation that are attested in the corpus are not represented, namely orthographic variation, dialectal variation, and tonal variation. To eliminate these types of variation only orthographically normalized forms are used throughout the dataset. All variants of the same lemma (dialectal, tonal, phonetic, etc.) were transformed to the canonical form, which is operationalized as the first variant listed for a lexical entry in the Bamadaba dictionary[5].

To make the most of the structural information available in the annotation, the basic positional skipgrams model presented above is supplemented with the n-grams based on the morpheme-level grammatical information. To keep data sparsity at a manageable level, the principle of limiting the right-hand side of the n-grams to the closed-class and frequent phenomena was observed (see section 2 for details). Thus out of the morpheme-level annotation layer only morphological tags from a standard list of glosses were taken into account. The resulting list of skipgrams includes the pairs of the following form:

- wordform (or lemma) — part of speech tag + position
- part of speech tag — part of speech tag + position
- wordform (or lemma) — standard gloss + position
- standard gloss — standard gloss + position

Numerals and punctuation are not included as the left-hand side items in the n-grams, but may appear as positional collocates on the right-hand side. The context window width for building skipgrams is defined to be five tokens on each side of the word, but is not allowed to cross sentence boundaries. Sentence boundaries are included in the list of positional collocates using a conventional *SENT* tag. The choice of five tokens as a context window width is arbitrary, although it is in accord with the common practice in other n-gram-based models. It is reasonable to expect that clause length in Bambara will not frequently exceed this width, so that not much useful statistics could be collected with a wider context window.

[4] See the full list of the glosses for grammatical morphemes and auxiliaries for Bambara at: http://cormand.huma-num.fr/gloses.html.

[5] See information on the dictionary at http://cormand.huma-num.fr/bamadaba.html.

For the convenience of dataset users, the skipgram frequency data is presented in several variants. First, the data is split according to the basic lexical item used for building skipgrams that is either an orthographically normalized wordform, or a canonical lemma. Second, frequency data on both wordfrom-based and lemma-based skipgrams are presented in two forms: an aggregated variant showing total counts for a whole corpus, and a disaggregated variant showing document-level frequencies.

The data is presented in a text-based tabular format. Skipgram frequency tables are in the TSV (tab separated values) format and contain the following columns:

- lexical item, tag or standard gloss;
- its positional collocate;
- total frequency of the lexical item/tag/gloss;
- total frequency of the collocate;
- frequency of the co-occurrence of the item with the collocate (n-gram frequency);
- a label indicating the type of the collocate (word–tag, tag–tag, etc.) to facilitate filtering.

The document-level frequency data has an additional column with the document ID that precedes the list. Document-level metadata are provided as a separate CSV file that can be linked to the document-level skipgram frequency tables based on the value of the document ID field.

5. Possible applications

This section presents a few examples of the ways in which information contained in the positional skipgrams can be rearranged and explored to address linguistic queries. The statistical processing of the data in the examples is intentionally kept to a minimum, in order to demonstrate conceptual simplicity and interpretability that the lists of positional skipgrams can offer by themselves. The examples presented in this section are neither an exhaustive list of the uses for positional skipgrams, nor a set of finished linguistic case-studies in Bambara; they are meant to serve just as illustrations of possible applications.

Lexical comparison

Let’s start with a simple query on lexical semantics where the application of the positional skipgrams is quite straightforward. In Bambara, there is a pair of moderately frequent verbs, *gòsi* and *bùgɔ*, both of which mean ‘to hit’. Having corpus data at hand, we may make inferences about the semantic differences of these two verbs based on

the differences in their context distributions. In addition to the traditional reading of the concordance for both verbs, positional skipgrams can offer a summary of morphosyntactic positions of each verb together with frequency statistics (see table *1*).

Table 1. Top 8 frequent positional skipgrams for *bùgɔ* and *gòsi*

item	collocate	freq1	freq2	ngram
bùgɔ_v	v:0	187	188431	187
bùgɔ_v	pm:-2	187	146505	126
bùgɔ_v	pers:-1	187	179651	76
bùgɔ_v	c:1	187	130483	64
bùgɔ_v	pers:-3	187	145917	55
bùgɔ_v	3SG:-1	187	81640	43
bùgɔ_v	INF:-2	187	49982	42
bùgɔ_v	n:-1	187	309187	38
gòsi_v	v:0	285	188431	285
gòsi_v	pm:-2	285	146505	179
gòsi_v	n:-1	285	309187	91
gòsi_v	PFV.TR:-2	285	21125	84
gòsi_v	num:3	285	20081	75
gòsi_v	n.prop:-1	285	41012	74
gòsi_v	conj:2	285	45496	73
gòsi_v	pers:-1	285	179651	71

Columns indicate: freq1 — the frequency of the verb itself; freq2 — total frequency of a collocate in a corpus; ngram — frequency of co-occurrence of a collocate with the verb.

The data in table *1* essentially presents an excerpt from the unaltered table of aggregated counts of positional skipgrams on the whole Bambara corpus. The only operations needed to get this view are just proper filtering (all lines including *gòsi_v* and *bùgɔ_v*) and sorting (in the descending order of skipgram frequency). Yet even this simple frequency list immediately reveals differences in use that point to the semantic contrast between these two lexical items. While the first two positions in the list for both verbs are trivial in that they just reflect the part of speech and the position of the verb in a clause (S AUX O V), the third position is of particular interest because it reflects the position of the direct object, and is different for the two verbs. Taken

together, all n-grams that refer to that position in the top of the lists indicate, that for *bùgɔ*, personal pronouns (especially 3SG) dominate over nouns in the position of the direct object, while for *gòsi* the position of a direct object is more equally distributed among nouns, proper nouns, and personal pronouns. Thus a hypothesis may be formulated that *bùgɔ* is preferred when talking about hitting people, while *gòsi* is more general and probably more suitable in talking about hitting objects.

Interpretation of raw frequency data may be suggestive, but it is misleading in many cases. While frequencies of the two verbs in question are on the same order of magnitude, they still differ by a factor of 1.5, which makes numbers in the two lists not directly comparable. A more principled way to identify differences in usage would require some sort of a statistical model that takes into account the differences in frequencies. There are plenty of approaches to this task in natural language processing. For the purposes of this demonstration we adopt a weighted log-odds model suggested in Monroe, Colaresi & Quinn (2008).

To put it simply, the weighted log odds method is used to compare relative frequencies of two events. For the sake of example, let's consider the frequency of occurrence of the personal pronoun before the verbs *bùgɔ* and *gòsi*, respectively. The values of these frequencies can be found in the rows for *pers:-1* collocate in table *1*. To decide which verb personal pronouns co-occur with more often, the overall frequencies of the verbs should be taken into account. This can be done by transforming frequencies into odds, that is the ratio of the number of cases when there is a pronoun in that position to the number of cases when there is something else. This gives us 76:(187-76)=0.68 for *bùgɔ*, and 71:(285-71)=0.33 for *gòsi*. By taking the ratio of these two values (the odds ratio), we immediately find that personal pronouns are approximately two times more likely to occur before *bùgɔ* than before *gòsi*. It is conventional to take the logarithm of the odds ratio (log-odds) to make the measurement symmetrical with respect to the order of values. In the example above, if we were to divide odds for *gòsi* by odds for *bùgɔ*, the result would be close to 1/2. But the logarithm of 2 is 0.69 while the logarithm of 1/2 is –0.69, which reflects the fact that the magnitude of the difference is the same in both cases, and the sign indicates whether the feature in question is preferred or avoided by the verb that is on top of the ratio. The important intuition behind the *weighted* log odds is that for rare events we may observe the frequencies 2 and 1 that produce the same ratio, but this observation is much less reliable compared to the case of observed frequencies of, for instance, 100 and 50. The magnitude and even the direction of difference in the former case is more likely to be due to sampling error. Hence the method includes a correction term in the formula that puts more weight on those frequency differences that are supported by more evidence (examples). The values of the weighted log odds for *gòsi* vs. *bùgɔ* are

shown in the last column of table 2. Positive values indicate the prevalence of the collocate with *gòsi*, and negative values correspond to higher co-occurrence rate with *bùgɔ*.

Table 2. Collocates for the two preceding positions for gòsi and bùgɔ, ordered by weighted log-odds

collocate	**ngram_bùgɔ_v**	**ngram_gòsi_v**	**log_odds_gòsi_v**
TOP:-1	2	67	3.69
n.prop:-1	13	74	2.87
PFV.TR:-2	28	84	2.03
n:-1	38	91	1.59
v:-2	4	20	1.40
prn:-1	25	9	-2.16
RECP:-1	13	2	-2.00
NOM.F:-1	10	1	-1.87
pers:-1	76	71	-1.48
PFV.NEG:-2	11	4	-1.43

Positive log-odds indicate prevalence of a collocate with gòsi, negative — with bùgɔ. Only collocates with overall frequency of 10 or more are included in the list.

Table *2* shows a list of the positional collocates in the two preceding positions for both verbs, ranked by the magnitude of the frequency difference as evaluated by weighted log-odds.[6] These data support the observation that pronouns preferentially occur in the position of the direct object with *bùgɔ*. The list also demonstrates that most of the proper nouns that fill the position of the direct object for *gòsi* are toponyms.

The above example demonstrates that positional skipgrams may serve as a tool to focus the attention and guide the analysis of differences in lexical usage, though they cannot directly show what the difference is. In particular, it helps to construct specific hypotheses in terms of the positional collocates. The tentative hypotheses built using positional skipgrams may be further explored with a classical concordance or more sophisticated statistical modeling.

[6] The computation was performed using the tidylo R package (Schnoebelen & Silge 2019).

Subcorpora comparison

Analysis of positional skipgrams need not be limited to the individual lexical units. The n-gram frequency easily lends itself to aggregation by any relevant metatextual properties. As a result, it is easy to obtain a frequency list of positional skipgrams for a subcorpus of texts that are comparable in some respect. In effect, this method allows for comparison of positional distributions of lexical items and grammatical tags across genres, time periods, regions, etc.

The idea that a frequency list of n-grams for a certain corpus characterizes the language variety used in the texts is not new. In the literature on natural language processing and on stylometry it is known as *n-gram profile*. N-gram profiles can be used to formally distinguish between different language varieties, provided that corresponding textual corpora are available to collect n-grams. It was successfully applied, for instance, in tasks to detect language by script (with character n-grams) (Cavnar, Trenkle & others 1994), and in authorship attribution (Koppel & Schler 2003).

In the following example, two subsets of the Bambara corpus are contrasted using n-gram profiles built from positional skipgrams: folkloric texts versus news articles. These two broad genres can be reasonably expected to differ in many respects of language use, some of which should clearly manifest itself in the prevailing syntactic patterns as well as in frequency distributions of part of speech tags, grammatical categories, and lexical items. The point is not to use positional skipgrams in a statistical classification setting (predictive modeling), but to employ them as a guide in the search for linguistically meaningful contrasts in language use.

The disambiguated part of the Bambara corpus contains 148 files classified as folklore (0.25M words in total), and 834 files of news articles (0.36M words). The n-gram profiles for the two subcorpora were built using the file-level positional-skipgrams data and the metadata table. Even a quick inspection of the top-frequency skipgrams lists for the two genres shows an appreciable difference in the syntactic patterns of the two subcorpora. The folkloric subcorpus has the first person singular pronoun *à* as the most frequent feature and the top 10 is dominated by n-grams involving verbs and personal pronouns. Contrariwise, all top 10 n-grams for news include a noun, and most of them consist of two nouns in some positional relationship. The third person singular occurs only on the 13th line. This clearly attests to the higher frequency of nouns and longer noun groups. When the weighted log-odds test discussed in the previous section is applied to the folklore/news dichotomy, it confirms that the syntactic differences in the narrative and reported speech versus noun groups is the most prominent contrast (see table *3*).

Table 3. Skipgrams most characteristic of folklore and news subcorpora, ordered by weighted log-odds

skipgram	**log_odds_folk**	**log_odds_news**	**f_folk**	**f_news**
pers – pers:-2	30.58	–30.58	7046	3716
pers – pm:1	30.11	–30.11	10572	7216
kó_cop – pers:-1	28.69	–28.69	2640	463
pers – v:3	28.23	–28.23	7694	4752
3SG – QUOT:1	27.19	–27.19	2156	286
kó_cop – 3SG:-1	27.18	–27.18	2155	286
n – n:1	–28.66	28.66	5435	17621
n.prop – n.prop:1	–25.79	25.79	608	5200
n – num:1	–25.56	25.56	1780	8243
n.prop – n.prop:2	–24.17	24.17	339	3978
n – n:4	–23.66	23.66	7075	18918

Table 4. Skipgrams that include nominalization, ordered by the weighted log-odds difference between folklore and news. Items with overall frequency less than 10 are omitted

skipgram	**log_odds_folk**	**log_odds_news**	**f_folk**	**f_news**
sɔ̀sɔli_n – NMLZ:0	2.48	–2.48	19	3
dún_n – NMLZ:0	2.03	–2.03	141	139
nà_v – NMLZ:1	1.73	–1.73	13	4
kòlijí_n – NMLZ:0	1.53	–1.53	10	3
IPFV.NEG – NMLZ:1	1.47	–1.47	20	12
PL – NMLZ:1	–11.83	11.83	18	741
NMLZ – PP:1	–7.83	7.83	36	419
yé_pp – NMLZ:-1	–7.83	7.83	36	419
lá_pp – NMLZ:-1	–7.53	7.53	81	526
NMLZ – ADR:1	–7.33	7.33	27	353
ni_conj – NMLZ:2	–6.74	6.74	3	230

The same data and method may be used to explore subtler differences between these subcorpora, and to test more specific hypotheses about their differences. As an example, nominalized forms can be taken, since they are expected to be much more prominent in news. To get an overview of the differences between folklore and news in respect to nominalizations, it suffices to filter the skipgrams list to get the lines containing a reference to the nominalization marker (standard gloss — *NMLZ*). The differences here are not so pronounced, but they do exist (see table *4*).

6. Discussion

N-grams are among the earliest and most widely used methods in statistical language processing. Despite the criticism by Chomsky (Chomsky 1956) who showed that n-grams (along with other finite-state models) cannot fully model the syntax of English due to their inability to represent long-distance dependencies and parenthetical constructions, the approach thrived in practical applications[7]. Statistics on n-grams of adjacent letters and phomenes proved useful for optical character recognition and speech recognition as early as the 1970s (Robertson & Willett 1998). When textual data became abundant in the 1990s, using n-grams of words turned into a well-established technique in applied language modeling tasks, such as part of speech tagging (Brants 2000), or for more linguistically oriented tasks like collocation extraction (Manning & Schütze 1999).

The computational and conceptual simplicity of the "default" bi- and tri-grams of adjacent words favored their usage wherever the performance of the resulting model was acceptable. Yet it is clear that related words are not necessarily positioned next to each other. As computational power and storage capacity grew, the idea of using word *skipgrams* gained traction as a way to capture long-distance dependencies (Guthrie et al. 2006). *Concgrams*, suggested in (Cheng, Greaves & Warren 2006), relaxed constraints not only on the distance between collocates, but also on their relative order, thereby going even further to alleviate the problem of variability in surface structures. These generalizations of n-grams clearly widen the scope of syntactic phenomena that n-grams are able to represent, but simultaneously introduce much more noise in the frequency data.

The "noise" here means unrelated or indirectly related words appearing together in the n-gram, while n-grams carrying information on meaningful regularities would constitute a useful "signal". In the example (1) above 'You should know that a greedy woman won't cook a good sauce', the skipgram *mùso – tóbi* ('woman' – 'to cook')

[7] See Church (2011) for a discussion of the n-gram based language models in a wider context of rationalist/empiricist debate in computational linguistics.

reflects a very relevant subject – verb relation, while the structurally identical skipgram *mùso – dɔ́n* (‘woman’ – ‘to know’) conveys only a much less direct link between the verb of the main clause and the subject of the embedded clause. The problem is that the two skipgrams and the co-occurrence events they represent get the same weight (the latter skipgram will even get more weight if we take distance between words in text into account). Hence the problem of noise is a direct consequence of the simplistic way of treating context relationships that reduces any syntactic structure to plain word sequence.

Two opposite ways to maintain a decent signal-to-noise ratio in n-gram data can be attested in the recent literature. One way is to collect ever more data to let the noisy co-occurrences be dwarfed by the relevant ones. This became a trend after the advent of neural networks in language modeling that followed the success of word2vec (Mikolov et al. 2013). The other approach is to reduce noise sources in terms of the surface structure by adding more linguistic structure to the input data. The idea of building *syntactic n-grams* based on relations in the syntactic tree rather than the word sequence is characteristic of this latter position (Sidorov et al. 2014). The downside of the first method is that it requires large amounts of textual data to be available for training. The obvious drawback of the second is that a reliable syntactic parser is required for it to work. Both of these are serious, if not blocking, limitations in the context of low-resourced languages.

The *positional skipgrams* method suggested in this paper sits somewhere in between the above approaches in terms of balancing signal and noise in the n-gram data. Contrary to the word2vec approach, positional skipgrams do require linguistically annotated data for the input, but the annotation can be rather shallow, like part of speech tags in the examples above. By using tags and their relative positions, the skipgrams are able to capture the signal on syntactic regularities from the tags. Part-of-speech tags and other morphological annotation, in their turn, accumulate information from the dictionary that relates wordforms to lemmas and grammatical categories for many infrequent words for which textual examples in a corpus would be insufficient to infer categories with statistical models. Another source of information contained in the tags is language knowledge added by human annotators in case if annotation was checked manually. That is exactly the kind of signal for which sparse lexical data would be insufficient in the absence of the huge training corpora. At the same time, given the current state of the art in part of speech tagging, it seems reasonable to assume that such annotation can be obtained for significant amounts of text even for lower-resourced languages. Collecting more shallow-annotated data of this kind can compensate for the noisy way of capturing morphosyntactic structures offered by n-grams. Positional skipgrams are not an exception in this respect. But given the higher

frequency and lower diversity of the part of speech tags, the signal can be expected to overcome noise much sooner compared to training on raw words.

To recapitulate, the positional skipgrams were introduced here for the case of a low-resourced language in order to alleviate the problem of data sparsity. Word-level skipgrams deliberately throw out information encoded in the exact positioning of words in the data, regarding it as noise. This may indeed work when there is a large amount of data. But when the data is relatively scarce, sampling errors with sparse lexical items is a very serious concern. Traditionally, co-occurrence analysis treats lexicon as signal and syntax as noise, whereas in this work I suggest to switch sides. By tracking positional co-occurrences with higher-level grammatical categories with positional skipgrams, patterns of the local surface structures emerge that stem from the information obtained using dictionaries and human language competence. This can provide a statistical signal that is more reliable and broad than plain lexical co-occurrence.

7. Conclusion

This article presented a new dataset built using a morphologically-annotated and manually disambiguated subcorpus of the Bambara Reference Corpus, and demonstrated several ways in which this dataset may help in formulating linguistic hypotheses about various contrasts present in textual data. This quantitative data (with metadata) enables testing hypotheses and building models based on lexico-grammatical distributions in various parts of the corpus. The goal of publication of this dataset is to provide a wider audience with access to the data on linguistic regularities observed in the Bambara corpus for linguistic research and development of NLP applications.

The data is organized in the form of frequency lists of *positional skipgrams*, which is a framework for building n-grams suggested in this article. These n-grams capture information about co-occurrence of lexical items with grammatical categories at various relative positions. Researchers are free to download the data and to use it both as an aid for linguistic queries for the corpus, and as a basis for building applications for the natural language processing tasks.

The approach to create an n-gram corpus suggested in this article is suited well to less-resourced settings where overall textual data is not easily available but some annotated texts are present. For linguists, positional skipgrams may serve as an exploratory tool which, much like the concordance, reorganizes the textual data in a non-linear fashion in order to reveal regularities. This is intended not to replace, but to supplement other views of the corpus, including online search and concordancing.

References

Brants, Thorsten. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, 224–231. Association for Computational Linguistics.

Brants, Thorsten & Alex Franz. 2006. The Google Web 1T 5-gram corpus version 1.1. *LDC2006T13*.

Brown, Peter F, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra & Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18(4). 467–479.

Cavnar, William B, John M Trenkle & others. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, 161–175.

Cheng, Winnie, Chris Greaves & Martin Warren. 2006. From n-gram to skipgram to concgram. *International journal of corpus linguistics* 11(4). 411–433.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on information theory* 2(3). 113–124.

Church, Kenneth. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology* 6(5). 1–27.

Davydov, Artem. 2010. Towards the Manding corpus: Texts selection principles and metatext markup. In *Proceedings of the Second Workshop on African Language Technology AfLaT*, 59–62.

Gries, Stefan Th & Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. *New directions in cognitive linguistics* 57–75.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie & Yorick Wilks. 2006. A Closer Look at Skip-gram Modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/357_pdf.pdf.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1(1). 75–98.

Jaggi, Martin, Fatih Uzdilli & Mark Cieliebak. 2014. Swiss-chocolate: Sentiment detection using sparse SVMs and part-of-speech n-grams. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 601–604.

Kopotev, Mikhail, Lidia Pivovarova, Natalia Kochetkova & Roman Yangarber. 2013. Automatic Detection of Stable Grammatical Features in N-Grams. In *NAACL HLT 2013: Proceedings of the 9th Workshop on Multiword Expressions*, 73–81.

Atlanta, Georgia, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-1011.

Koppel, Moshe & Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, vol. 69, 72–80.

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225. https://doi.org/10.1162/tacl_a_00134.

Ling, Wang, Chris Dyer, Alan W. Black & Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1142. https://www.aclweb.org/anthology/N15-1142.

Manning, Christopher D & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

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.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Monroe, Burt L., Michael P. Colaresi & Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16(4). 372–403. https://doi.org/10.1093/pan/mpn018.

Plecháč, Petr. 2019. *Atribuce autorství básnických textů [Authorship attribution of poetic texts]*. Praha: Univerzita Karlova phd.

Robertson, Alexander M & Peter Willett. 1998. Applications of n-grams in textual information systems. *Journal of Documentation* 54(1). 48–67.

Rosenfeld, Ronald. 2000. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE* 88(8). 1270–1278.

Schnoebelen, Tyler & Julia Silge. 2019. *tidylo: Tidy Log Odds Ratio Weighted by Uninformative Prior*. http://github.com/juliasilge/tidylo.

Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh & Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning

features for natural language processing. *Expert Systems with Applications* 41(3). 853–860.

Stefanowitsch, Anatol & Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics* 8(2). 209–243.

Taylor, Paul & Alan W Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech & Language* 12(2). 99–117.

Vydrin, Valentin. 2013. Bamana Reference Corpus (BRC). In Chelo Vargas-Sierra (ed.), *Procedia - Social and Behavioral Sciences: Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013)*, vol. 95, 75–80. Alicante: Elsevier. http://www.sciencedirect.com/science/journal/18770428.

Kirill Maslinsky

Positional skipgrams for Bambara: a resource for corpus-based studies

This article presents a new online dataset of linguistically rich n-gram frequency data for Bambara based on the disambiguated part of the Bambara Reference Corpus. The n-grams in the dataset are *positional skipgrams* that capture information about co-occurrence of lexical items with grammatical categories at various relative positions. These n-grams were constructed with the aim to leverage those types of information that are available in the morphologically annotated corpus of Bambara given the limited amount of textual data. The methodology and data used for constructing n-grams for Bambara are discussed, followed by brief illustrations of how the positional skipgrams data may be employed in corpus-based linguistic research.

Keywords: Bambara, corpus, n-grams, shared data

Les *skipgrams* positionnels pour le bambara : une ressource pour la recherche linguistique orientée corpus

L'article présente un nouveau paquet de données linguistiques de fréquences de n-grams pour le bambara, basé sur le sous-corpus désambiguïsé du Corpus Bambara de Référence. Les n-grams sont de *skipgrams positionnels* qui capturent l'information sur la co-occurrence des lexèmes avec des catégories grammaticales à des positions différentes. Ces n-grams ont été conçus pour tirer profit de ce type d'informations disponibles dans le corpus bambara morphologiquement annoté, vu le volume limité des données textuelles. La discussion de la méthodologie et les données utilisées pour la construction des n-grams pour le bambara est suivie par quelques illustrations

d'utilisation des skipgrams positionnels dans des recherches linguistiques basées sur un corpus.

Key words : bambara, corpus, n-gram, données partagées

Kirill Maslinsky

Les *skipgrams* positionnels pour le bambara : une ressource pour la recherche linguistique orientée corpus

L'article présente un nouveau paquet de données linguistiques de fréquences de n-grams pour le bambara, basé sur le sous-corpus désambiguïsé du Corpus bambara de référence. Les n-grams sont de *skipgrams positionnels* qui capturent l'information sur la co-occurrence des lexèmes avec des catégories grammaticales à des positions différentes. Ces n-grams ont été conçus pour tirer profit de ce type d'informations disponibles dans le corpus bambara morphologiquement annoté, vu le volume limité des données textuelles. La discussion de la méthodologie et les données utilisées pour la construction des n-grams pour le bambara est suivie par quelques illustrations d'utilisation des skipgrams positionnels dans des recherches linguistiques basées sur un corpus.

Mots clé : bambara, corpus, n-gram, données partagées

Кирилл Александрович Маслинский

Позиционные скип-граммы для бамана: ресурс для корпусных исследований

В статье представлен новый доступный онлайн набор данных: корпус н-грамм слов на основе подкорпуса со снятой омонимией Справочного корпуса бамана. В наборе данных представлены частотные списки позиционных скип-грамм, в которых отражена информация о совместной встречаемости лексем с грамматическими категориями на различных относительных позициях в тексте. Данный тип н-грамм разработан для того, чтобы более полно отразить лингвистическую информацию, содержащуюся в морфологически аннотированном корпусе бамана. В статье обсуждается методология подготовки корпуса н-грамм для бамана и представлено несколько кратких иллюстративных примеров использования данных о частотности позиционных скип-грамм в корпусных лингвистических исследованиях.

Ключевые слова: Бамана, корпус, н-граммы, открытые данные