Adaptor Grammars for the Linguist: Word Segmentation Experiments for Very Low-Resource Languages

Pierre Godard[*], **Laurent Besacier**[†], **François Yvon**[*], **Martine Adda-Decker**[⋆],
Gilles Adda[*], **Hélène Maynard**[*], **Annie Rialland**[⋆]
[*]LIMSI, CNRS, Université Paris-Saclay / Orsay, France
[†]LIG, UGA, G-INP, CNRS, INRIA / Grenoble, France
[⋆]LPP, CNRS / Paris, France
[*]{godard,yvon,gadda,hbm}@limsi.fr
[†]laurent.besacier@univ-grenoble-alpes.fr
[⋆]{martine.adda-decker,annie.rialland}@univ-paris3.fr

Abstract

Computational Language Documentation attempts to make the most recent research in speech and language technologies available to linguists working on language preservation and documentation. In this paper, we pursue two main goals along these lines. The first is to improve upon a strong baseline for the unsupervised word discovery task on two very low-resource Bantu languages, taking advantage of the expertise of linguists on these particular languages. The second consists in exploring the Adaptor Grammar framework as a decision and prediction tool for linguists studying a new language. We experiment 162 grammar configurations for each language and show that using Adaptor Grammars for word segmentation enables us to test hypotheses about a language. Specializing a generic grammar with language specific knowledge leads to great improvements for the word discovery task, ultimately achieving a leap of about 30% token F-score from the results of a strong baseline.

1 Introduction

A large number of the world's languages are expected to go extinct during this century – as much as half of them according to Crystal (2002) and Janson (2003). Such predictions have subsequently fostered a growing interest for a new field, Computational Language Documentation (CLD), as it is now clear that traditional field linguistics alone will not meet the challenge of preserving and documenting all of these languages.

CLD attempts to make the most recent research in speech and language technologies available to linguists working on language preservation and documentation (e.g. (Anastasopoulos and Chiang, 2017; Adams et al., 2017)). A remarkable effort in this direction has improved the data collection tools to be used on the field (Bird et al., 2014; Blachon et al., 2016), enabling to collect corpora for several endangered languages (Adda et al., 2016). In parallel, the language technology community is investing more efforts to design methodologies tailored for the new challenges posed by the analysis of such linguistic material: the extreme variability of the orthographic representation, the scarcity of annotated data (both written and oral), as well as the modeling of complex tonal systems.

This effort could greatly benefit from a tighter collaboration between the two main research communities involved in this endeavor, which often struggle to cooperate efficiently. Knowledge background differs between linguists and computer scientists; the definition of why a problem is interesting or not may not be the same for the two communities, theoretical and experimental platforms do not intersect much, etc. Consequently, for lack of investing enough energy working on the same problems with the same tools and towards the same goals, we might not achieve the efficiency that is needed, as time is running out for many languages. This view constitutes the underlying motivation of the work reported here.

We pursue two main goals in this spirit. The first one is to improve upon a strong baseline (Goldwater et al., 2009) for the unsupervised word discovery task[1] on two low-resource languages, by teaming up with linguist experts. A natural idea to achieve this goal is to engage them in formalizing their linguistic knowledge regarding the languages or language families under study, in the hope that it will compensate for the small amount of available data. In our case, this expertise corresponds to morphological and phonotactic constraints for

[1] We indifferently use the terms *word discovery* and *word segmentation* to denote the task defined in Section 2.2.

Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 32–42
Brussels, Belgium, October 31, 2018. ©2018 The Special Interest Group on Computational Morphology and Phonology
https://doi.org/10.18653/v1/P17

two Bantu languages displaying very similar structures (see Section 3). For one language, we were also able to elicit a list of prefixes and some additional knowledge regarding the consonantal system. Such expert knowledge can readily be integrated in grammar rules using the framework of Adaptor Grammars (see Section 6). Another interesting property of this framework is its compatibility with two strategies that are usually thought as being mutually exclusive: rule-based learning, still in wide use inside the linguistics community, and statistical learning, prevalent in natural language processing circles.

Our second goal is to study ways to help linguists explore language data when little expert knowledge is available. Our proposal is to complement the grammatical description activity with task-oriented search procedures, that will speed up the exploration of competing hypotheses. The intuition is that better grammars should not only truthfully match the empirical data, but also improve the quality of automatic analysis processes. The word discovery task considered below should thus be viewed as an extrinsic validation procedure, rather than a goal in and of itself. This process might also yield new linguistic insights regarding the language(s) under focus.

To sum up, the main contribution of this paper is a methodology for systematically exploring (a subpart of) the space of possible grammars, refining grammar rules (from the most generic to the most language specific) at four levels of description (see Section 4). This results in a comparison of 162 alternative accounts of the grammar for two languages. Our results (analyzed in Section 5) show that enriching grammar rules with language specific knowledge has a consistent positive impact in performance for the segmentation task. They validate our hypotheses that (a) improved grammatical descriptions actually correlate with better automatic analysis; (b) Adaptor Grammars provide a framework around which linguists and computer scientists can effectively collaborate, with tangible results for both communities.

2 Adaptor Grammars for Word Discovery

2.1 Adaptor Grammars

Formal grammars, and notably Context-Free Grammars (CFGs), are a cornerstone of linguistic description and provide a model for the structural description of linguistic objects. Our grammars capture simple aspects of the syntax and some less trivial aspects of the morphological and phonological structures. As discussed below, both levels of descriptions are useful for word discovery.

A CFG is a 4-tuple $G = (N, W, R, S)$ where N and W are respectively the non-terminal and terminal symbols, R a finite set of rules of the form $A \rightarrow \beta$, with $A \in N$ and $\beta \in (N \cup W)^*$, and $S \in N$ the start symbol. Our grammars will be used to analyze the structure of complete utterances and the start symbol S will always correspond to the sentence top-level. Assuming that S, Words, and Word belong to N, the top level rules will typically look like: $S \rightarrow \mathrm{Words}; \mathrm{Words} \rightarrow \mathrm{Word}\,\mathrm{Words}; \mathrm{Words} \rightarrow \mathrm{Word}$, the last two rules abbreviated as $\mathrm{Words} \rightarrow \mathrm{Word}+$.

Probabilistic CFGs (PCFGs) (Johnson, 1998) extend this model by associating each rule with a scalar value $\theta_{A\rightarrow\beta}$, such that for each $A \in N$, $\sum_\beta \theta_{A\rightarrow\beta} = 1$. Under some technical conditions (Chi, 1999), PCFGs define probability distributions over the set of parse trees, where the probability of a tree is a product of the probability of the rules it contains. PCFGs can be learned in a supervised way from treebanks or in a unsupervised manner using, for instance, the EM algorithm (Lari and Young, 1990).

PCFGs make unrealistic independence assumptions between the different subparts of a tree, an observation that has yielded many subsequent variations and extensions. Adaptor grammars (AGs) (Johnson et al., 2007) define a powerful mechanism to manipulate PCFG distributions to better match the occurrences of trees and subtrees observed in actual corpora. Informally, an AG is a CFG where non-terminals have the possibility to be *adapted*: when non-terminal A is adapted, all subtrees rooted in A are “reified”, meaning that they are no-longer viewed only as decomposable objects, but can also be manipulated and stored as a whole. In our grammars below, adapted non-terminals are underlined, and optional non-terminals appear between brackets. Following (Johnson et al., 2007), we only adapt non-recursive non-terminals.[2]

AGs define a framework to implement Bayesian nonparametric learning of grammars, and are usually trained in an unsupervised manner using sampling techniques (Monte-Carlo Markov Chain,

[2] A non-terminal A is recursive if R contains a rule where A appears both in the left and right-hand sides.

MCMC). A typical run will produce, for each sentence, a distribution of possible parses under the grammar, from which we can then retain the most frequent one as the "best" possible analysis.[3]

2.2 Word Segmentation using AGs

In this work, we are interested in the word segmentation task: from an unsegmented stream of symbols, the system must output delimited sequences corresponding to actual words in the language. For this, we assume a linguistic grammar G, which parses sequences of letters (or phones) as being organized into Words, which themselves recursively decompose into smaller units such as Morphs, $\mathrm{Syllables}$, etc. To induce word segmentation from parse trees, we will consider that each span covered by the non-terminal symbol Word defines a linguistic word, even though in a fully unsupervised setting, this non-terminal might actually correspond to larger or smaller linguistic units. Figure 2 illustrates this on two example parses.

Likewise, when examining the output of the training process, we are in a position to collect sets of word types (or morph types, syllable types, etc.) and will do so based only on the identity of the root symbol, i.e. without any certainty regarding the linguistic status of the collected sequences.

3 Linguistic material

3.1 Mboshi and Myene

We experiment with two Northwestern Bantu Languages: Mboshi (Bantu C25), a language spoken in Congo-Brazzaville, and Myene (B10, Gabon), a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu) spoken at the coastal areas and around the town of Lambarene in Gabon.[4] Unlike southern Bantu relatives such as Swahili, Sotho or Zulu, Mboshi and Myene are scarcely studied, protected, and resourced. We briefly describe the main aspects related to phonetics, phonology, morphology, and tonology of these languages.

Phonetics and phonology. Mboshi and Myene both have a seven vowel system (i, e, ɛ, a, ɔ, o, u). Mboshi has an opposition between long and short vowels, which does not exist in Myene. Mboshi consonantal system includes the following phonemes: p, t, k, b, d, β, l, r, m, n, ɲ, mb, nd, ndz, ng, mbv, f, s, ɣ, pf, bv, ts, dz, w, j. It has a set of prenasalized consonants (mb, nd, ndz, ng, mbv) which is common in Bantu languages (Embanga Aborobongui, 2013; Kouarata, 2014). Myene includes the following phonemes: p, t, k, b, d, β, l, r, m, n, f, s, g, y, v, ŋ, w, z – many of them with variants of realization. Prenasalized consonants exist also in Myene (Ambouroue, 2007).

While both languages can be considered as rarely written, linguists have nonetheless defined a non-standard graphemic form for them, considered to be close to the language phonology. Affricates and prenasalized plosives are coded using multiple symbols (e.g. two symbols for dz, three for mbv). For Mboshi, long and short vowels are coded respectively as V and as VV. In Myene, the transcription of the corpus involves not only the phoneme set, but also the main variants (ɲ, tʃ, dz) and some marginal sounds found in loanwords.

Both languages display a complex set of phonological rules. The deletion of a vowel before another vowel in particular, common in Bantu languages, occurs at 40% of word junctions in Mboshi (Rialland et al., 2015). This tends to obscure word segmentation and introduces an additional challenge for automatic processing.

Morphology. Words are composed of roots and affixes, and almost always include at least one prefix, while the presence of several prefixes and one suffix is also very common. The suffix structure mostly consists of a single vowel V (e.g. -a or -i) whereas the prefix structure may be both CV or V (or CVV in Mboshi). The most common syllable structures are V and CV in both languages. CVC also occurs in Myene, and CVV in Mboshi.[5]

The noun class prefix system is another feature typical of Bantu languages. For both languages, the structure of the verbs, also common in Bantu languages, is as follows: Subject Marker — Tense/Mood Marker — Root-derivative Extensions — Final Vowel. A verb can be very short or quite long, depending on the markers involved.

Tonology. Prosodic systems for both Mboshi and Myene involve tones, but the transcribed data used for this work do not encode tone markers. Experiments to assess the usability of tonal information for word segmentation were conducted in (Godard et al., 2018b).

[3] In practice, we will retain the most frequent segmentation rather than the most frequent parse (see Section 2.2).

[4] Our Myene data correspond to the Orungu variant.

[5] CCV may also arise due to the presence of affricates and prenasalized plosives mentioned in this section.

language	#sent	#tokens	#types	avg. token length
Mboshi	5130	30,556	5,312	4.19
Myene	4,579	18,047	4,190	4.72

Table 1: Corpora Statistics

3.2 Corpora for Mboshi and Myene

Corpora for Mboshi and Myene were collected following a real language documentation scenario, using a mobile app dedicated to fieldwork language documentation (Blachon et al., 2016). These corpora contain manual transcriptions in the form of a non-standard graphemic form close to the languages' phonology. The correct word segmentations for these transcripts were also annotated by linguists. Basic statistics are in Table 1. The Mboshi corpus is more comprehensively described in (Godard et al., 2018a).[6]

4 Grammars

4.1 Structuring Grammar Sets

Our starting point is the set of grammars used in (Johnson and Goldwater, 2009) and (Eskander et al., 2016) which we progressively specialize through an iterative refinement process involving both field linguists and computer scientists. As we wish to evaluate specific linguistic hypotheses, the initial space of interesting grammars has been generalized in a modular, systematic, and hierarchical way as follows. We distinguish four sections in each grammar: sentence, word, syllable, character. For each section, we test multiple hypotheses, gradually incorporating more linguistic structure. Every hypothesis inside a given section can be combined with every hypothesis of any other section,[7] thereby allowing us to explore a large quantity of grammars and to analyze the contribution of each particular hypothesis.

4.2 The Full Grammar Landscape

All the grammar sections (sentence, word, syllable, character) experimented in this paper are detailed in Figure 1. We describe below the way each section was designed.

- sentence level: we model 3 different hierarchies of words. We introduce first the `flat` variety with two rules generating right-branching parse trees. `colloc` adds a single level of word collocation, aimed to capture recurrent local word associations (such as frequent bigrams); `colloc3` displays a deeper hierarchical structure with three levels of collocations. Exploring more realistic syntactic structures is left for future work.

- word level: here we propose 6 competing hypotheses. `flat` is similar to previous sentence variety but at the word level instead of the sentence level. `generic` corresponds to a more structured version of `flat`, as the specification of a sequence of 5 adapted morphemes allows, in principle, the Adaptor Grammar to learn some morphotactics. `bantu` defines a generic morphology for Bantu languages. `basaa` implements the morphology of a well-studied Bantu language, Basaa (A43 (Hamlaoui and Makasso, 2015)). `mboshi/myene` corresponds to a somewhat crude morphology of Mboshi, also applicable to Myene. Last `mboshi/myene_NV` refines `mboshi/myene` with a specification of the morphology of nouns and verbs. Additionally, for `basaa`, `mboshi/myene` and `mboshi/myene_NV` which introduce a notion of prefix, we also test a variant (called respectively `basaa+`, `mboshi/myene+` and `mboshi/myene_NV+`) containing an explicit list of prefixes in Mboshi.

- syllable level: we contrast 3 hypotheses : `flat` is similar to previous sentence and word varieties but at the syllable level, defining the syllable as a mere sequence of characters. `generic/basaa` is a generic set of rules modeling phonotactics applicable to a wide scope of languages (including Basaa mentioned in the preceding level). `bantu/mboshi/myene` displays a set of rules more specific to Mboshi and Myene.[8]

- character level: rules in the `chars` set simply rewrite the characters (terminals) ob-

[6] This dataset has already been used in several studies targeting endangered languages and is available at `http://www.islrn.org/resources/747-055-093-447-8/`.

[7] Note that if a non-terminal is absent from a hypothesis (e.g. Syllable in a word level hypothesis), the corresponding non-terminal in the subsequent hypotheses (e.g. at the syllable level) will be ignored.

[8] In theory, we should not include a coda in this last hypothesis, but loanwords and proper names in our data made the Adaptor Grammar fail to parse without a coda. To decrease the impact of this rule, we chose not to adapt the corresponding non-terminal, in contrast to `generic/basaa`.

Sentence level (A)

Words → Word+

flat(A1)

Collocs → Colloc+
Colloc → Words
Words → Word+

colloc(A2)

Colloc3s → Colloc3+
Colloc3 → Colloc2s
Colloc2s → Colloc2+
Colloc2 → Collocs
Collocs → Colloc+
Colloc → Words
Words → Word+

colloc3(A3)

Word level (B)

Word → Morphs
Morphs → Morph+
Morph → Chars

flat(B1)

Word → M1 (M2 (M3 (M4 (M5))))
M1 → Chars
M2 → Chars
M3 → Chars
M4 → Chars
M5 → Chars

generic(B2)

Word → (Prefixes) Stem (Suffixes)
Prefixes → Chars
Stem → Chars
Suffixes → Chars

bantu(B3)

Word → (Prefix) Stem (Suffix)
Prefix → Syllable
Suffix → Syllable
Stem → Syllable
Stem → Syllable Syllable

basaa(B4)

Word → (Prefix1 (Prefix2)) Stem (Suffix)
Prefix1 → Syllable
Prefix2 → Syllable
Suffix → Syllable
Stem → Syllable (Syllable)

mboshi/myene(B5)

Word → Noun
Word → Verb
Word → Chars
Noun → (PrefixNoun) Stem (Suffix)
Verb → (Prefix1 (Prefix2)) Stem
PrefixNoun → Syllable
Prefix1 → Syllable
Prefix2 → Syllable
Suffix → Syllable
Stem → Syllable (Syllable (Syllable)

mboshi/myene_NV(B6)

Syllable level (C)

Syllable → Chars
Chars → Char+

flat(C1)

Syllable → (Onset) Rhyme
Rhyme → Nucleus (Coda)
Onset → Consonants
Nucleus → Vowels
Coda → Consonants
Consonants → Consonant+
Vowels → Vowel+
Chars → Char+

generic/basaa(C2)

Syllable → (Onset) Rhyme
Rhyme → Nucleus (Coda)
Onset → Consonants
Nucleus → Vowel (Vowel)
Coda → Consonants
Consonants → Consonant+
Chars → Char+

bantu/mboshi/myene(C3)

Character level (D)

Char → Vowel
Char → Consonant
Vowel → u
Vowel → o
Vowel → i
Vowel → a
Vowel → e
...

chars(D1)

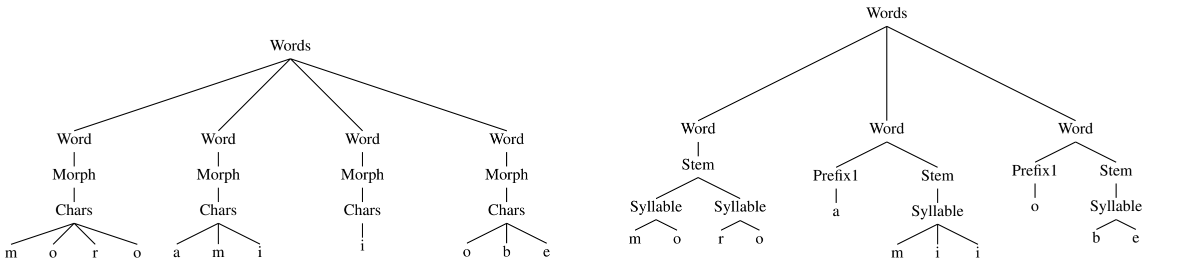
...
Consonant → m b
Consonant → n d
Consonant → n d z
...

chars+(D1+)

...
Prefix → o
Prefix → i
Prefix → e
Prefix → a
Prefix → l e
Prefix → l a
Prefix → l i i
...

{basaa,mboshi/myene,mboshi/myene_NV}+
(B{4,5,6}+)

Figure 1: Grammar rules for all the hypotheses presented in Section 4.



A3B1C1D1, *"moro ami i obe"*

A3B5C2D1+, *"moro amii obe"* (correct word segmentation)

Figure 2: Examples of parses – some non-terminals have been omitted for readability – obtained with two grammars, and the corresponding word segmentation for Mboshi sentence *"Moro a-mii o-be"*. (CL1.man 3SG-swallow.PST CL14-bad; since *Moro* is an irregular noun, the prefix and the stem are difficult to separate, which is signaled by a dot, following the Leipzig glossing rules.)

served in our data. chars+ adds rules to capture the digraphs or trigraphs occurring in Mboshi (see details in Section 3).

5 Experiments and Discussion

We now experiment along the methodology presented in Section 4. We report word segmentation performance using precision, recall, and F-measure on tokens (WP, WR, WF), and types (LP, LR, LF). We also report the exact-match (X) metric which calculates the proportion of correctly segmented utterances.[9]

In all the figures, and in this section, we use the following compact names for grammatical hypotheses at each level:

- A1 (flat), A2 (colloc), A3 (colloc3),
- B1 (flat), B2 (generic), B3 (bantu), B4 (basaa), B5 (mboshi/myene), B6 (mboshi/myene_NV), with additional "+" variants for B4, B5, and B6 when a list of prefixes is provided, for instance B6+ (mboshi/myene_NV+),
- C1 (flat), C2 (generic/basaa), C3 (bantu/mboshi/myene),
- D1 (chars), D1+ (chars+).

For each language, we evaluate our 162 grammar configurations using Mark Johnson's code,[10] collecting parses after 2,000 sampling steps.[11] We adapt all non-recursive non-terminals and use a Dirichlet prior to estimate the rule probabilities. We place a uniform Beta prior on the discount parameter of the Pitman-Yor process, and a vague Gamma prior on the concentration parameter.

Figure 3 presents token metrics (WP, WR, WF) and type metrics (LP, LR, LF), as well as sentence exact-match (X) for both corpora on all grammars.

5.1 Word Segmentation Results

Impact of sentence level variants We can see in Figure 3 that A2 and A3 hypotheses globally yield better results than A1 in both languages. For Mboshi, the benefit of A3 vs. A2 appears especially on token metrics (WP, WR, WF), but this contrast is less clear on Myene. For both languages, however, our results confirm that modeling collocation-like word groups at the sentence level is important. These word dependencies seem indeed related to a universal linguistic property.

Impact of word level variants If we now focus solely on the A3 hypothesis for Myene in Figure 3, we observe a general trend upwards for all metrics. The benefit of gradually using more language-specific grammars, from B1 to B6+, is clear. While this trend is also observed for Mboshi, the less specific B3 hypothesis yields the strongest results on token metrics (WP, WR, WF). Precision on types (LP) with B3 is also the strongest, but B6+ achieves better performance on type recall and F-measure (LR and LF). The contrast between B1 and B2 for all metrics on both languages (keeping a focus on A3, but this can also be seen for A1 and A2) highlights the benefit of modeling some morphotactics inside the word-level hypotheses, which seems to correspond to another universal linguistic property (the dependency between morphemes inside a word).

Impact of syllable level variants It is difficult to see a clear trend for the impact of syllable-level variants in Figure 3. Importantly, the syllable level will only be effective when combined with word level variants B4, B5 and B6 (and their "+" versions) which model the concept of syllable: when combined with B1, B2 or B3, each C level hypothesis will default to its "Chars $\rightarrow$ Char +" rule. Figure 4 illustrates the impact of C1, C2, and C3 by averaging type and token F-measures (LF and WF) over all grammar sections with a syllable non-terminal (B4, B4+, B5, B5+, B6, and B6+). The benefit of C2/C3 vs. C1 appears more clearly, especially on type F-measures and on Myene.[12] Nevertheless, the impact of the syllable level, and the capacity to incorporate phonotactics in our models, seems of less significance for word segmentation than choices made at the word and sentence levels.

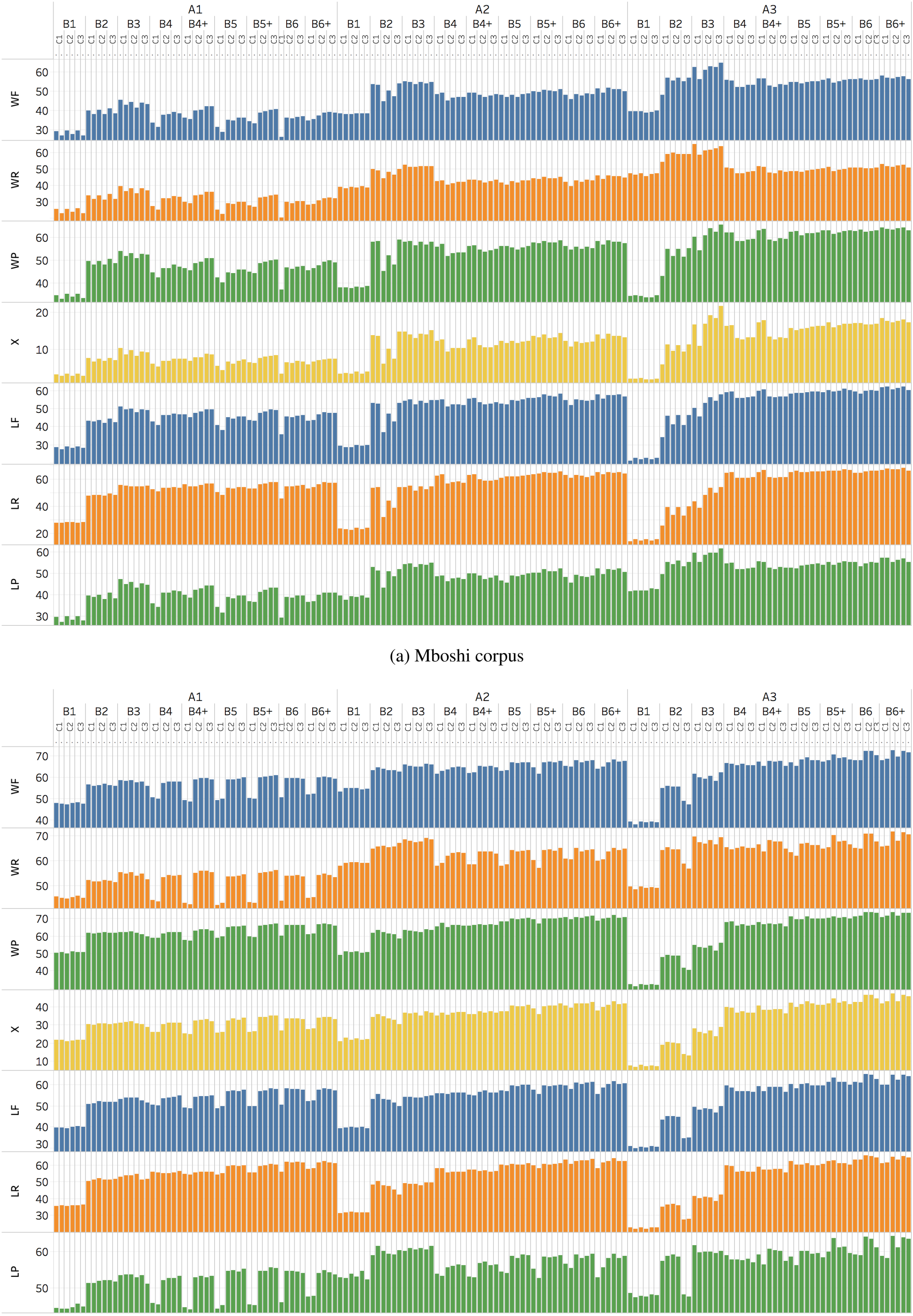
Impact of character level variants In Figure 3, it is also hard to see if there is any benefit in using D1+ over D1, i.e. adding digraphs or trigraphs

[9] The exact-match metric includes single-word utterances.

[10] http://web.science.mq.edu.au/~mjohnson/Software.htm

[11] The large number of experiments we are dealing with did not allow us to average over several runs. Stable results were obtained on a subset of grammars. Two particular configurations in Mboshi (A3-B6-C3-D1+ and A1-B6-C1-D1) did not reach 2,000 iterations within the maximum wall clock time allowed by the cluster used for these experiments (2 weeks), and are left out of the discussion.

[12] The differences between C3 and C2, two very similar hypotheses, are hardly significant.



(a) Mboshi corpus

(b) Myene corpus

Figure 3: Word segmentation performance evaluated with token metrics (WP, WR, WF), type metrics (LP, LR, LF), and sentence exact-match (X) for Mboshi (top) and Myene (bottom). All grammars are broken down by A, B, C, and D levels (D1 shown before D1+).

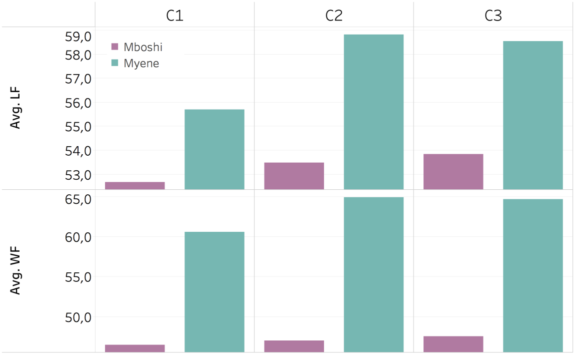


Figure 4: Impact of C variants on Mboshi and Myene. Token F-measure (WF) and type F-measure (LF) are averaged over hypotheses B4, B4+, B5, B5+, B6, and B6+.

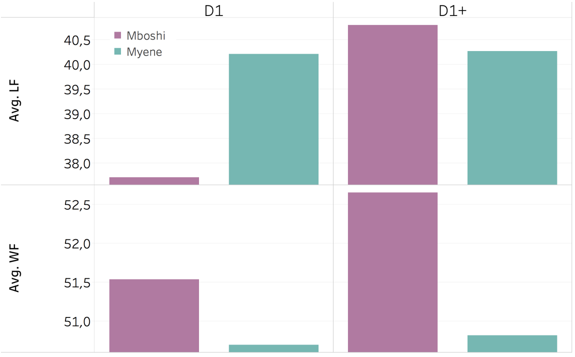


Figure 5: Impact of D variants on Mboshi and Myene. Token F-measure (WF) and type F-measure (LF) are averaged over hypotheses with A level set to A3 and B level set to B1, B2, or B3.

to the consonant inventory. Averaging over all hypotheses at the A, B, and C levels do not exhibit any clearer impact. It is likely that refined models at the syllable level (C) compensate for a less accurate consonant inventory through the adaptation of their non-terminals, and do learn some phonotactics. This would explain the weak contribution of D1+. To test this hypothesis, we set the sentence level to A3 (the best compromise for Mboshi and Myene) and the word level to B1, B2, or B3 (levels without a Syllable non-terminal, which cancels the effect of the syllable level C). The token and type F-measures averaged over the considered hypotheses are shown Figure 5. We do observe a benefit in using the D1+ character variant in Mboshi, but not in Myene. This is not surprising, as the digraph and trigraph rules added by the D1+ variant are specific to Mboshi and do not cover the inventory for Myene.

Stronger results in Myene Segmentation performance is globally superior in Myene. This can probably be explained by corpus statistics (see Table 1), as the average number of words per sentence is 3.94 in Myene, and 5.96 in Mboshi. Since sentence boundaries are also word boundaries, the proportion of already known word boundaries is higher in Myene, which makes word segmentation a harder task in Mboshi. Figure 3 also reveals an interesting contrast: token results are higher than type results in Myene, while the converse is true in Mboshi. The token/type ratio (5.75 tokens for one type in Mboshi, and 4.30 in Myene) indicates a higher lexical diversity in Myene, which might explain weaker results on types. Strong results on types for Mboshi, on the other hand, show the capacity of AGs to generalize well on low-frequency events, a property of particular interest in the low-resource scenario.

Comparison to an existing baseline Overall, our best performing grammars are A3-B3-C3-D1+ for Mboshi (64.78% token F-measure) and A3-B6+-C2-D1 for Myene (72.62% token F-measure). This result is about 30 points higher than a strong Bayesian baseline, the Dirichlet process-based bigram word segmentation system of Goldwater et al. (2006, 2009), [13] which yields 34.34% token F-score on Mboshi and 44.48% on Myene.

5.2 How Can This Help a Linguist?

Our second goal is to understand more precisely how such experiments can be useful for linguists, beyond the benefit of having access to better automatic word segmentation tools for their data.

Phonological status of complex consonants In the analysis of the results (Section 5.1 above) we showed the benefit of integrating digraphs or trigraphs in the consonants inventory for Mboshi. This result is of special interest for linguists, since it is in line with the most recent phonological analyses of Mboshi (Embanga Aborobongui, 2013; Kouarata, 2014; Amboulou, 1998) which agree in recognizing complex consonants (represented by digraphs or trigraphs) in the phonological inventory of this language. The analysis of complex consonants, in particular prenasalized consonants, generated many debates in Bantu linguistics

[13] https://homepages.inf.ed.ac.uk/sgwater/resources.html.

(Odden, 2015; Herbert, 1986; Downing, 2005). The present experiments provide more substance to support the integration of complex consonants in the phonological inventory of Mboshi.

Learning prefixes without supervision Since parses are produced to segment sentences into words, it is possible to extract the most frequent prefixes or suffixes (for B variants introducing such a concept). The precision on the 20 most frequently found prefixes for grammars without prefix-supervision (B3, B4, B5 and B6)[14] reaches 58.21% in Mboshi, and 61.21% in Myene. The capacity of AGs to learn true prefixes without supervision could thus help linguists in the process of documenting a new language. On the supervised variants (B4+, B5+, and B6+), the precision achieved in Mboshi is 61.11%, and 63.07% in Myene: the benefit of the supervision is limited; token measures for Mboshi with these variants (Figure 3) nevertheless indicate a benefit for word segmentation.

6 Related Work

AGs have been used to infer the structure of unsegmented sequences of symbols, offering a plausible modeling of language acquisition (Johnson, 2008b; Johnson and Goldwater, 2009); they have also been used for the unsupervised discovery of word structure, applied to the Sesotho language by Johnson (2008a). One notable outcome of this latter study was to demonstrate the effectiveness of having an explicit hierarchical model of word internal structure ; an observation that was one of our primary motivations for using AGs in our language documentation work. In this series of studies, AGs are shown to generalize models of unsupervised word segmentations such as the Bayesian nonparametric model of Goldwater (2006), delivering hierarchical (rather than flat) decompositions for words or sentences.

While AGs are essentially viewed as an unsupervised grammatical inference tool, several authors have also tried to better inform grammar inference with external knowledge sources. This is the case of Sirts and Goldwater (2013), who study a semi-supervised learning scheme combining annotated data (parse trees) with raw sentences. The linguistic knowledge considered in (Johnson et al., 2014) aims to better model function words in a language acquisition setting: explicitly representing the occurrence of these short (typically monosyllabic) tokens in front of content-bearing words was shown to improve the resulting word segmentations. The work of Eskander et al. (2016) considers the use of additional dictionaries, storing partial lists of prefixes or suffixes collected either on the Internet, or discovered during a first round of training. We study similar complementary information, which are collected in close collaboration with linguistic experts.

Various other extensions or applications of AGs are worth mentioning, such as O'Donnell et al. (2009), which generalizes AGs so as to adapt fragments of subtrees (rather than entire subtrees). Botha and Blunsom (2013) consider the adaptation of grammars from a more general class than context-free grammars (mildly context-sensitive grammars), in order to model discontinuous fragments in non-concatenative morphology. Finally, Börschinger and Johnson (2014) propose to model the role of stress cues in language learning.

7 Conclusion

This paper had two main goals: (1) improve upon a strong baseline for the unsupervised discovery of words in two very low-resource Bantu languages; (2) explore the Adaptor Grammar framework as an analysis and prediction tool for linguists studying a new language.

Systematic experiments with 162 grammar configurations for each language have shown that using AGs for word segmentation is a way to test linguistic hypotheses during a language documentation process. Conversely, we have also shown that specializing a generic grammar with language specific knowledge greatly improves word segmentation performance. In addition, our paper reports word segmentation results that are way higher than a Bayesian baseline. These results invite us to further this collaboration, and to analyze more thoroughly the usability of output parses in speeding up the documentation process.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We also thank Ramy Eskander for his help in the early stages of this research. This work was partly funded by French ANR and German DFG under grant ANR-14-CE35-0002 (BULB project).

[14] We include B3 variant, interpreting its non-terminal Prefixes as a prefix.

References

Oliver Adams, Trevor Cohn, Graham Neubig, and Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 53–60.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The Bulb project. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia.

Célestin Amboulou. 1998. *Le Mbochi: Langue Bantu Du Congo-Brazzaville (Zone C, Groupe C20)*. Ph.D. thesis, INALCO, Paris.

Odette Ambouroue. 2007. *Éléments de description de l'orungu, langue bantu du Gabon (B11b)*. Ph.D. thesis, Université Libre de Bruxelles.

Antonios Anastasopoulos and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Honolulu. Association for Computational Linguistics.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. *ACL 2014*.

David Blachon, Élodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. *Procedia Computer Science*, 81:61–66.

Benjamin Börschinger and Mark Johnson. 2014. Exploring the Role of Stress in Bayesian Word Segmentation using Adaptor Grammars. *Transactions of the Association of Computational Linguistics*, 2:93–104.

Jan A. Botha and Phil Blunsom. 2013. Adaptor Grammars for Learning Non-Concatenative Morphology. In *EMNLP*, pages 345–356.

Zhiyi Chi. 1999. Statistical properties of probabilistic context-free grammars. *Comput. Linguist.*, 25(1):131–160.

David Crystal. 2002. *Language Death*. Cambridge University Press. Cambridge Books Online.

Laura J. Downing. 2005. On the ambiguous segmental status of nasals in homorganic NC sequences. In *The Internal Organization of Phonological Segments*, pages 183–216.

Georges Martial Embanga Aborobongui. 2013. *Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25)*. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the Use of Adaptor Grammars for Unsupervised Morphological Segmentation of Unseen Languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910, Osaka, Japan. The COLING 2016 Organizing Committee.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito. 2018a. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Proceedings of LREC*, Miyazaki, Japan.

Pierre Godard, Kevin Loser, Alexandre Allauzen, Laurent Besacier, and Francois Yvon. 2018b. Unsupervised learning of word segmentation: Does tone matter ? In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING), Hanoi, Vietnam.*

Sharon Goldwater. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual Dependencies in Unsupervised Word Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia. Association for Computational Linguistics.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Fatima Hamlaoui and Emmanuel-Moselly Makasso. 2015. Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. *Lingua*, 154:35–64.

Robert K. Herbert. 1986. *Language Universals, Markedness Theory, and Natural Phonetic Processes*. De Gruyter Mouton, Berlin, Boston.

Tore Janson. 2003. *Speak: A Short History of Languages*. Oxford University Press.

Mark Johnson. 1998. PCFG models of linguistic tree representations. *Comput. Linguist.*, 24(4):613–632.

Mark Johnson. 2008a. Unsupervised Word Segmentation for Sesotho Using Adaptor Grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio. Association for Computational Linguistics.

Mark Johnson. 2008b. Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure. In *Proceedings of ACL-08: HLT*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Mark Johnson, Anne Christophe, Emmanuel Dupoux, and Katherine Demuth. 2014. Modelling function words improves unsupervised word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 282–292, Baltimore, Maryland. Association for Computational Linguistics.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A Framework for Specifying Compositional Nonparametric Bayesian Models. In *Advances in Neural Information Processing Systems 19*, pages 641–648, Cambridge, MA. MIT Press.

Guy-Noël Kouarata. 2014. *Variations de formes dans la langue mbochi (Bantu C25)*. Ph.D. thesis, Université Lumière Lyon 2.

Kamran Lari and Steve J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

David Odden. 2015. Bantu Phonology. *Oxford Handbooks Online*.

Timothy J. O'Donnell, Joshua B. Tenenbaum, and Noah D. Goodman. 2009. Fragment grammars: Exploring computation and reuse in language. Technical report, Massachusetts Institute of Technology.

Annie Rialland, Georges Martial Embanga Aborobongui, Martine Adda-Decker, and Lori Lamel. 2015. Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In *Proceedings of the 44th ACAL meeting*, pages 221–230, Somerville. Cascadilla.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.