Design and Development of Part-of-Speech-Tagging Resources for Wolof (Niger-Congo, spoken in Senegal)

Cheikh M. Bamba Dione[*] **Jonas Kuhn**[*/**] **Sina Zarrieß**[**]

[*]Department of Linguistics
University of Potsdam
Germany

[**]Institute for Natural Language Processing (IMS)
University of Stuttgart
Germany

dione@uni-potsdam.de, jonas.kuhn@ims.uni-stuttgart.de,
sina.zarriess@ims.uni-stuttgart.de

Abstract

In this paper, we report on the design of a part-of-speech-tagset for Wolof and on the creation of a semi-automatically annotated gold standard. The main motivation for this resource is to obtain data for training automatic taggers with machine learning approaches. Hence, we take machine learning considerations into account during tagset design and present training experiments as part of this paper. The best automatic tagger achieves an accuracy of 95.2% in cross-validation experiments. We also wanted to create a basis for experimenting with annotation projection techniques, which exploit parallel corpora. For this reason, it was useful to use a part of the Bible as the gold standard corpus, for which sentence-aligned parallel versions in many languages are easy to obtain.

1. Introduction

This paper[1] presents work on the design and development of annotated corpus resources supporting part-of-speech-(PoS-)tagging for Wolof, a language from the Niger-Congo family mainly spoken in Senegal. Specifically, we discuss the design of tagsets of various granularity, created for automatic tagging purposes, we report on a process of successive improvement of a manually corrected gold-standard annotation of training data, and we show the results of a number of machine learning experiments based on this resource. This work is, to our knowledge, the first effort in building a publicly available NLP resource for the Wolof language. More generally, there has recently been a growing interest in NLP technologies for African languages, see, e.g. (Pauw et al., 2009) for current developments in this field. We believe that to a large degree, the techniques we report can be generalized to similar efforts for other lesser-studies languages, although we are in a somewhat special situation, as the first author combines (i) the expertise of a computational linguist and (ii) native speaker knowledge of the language under consideration in one person. It may not always be possible to find such a person for the task of gold standard design and development.

This research is situated in a large collaborative research programme on information structure (SFB 632)[2]. In the context of this programme, the building of linguistically annotated resources for lesser-studied languages is supporting empirical, corpus-based investigations of the crosslingual realization of information structure. A discussion of the annotation infrastructure and its application in research on information structure is found in (Chiarcos et al., 2009). From an NLP perspective, these lesser-studied languages are (i) an interesting test-bed for the annotation and training techniques established for well-studied languages, and (ii) a challenge for weakly supervised annotation techniques (such as cross-lingual projection on a parallel corpus) as a way of avoiding the high manual annotation effort that cannot realistically be spent on all languages that would be of interest. Our project addresses both aspects, but this paper focuses on (i), i.e., the development of a relatively large gold-standard resource without recourse to weak supervision techniques. This resource has a high value of its own, in particular as a sufficiently large training set for established statistical PoS-taggers. But in addition, it is a prerequisite for a systematic study of aspect (ii), as the usefulness of weak supervision techniques can only be judged in comparison with more traditional, supervised approaches. A detailed account of the design and development process is given in (Dione, in preparation).

With practically no NLP resources available for Wolof, we had to design a tagset and create a PoS-annotated gold standard from scratch. The main purpose of the gold standard is to serve as training data for automatic tagging, using various learning techniques. So the tagset design followed two higher-level goals: (a) linguistic distinctions relevant for expressing effective PoS search patterns should be covered, taking into account typological peculiarities of the language; (b) automatic tagging based on the tagset should deliver high-accuracy performance. This means that notoriously indistinguishable aspects should be represented in an underspecified way that is transparent to the user. Due to these twofold goals, the development process of the Wolof tagset was closely interleaved with error analyses of automatic PoS annotation.

Besides the discussion of the tagset design, this paper presents the process of creating a semi-automatically annotated gold standard, exploiting available lexical resources and using purpose-built heuristic tools for stemming and

[1]The work reported in this paper was in part supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in (i) the Emmy Noether project PTOLEMAIOS, on Grammar Induction from Parallel Corpora, and (ii) SFB 632 on Information Structure, project D4 (Methods for interactive linguistic corpus analysis).

[2]www.sfb632.uni-potsdam.de/

guessing of word forms. Finally, we demonstrate the use of the gold standard resource in machine learning experiments, providing a comparison of results achieved by statistical state-of-the-art PoS taggers on our gold standard and a brief summary of experiments making use of cross-lingual projection on the parallel corpus data.
Section 2. provides some background on Wolof, section 3. introduces the corpus we use as our basis for annotation. In section 4., we discuss the considerations behind our tagset design, section 5. reports on the semi-automatic process of gold standard annotation. In section 6., we summarize some of our machine learning experiments using the gold standard resource, before concluding in section 7.

2. Wolof Language

With about 4 million native speakers, Wolof is one of the most widely spoken languages within the West Atlantic branch of the Niger-Congo language family. Moreover, Wolof is used as a lingua franca in Senegal such that 80% of the population are assumed to speak Wolof.
The West Atlantic languages have attracted theoretical interest at least since the 1970's (their classification goes back to (Sapir, 1971)). The phenomena cited as characteristic in the literature include (i) the complex systems of nominal classifiers, (ii) consonant mutations, i.e. regular alternations of consonants in the morphological paradigm of a noun or verb, (iii) verbal extensions, i.e. systems of morphemes (suffixes) which can be affixed to a verb and change its syntactic behaviour, and (iv) the interaction between focus and inflectional markers/pronominals/clitics (Becher, 2002; Zribi-Hertz and Diagne, 2002; Russell, 2006).
Among the West Atlantic languages, Wolof ranks among the linguistically well documented languages. Two main aspects of Wolof's grammar have been mostly studied in the literature: First, Wolof has a very productive derivation morphology for nouns and verbs allowing to alter the category, valence, and semantics of a nominal or verbal base (Becher, 2002). An example is given in (1) where the *-al* affix allows the verb *togg* "cook" to select for a benefactive argument. Note also that the main verb in (1) does not itself carry inflectional markers. This is quite typical, and in the following sections, we will show that this is a notoriously difficult problem for automatic PoS annotation.

(1) Togg-al naa xale bi ceeb.
Cook-APPL 1SG child DET rice.
I cooked rice for the child.

Second, Wolof exhibits a complex system of obligatory inflectional elements, pronouns or clitics[3] that appear as separate tokens or as verbal suffixes, i.e. they mainly replace verbal inflection. The inflectional elements express the person of the verb's subject, aspect, tense, polarity, and – what makes Wolof particularly interesting for research in information structure – the focus in the sentence (e.g., verb focus, subject focus, object focus). Some examples, due to (Robert, 2000), are given in (2). In (2-a), the perfective aspect of the verb is indicated on the subject pronoun which also carries person and number information. If the verb *lekk* "eat" is to be negated, it has to be affixed by the morpheme *-ul* that also inflects for person and number (2-b-c).

(2) a. Lekk nga.
Eat PERF.2SG.
You have eaten.
b. Lekk-uloo.
Eat-NEG.2SG.
You have not eaten.
c. Lekk-ul.
Eat-NEG.3SG.
He has not eaten.

Several different paradigms of the inflectional markers are available, depending on which part of the sentence is focussed. Thus, the information structure of a sentence is generally explicitly marked in the syntax, leading to interesting interactions between focus and, e.g., aspect, polarity or interrogation (Perrin, 2005; Robert, 2000). Sentences (3-a-c) illustrate the overt focus marking on the subject pronominal. Sentence (3-a) does not have an explicit focus marker, *na* only expresses perfective aspect. The sentence could thus be used in an all-focus context. In (3-b) the subject is focussed, due to *moo* (we use the gloss FOC-S to mark this). Sentence (3-c) illustrates a verb focus.

(3) a. Peer lekk na.
Peer eat PERF.3SG.
Peer has eaten.
b. Peer moo ko lekk.
Peer FOC-S.3SG pro eat.
PETER ate it./It was Peer who ate it.
c. Peer dafa ko lekk.
Peer FOC-V.3SG pro eat.
Peter ATE it.

Moreover, Wolof lacks adjectives whose role is taken over by stative verbs. The contrast between non-stative and stative verbs is illustrated in (4). Whereas the perfective of a non-stative verb in (4-a) is interpreted as past tense, the perfective of a stative (adjectival) verb in (4-b) receives a present tense interpretation. See (McLaughlin, 2004) for a discussion of adjectives in Wolof. In our tagset, we do not include a part-fo-speech category for adjectives.

(4) a. Dem naa Ndakaaru.
Go PERF.1SG Dakar.
I went to Dakar.
b. Sonn naa.
Tired PERF.1SG.
I am tired.

General references on Wolof include (Diagne, 1971; Jean Léopold Diouf, 1991; Ndiaye, 2004).

[3]We remain neutral as to the exact syntactic status of these elements. Since tagging operates at the word/token level, it is required to assign a tag to each token, and the goal is to design a tagset that is both reliable and informative with respect to the syntactic function of the linguistic elements. This means that wherever possible, special functions and distributional characteristics of the inflectional markers/pronouns/clitic should be included in the annotation; the category label for the elements is less important.

3. Wolof Corpora

Besides the Wolof Wikipedia[4] and some linguistic web pages[5] it is currently still difficult to obtain larger amounts of electronic Wolof texts. Morevover, divergences in spelling conventions are an issue for heterogenous text collections. Our goal for the current effort was to obtain a relatively large dataset of homogenous text based on consistent conventions. Moreover, because of the annotation projection experiments we are planning, we needed a parallel corpus. For these reasons, we decided to use the Wolof translation of the Bible (New Testament) as our main corpus. It has a consistent orthography and can be straightforwardly used as a parallel corpus with accurate sentence (i.e., verse) alignment. The Bible corpus contains 203,200 tokens in total, for our (semi-)automatic PoS annotation experiments, we selected 28 chapters of the Matthew gospel (26,846 tokens).

For the tagset design and development, we made sure that different language registers were taken into account. Hence, we also took orthographically transcribed contemporary dialogue data into account, in particular in the early stages of tagset design. For this, we could use a Map Task spoken language corpus, kindly provided to us by Uli Reich from Freie Universität Berlin.

4. Tagset Design

Obviously, the first requirement for manual or automatic PoS annotation is a consistent and complete tagset for the language under consideration. We designed the tagset from scratch, following the EAGLES guidelines for PoS models (Leech and Wilson, 1996) wherever possible. Even at the level of coarse-grained word categories, there is no established PoS inventory for Wolof. Since the behaviour of certain word classes in Wolof substantially differs from Indo-European languages (see section 2.), or certain word classes may simply not exist (e.g. adjectives), it is not a trivial task to adapt the design of the major, well-established tagsets to Wolof. We used the existing dictionaries and grammars (Diagne, 1971; Jean Léopold Diouf, 1991; Ndiaye, 2004) as a guidance, however, these resources are not always consistent, especially with respect to difficult categories like verbs. Table 1 summarises the main lexical categories we used in our annotation experiments.

Adverbs	dell (*fully*), tey (*today*)
Prepositions	ci (*in,on*)
Articles	cib (*in the*), cab (*in a*)
Comparatives	ni (*like*)
Conjunctions	ak (*and*)
Determiners	ban, gan (*which*)
Inflection markers	maa, yaa
Nouns	téere (*book*)
Pronouns	googu (*those*)
Particles	woon (past tense particle)
Verbs	war (*shall*)
Reflexives	boppam (*himself*)
Foreign language material	ràbbuni (*"my God"*)
Punctuation	

Table 1: Inventory of Lexical Categories in Wolof

4.1. PoS categories for Verbs

The difficulties of tagset design for Wolof can be well illustrated for verbs. Whereas the major tagsets for European languages distinguish various verbal categories for verbs according to their finiteness, the issue is less clear in Wolof. A possible approach would be to follow the literature, in particular the work by (Zribi-Hertz and Diagne, 2003), who propose three categories of verb finiteness. (Zribi-Hertz and Diagne, 2003) distinguish between (i) finite verb occurrences (with inflections for person, aspect, tense, and polarity present in the clause), (ii) "deficiently finite" verb occurrences whose inflection does not indicate the verb arguments' person but only aspect, tense and polarity markers, and (iii) non-finite verbs which occur when no inflection is present in the clause. All three types of verbs can function as the main predication of the sentence. It may seem natural to always encode this three-way distinction in the tag for the main verb, since this is the one category that is reliably present in a clause.

We conducted some preliminary experiments adopting this scheme and integrated three PoS labels in our tagset (VVFIN, VVNFN, VVINF) corresponding to finite, deficiently finite, and infinite verbs. In the manual annotation, these tags were used to reflect the presence of the respective inflection within the clause, i.e. on the verb itself or in the verb's context. However, when we performed machine learning experiments on our Bible corpus, it turned out that the distinction between these categories was very hard to pick up for standard machine learning approaches. Table 2 presents the ten most frequent errors made by the TreeTagger trained on annotated data distinguishing 3 verb finiteness categories. This is not too surprising given the variance in the exact patterning of the inflectional marking, which is orthogonal to further distinctions needed for verbs (compare for instance example (2)). As a consequence, we conflated the three variants into one tag (VVBP) for verb bases without a token-internal inflection, besides a fine-grained distinction of token-internally inflected verbs on the one hand and separated inflectional markers on the other hand. Note that at the clause level, the relevant functional information can thus be recovered, and further processing steps, such as parsing, will benefit from more reliable decisions at the PoS level.

4.2. PoS categories for Focus Markers

In other cases, our tagset captures fine-grained, intra-categorial distinctions, e.g., the focus related types of pronominal markers. These fine-grained categories are easy to establish for human annotators and can be well recognized by automatic PoS annotation procedures since the pronominal markers are very frequent and exhibit little syncretism with respect to person, aspect and focus type. In consequence, our PoS annotated resource allows for searching focus constructions and their contexts. For instance, a study of Wolof interrogation could extract subject vs. object focus questions by means of the PoS patterns 'PW ICF' (in-

[4] wo.wikipedia.org/wiki

[5] www.linguistique-wolof.com/corpus.html

(incorr.) system tag	gold tag	error ratio wrt. gold tag	tokens affected in entire test corpus
VVFIN	VVNFN	5.88%	0.83%
VVNFN	VVINF	45.24%	0.72%
NC	VVNFN	4.28%	0.60%
VVNFN	VVFIN	30.43%	0.53%
NC	NP	12.22%	0.42%
VVNFN	VVRP	29.17%	0.26%
VVNFN	NC	2.23%	0.23%
VVINF	VVNFN	1.60%	0.23%

Table 2: Excerpt from confusion matrix for TreeTagger on a tagset that distinguishes 3 verb finiteness categories

terrogative pronoun, object focus inflection) vs. 'PW ISUF' (interrogative pronoun, subject focus inflection).

4.3. Multiword Units

A further difficulty for the definition of word classes is the treatment of multiword units which are very common in Wolof. For instance, the pronominals or focus markers and their corresponding inflection often appear as separated words in the text, e.g. the sentence focus *maa ngi* where *maa* is an inflectional marker which carries information about the subject's first person and the type of focus, whereas *ngi* is an invariant sentence focus particle, i.e. if a sentence had a 3rd person subject its focus particle would be *mu ngi*. In this case, we decided to strictly follow standard tokenization, assigning a tag to each space-separated element in standard orthography. We introduce a special tag for the first component of the multiword unit, in this case an "inflectional sentence focus marker" occurring in front of a "sentence focus particle". Thus, the multiword *maa ngi* is labelled as *maa/ISF ngi/UPSF* where ISF corresponds to sentence focus inflection marker and UPSF is a sentence focus particle.

4.4. Tagset Granularity

Depending on the application of the PoS annotation, different granularities of the tagset may be needed. A fine-grained tagging can of course be easily reduced to a coarser one. But in the context of machine learning of taggers, the accuracy of automatic tagging will be influenced by the size and granularity of the tagset. Hence, the goal is to find a compromise for a multi-purpose tagset that balances the needs of suitably fine-grained tag distinctions and reliability of automatic tagging.

The factors influencing automatic tagging quality are very subtle (a certain tagset may lead to a large number of ambiguous word forms, but standard tagging approaches may still be able to disambiguate them reliable in the contextual window). For this reason, we decided to start out with a rather fine-grained tagset of 200 different categories, which we used to annotate the entire gold standard corpus, using heurstics for semi-automatic annotation (see section 5.).

The fine-grained word class labels carry information about morphological categories like number, person or aspect for pronominals. Annotation at this level includes a fair amount of morphological analysis. The fully tagged gold standard can thus be used for studies at this detailed level.

In the context of machine learning oriented work however, the fine-grained tagset is seen as the point of departure for developing more compact candidate tagsets, which can be obtained by a systematic mapping from the detailed tagset. In section 6., we present tagger training experiments at three level of granularity: besides the full tagset of 200 categories, we explored a medium size tagset of 44 tags and a coarse one just consisting of the 14 major lexical categories. Examples for comparison are shown in table 3. Note for instance that in the reduced tagsets the distinction of nominal classes (b-class, y-class etc.) is dropped from the full tagset. These are lexically determined, so while it is relatively easy to reconstruct them from the lexicon for known words, it would be very hard to assign them reliably to unknown items in automatic tagging. So it is reasonable to leave this information out of the automatic tagging procedure.

As a consequence of various experiments and considerations, we finally designed a standard tagset of 80 tags (compare the last column in table 3) that adds "easily detectable" distinctions to the original medium size tagset, thus making it more useful for morphosyntactic studies of Wolof. For reference, the full standard tagset is also listed in the Appendix.

5. Heuristics for Semi-Automatic Annotation

To obtain a gold standard for training automatic PoS taggers, we annotated 26,846 tokens from the Matthew gospel of the Wolof Bible, using the GATE environment.[6] By automatically pre-annotating the corpus with guessed PoS tags, we could reduce the annotation effort and time by more than a half. Note however that all automatic annotation steps on the gold standard were carefully hand-checked in order to guarantee a high quality standard. The semi-automatically supported annotation process for the Matthew gospel took a little more than one person month, with an average rate of 700 tokens (or 23 sentences) per day.

As a basis for pre-annotation, we first compiled a lexicon consisting of 1700 entries for closed-class lexemes from (Diagne, 1971; Jean Léopold Diouf, 1991; Ndiaye, 2004; Ka, 1994). The lexicon also includes a list of proper names available at `www.senegalaisement.com`.

For improved pre-annotation of open-class lexemes, we used a set of heuristics taking advantage of morphological patterns to build an extended full form lexicon. In a cyclic process, we identified known inflectional and derivational suffixes (Ka, 1994) in the word forms occurring in the corpus. After manual checking, the PoS category marked by the respective morpheme was added as a possible category for the respective word form. Moreover, we used the word stems obtained by cutting the known suffixes off in order to generate additional word forms based on regular patterns of inflectional and derivational morphology. In order to avoid overgeneration, we took vowel harmony into account in the morphological generation process.

[6] `gate.ac.uk/`

Detailed tagset 200 Tags	Description	Medium 44 tags	General 14 tags	Standard 80 tags
ATDs.b.P	def. art., SG, b-class, proximal	ATDs	AT	ARTD
ATDp.y.R	def. art., PL, y-class, remote	ATDp	AT	ARTD
ATDs.b.SF	def. art., SG, b-class, sent. focus	ATDSF	AT	ARTF
ATDs.w.SF	def. art., SG, w-class, sent. focus	ATDSF	AT	ARTF
ATDp.ñ.SF	def. art., PL, ñ-class, sent. focus	ATDSF	AT	ARTF
I.1p.CF.PF	infl. marker, 1SG, compl. focus, perf	ICF	I	ICF
I.1p.DiFut.IMPF	infl. marker, 1SG, di future, impf	IFUT	I	IFUT
I.3p.NF.PF	infl. marker, 3PL, no focus, perf.	INF	I	INF
I.1p.VF.PF	infl. marker, 1PL, verb. focus, perf.	IVF	I	IVF
I.1s.SuF.IMPF	infl. marker, 1SG, subj. focus, imperf.	ISUF	I	ISuF
I.3p.SF	infl. marker, 3PL, sent. focus, perf	ISF	I	ISF
Pind.1p	free pron., 1SG	PRON	PD	PERS
Pind.2p	free pron., 2SG	PRON	PD	PERS
Pobj.3s.IMPF	object pron., 3SG perf.	PRON	PD	PRO
Pobj.3s.PF	object pron., 3SG impf.	PRON	PD	PRO
Psub.1p.IMPF	subject pron., 1PL, imperf	PRON	PD	PRS
Psub.1p.PF	subject pron., 1PL, perf.	PRON	PD	PRS
VNEIMP.2s	imp. negative, 2SG	VNEIMP	V	VNEIMP
VXNEG.1p	modal aux. neg., 1PL	VXNEG	V	VXNEG

Table 3: Examples of tags in the tagsets of different granularities

After compiling the extended full form lexicon, we used a heuristic procedure to generate the input for manual checking. For known ambiguous word forms, the full choice of options was presented to the annotator to choose from. For unknown word forms, suffix-based guessing was applied, and again, the possible choices were presented. For example (5), the annotator would get an input as in (6).

(5) man de ab kanaara la fi gis.
I interj. ART:indef turkey O.3sg here see.
I can only see a turkey here.

(6) man_<PERS|DWQ> de_<IJ> ab_<ARTI> kanaara_<NC> la_<PRO|ICF|ARTD> fi_<AV> gis_<VBP>

We also experimented with contextually-driven automatic disambiguation rules in order to speed up the manual annotation process. Of course, the rules had to be formulated in a conservative way to avoid elimination of the correct reading. But for instance for certain category/particle combinations, disambiguation of one element also determines the other choice. Hence, taking advantage of explicit rules led to a moderate additional reduction of annotation effort.

6. Automatic PoS tagger training experiments

In this section, we present machine learning experiments where our Bible gold standard serves as training and test data for induction of an automatic PoS tagging system. First of all, we wanted to address the question whether available state-of-the-art PoS taggers that have been successfully used for numerous, although mostly European, languages obtain satisfactory results on our Wolof data. The setup and results are described in section 6.1. Second, we started investigating the question whether we can exploit the fact that the Bible is a parallel text for automatic PoS annotation. In section 6.2., we report on the experimental setup and some first, preliminary results.

6.1. State-of-the-art statistical PoS taggers

There are a number of available statistical PoS taggers which have been mainly trained and tested on the major European languages. We assessed the performance of two well-known available machine-learning taggers on our Wolof data:

1. TnT tagger (Brants, 2000), based on a trigram Hidden Markov model. The authors report 96.7% accuracy on the German NEGRA corpus.
2. TreeTagger (Schmid, 1994) implements a decision tree model (96.06% on NEGRA).
3. SVMTool (Giménez and Màrquez, 2004) implements a support vector machine classifier (97.1% on the Wall Street Journal). SVMTool uses a very rich, lexical feature model.

We also compare against a baseline, which assigns to each known word form its most frequent tag from the training set, and to each unknown word form the most frequent tag overall (the 'NP' proper name tag).
We investigated (i) the performance of the different taggers for Wolof, (ii) the performance of the taggers depending on the size of the tagset. The average number of ambigous categories for the most fine-grained annotation level was 5.173 per word. We carried out a ten-fold cross-validation on the gold standard Matthew corpus for the various tagsets, training each of the taggers on 90% of the corpus (26,846 tokens) and evaluating on the remaining 2650 tokens. The results are summarized in table 4.[7]

[7]Confidence intervals are given for $p > 0.05$. Note that due to the different tagsets, the accuracy numbers should not be compared directly across tagsets.

	Accuracy			
Tagset size	**200**	**44**	**15**	**80**
Baseline	85.7% ± 0.9	88.4% ± 0.8	89.5% ± 1.0	87.6% ± 0.8
TnT	92.7% ± 0.6	94.2% ± 0.4	94.8% ± 0.4	94.5% ± 0.4
TreeTagger	90.7% ± 0.8	93.6% ± 0.5	94.5% ± 0.6	93.8% ± 0.5
SVM Tool	93.1% ± 0.6	95.3% ± 0.4	96.2% ± 0.3	95.2% ± 0.4

Table 4: PoS-tagging accuracy scores for the different tagsets

As mentioned in section 4.4., we developed the fourth, "Standard" tagset as an extension of the medium-sized tagset, avoiding additional sources of error. As the results indicate, this seems to be the case.

Figure 5 displays the most frequent confusions of the TnT tagger on the Standard tagset. The most prominent error affects the distinction between verbs (VV) and nouns (NC), thus concerning the most frequent open-class words. The difficulty to distinguish verbs and nouns error is probably due to the flexible derivation morphology and the non-existing verb inflection.

(incorr.) system tag	gold tag	error ratio wrt. gold tag	tokens affected in entire test corpus
VV	NC	3.94%	0.42%
NC	VV	1.95%	0.38%
PREL	PERS	3.07%	0.34%
NP	NC	3.23%	0.34%
PREL	AT	5.59%	0.30%
AV	NC	2.51%	0.26%
NP	VV	1.17%	0.23%
AT	AP	2.37%	0.15%

Table 5: Excerpt from confusion matrix for TnT

6.2. Towards an exploitation of parallel corpus data

Our research project that we sketched in the introduction also aims at investigating ways of exploiting parallel corpus resources as a way of "injecting" information in the annotation or training process. This is particularly interesting for lesser-studied languages since every way of facilitating or speeding up resource building will be highly welcome. A well-known technique of exploiting parallel corpora, or bitexts, is annotation projection as pioneered by (Yarowsky and Ngai, 2001). Here, the part of the bitext that is in a well-studied language like English is analyzed with a relatively reliable automatic tool, and a (statistical) word-alignment over the bitext is used to "project" word annotations, such as PoS tags, from this source language to a lesser-studied target language. On the target side, the annotation can then be used as training data for a (noise-robust) machine learning approach. Of course, the approach is confronted with multiple sources of errors and typically requires some amount of language-specific tuning in order to warrant useful results.

We hypothesize that for practical purposes, it may not be full annotation projection that is most useful, but rather some slightly different ways of "injecting" bitext information in the process of building analysis tools for lesser-studied languages. Here, we present some very preliminary experiments of a possible such approach. We assume that the target language tool, a PoS tagger in our case, is not trained on projected data exclusively, but on some – potentially very small – amount of genuine gold standard data for the target language. The bitext information is used in the training, essentially as additional feature information, which may make the small amount of training data more informative.

In our experiments we proceeded as follows: Since we chose parts of the Bible as our gold standard, we can apply standard statistical word alignment techniques (using GIZA++[8]) to align the Wolof words with the words from modern English and/or French Bible translations. The English and the French strings can be PoS-tagged using state-of-the-art taggers. For the classification decision of assigning a particular tag to a given Wolof word in its context, we can now not only exploit lexical and contextual knowledge from Wolof, but also correspondences in English and French, presumably mainly by relying on generalizations reflected in the PoS tags. Thus, in contrast to (Yarowsky and Ngai, 2001), we do not project annotations directly, but rather use them as an additional clue.

Example sentence (7) with the annotated, automatically word-aligned Wolof-English translation correspondence is displayed in figure 7. The sentence includes a number of nice correspondence links, but at the same time illustrates that direct projection of PoS categories may be problematic in cases where the translations are not as close as possible (*Yeesu – he*), or where multiple alignments for a single word form may be misleading (*indil-leen – them*); such configurations are very common, even between closely related languages, and for less related languages there are likely to be many more such cases. Note that PoS information that would be incorrect if projected as fixing the target language category may still be very informative as feature information for a machine learning classifier.

(7) Yeesu ne leen: "Indil-leen ma ko fii."
Jesus tells them bring-you me them here
Jesus tells them: "Bring them here to me."

We performed preliminary experiments with a MaxEnt tagger (in which the word alignment mediated information from the parallel languages is provided as features) and a variant of an HMM tagger. The latter is assumed to have more than one "output tape", omitting not only a word form in each state (corresponding to a PoS tag), but also zero to n foreign language tags (for the foreign words linked to the word by the word alignment).

Some results indicating the usefulness of the parallel corpus information are shown in table 6. Here, a statistically

[8] code.google.com/p/giza-pp/

	Training data size (tokens)		
	418	1249	4968
no parallel information	59.7% ± 1.1	68.3% ± 1.2	82.7% ± 0.9
information from English	62.6% ± 1.1	70.2% ± 0.6	84.0% ± 0.9
information from English and French	63.6% ± 1.2	70.6% ± 1.2	84.1% ± 1.0

Table 6: Training results for a "multi-tape" HMM tagger with and without information from the parallel corpus

NP	Yeesu	he	PP
VVBZ	ne	said	VVD
PRO	leen	:	:
$.	:	“	“
$(	“	bring	NP
VVIMPE	Indil-leen	them	PP
PRO	ma	here	RB
PRO	ko	to	TO
AVDEM	fii	me	PP
$.	.	.	SENT
$(	”	”	”

Table 7: Example for a Wolof sentence from Bible gold standard and its PoS tagged, word-aligned English translation

significant relative improvement due to parallel corpus information could be observed in a situation where very few gold standard data were used in training.[9] With larger sets of training data, the effect is no longer significant. A natural extension of the approach is to use a small seed set of data in a bootstrapping or active learning set-up for extending the set of reliable training data.

7. Conclusion

We discussed the development of PoS resources for a lesser-studied language. Our approach is oriented towards automatic tagging and combines manual tagset development and (semi-)automatic annotation, using very effective heuristics for pre-annotation. We consider the results achieved by state-of-the-art taggers on the gold standard quite satisfactory and plan to use the resource for further experimentation. This includes the exploration of further semi-automatic techniques, such as weak supervision techniques taking advantage of information from the parallel corpus set-up. Thus, we believe that our research has implications beyond resource development for Wolof itself, as many aspects of the Wolof scenario are quite comparable to other languages.

We also plan to explore the usability of the tagset, for instance for linguistic research on information structure. Since the gold standard corpus was carefully annotated with a fine-grained underlying tagset, it is also conceivable to make task-specific adjustments to the tagset (for which it will again be interesting to explore to what degree they can be picked up, possibly by a bitext-informed tagging procedure).

8. References

Jutta Becher. 2002. Verbalextensionen in den atlantischen Sprachen. *Hamburger afrikanistische Arbeitspapiere (HAAP)*, pages 1–38.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Christian Chiarcos, Ines Fiedler, Mira Grubic, Andreas Haida, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes, and Malte Zimmermann. 2009. Information structure in African languages: Corpora and tools. In Guy De Pauw, Gilles-Maurice de Schryver, and Lori Levin, editors, *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT 2009)*. The Association for Computational Linguistics.

Pathé Diagne. 1971. *Grammaire de wolof moderne*. Paris: Présence africaine.

Cheikh M. Bamba Dione. in preparation. Part-of-Speech-Tagging für die Sprache Wolof (Senegal). Diplomarbeit [equiv. of Master thesis] Universität Potsdam.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC*.

Marina Yaguello Jean Léopold Diouf. 1991. *J'apprends le wolof (Damay jang wolof)*. Editions Karthala, 75013 Paris.

Omar Ka. 1994. *Wolof phonology and morphology*. University Press of America Lanham, Maryland 20706.

Geoffrey Leech and Andrew Wilson. 1996. EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora. Technical report, Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-TCWG-MAC/R.

Fiona McLaughlin. 2004. Is there an adjective class in wolof? In R.M.W. Dixon and Alexandra Y. Aikhenvald, editors, *Adjective classes. A crosslinguistic typology.*, pages 242–262. Oxford University Press.

Moussa D. Ndiaye. 2004. *Éléments de morphologie du wolof*. LINCOM Studies in African Linguistics.

Guy De Pauw, Gilles-Maurice de Schryver, and Lori Levin. 2009. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT 2009)*. The Association for Computational Linguistics.

Loïc-Michel Perrin. 2005. *Des représentations du temps en wolof*. Ph.D. thesis, Université Paris VII Dénis Diderot, France.

Stéphane Robert. 2000. Le verbe wolof ou la grammaticalisation du focus. [Louvain: Peeters, Coll. Afrique et Langage, 229-267. Version non corrigée.].

Margaret A. Russell. 2006. *The Syntax and Placement of Wolof Clitics*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

David J. Sapir. 1971. West atlantic: an inventory of the languages, their noun class systems and consonant alternation. In T.A. Sebeok, editor, *Current Trends in Linguistics 7*, pages 45–112.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North*

[9] The confidence intervals are provided for $p > 0.05$.

American Chapter of the Association for Computational Linguistics on Language technologies 2001, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Anne Zribi-Hertz and Lamine Diagne. 2002. Clitic placement after syntax: Evidence from Wolof person and locative markers. *Natural Language and Linguistic Theory*, 20(4):823–884.

Anne Zribi-Hertz and Lamine Diagne. 2003. Déficience flexionnelle et temps topical en wolof. In Patrick Sauzet et Anne Zribi Hertz, editor, *Typologie des langues d'Afrique et universaux de la grammaire, volume 2: benue-kwa, wolof*, pages 205–231. L'Harmattan.

A Standard tagset

Code	Description	Examples
$,	comma	,
$.	sent.-final punct.	. ? ! ; :
$(	other punct.	- [,]()
ADV	general adverb	dell (*fully*), tey (*today*)
AVDEM	demonstrative adverb	foofu (*there*), noonu (*so*)
AVREL	subord. adverb	fu (*where*)
AVWQ	adverbial interr. pron.	nan (*how*)
AP	preposition	ci (*in,on*)
APART	contraction of preposition and article	cib (*in the*), cab (*in a*)
ARTD	def. article	bi (*the* b-class), gi (*the* g-class)
ARTF	def. article with focus	baa
ARTI	indef. article	ab (*a*), ay (indef. plural article)
CC	coord. conjunction	ak (*and*)
COMP	comparative particle	ni (*like*)
CSF	subord. conjunction with finite clause	ni (*such as*), su (*as* + conditional)
CSN	subord. conjunction with non-finite clause	ngir (*in order to*)
DWQ	interr. determiner	ban (*which* b-class), gan (*which* g-class)
ICF	infl. marker, objekt focus	laa (*I*+obj. focus), la (*he*+ obj. foc)
IFUT	infl. marker ('di' future element)	dinaa (*I will*)
IINJ	optative infl. marker	nanga
IJ	interjection	déedéet (*no*)
INF	infl. marker, no focus marking	naa (*I* no foc.)
ISF	infl. marker, sent. focus	maa (*I* + sent. foc.)
ISuF	infl. marker, subj. focus	moo (*he* + subj. focus)
IVF	infl. marker, verb focus	dama (*I*+Verb focus)
NC	normal noun	téere (*book*)
NCF	normal noun with copula	njàngalee (*lesson* + obj. foc.)
NP	proper name	kiriku (*Kiriku*)
NPF	proper name with copula	yeeso (*Jesus* + subj. foc.)
NVPS	normal noun with possessor	doomu (*son of*), baayu (*father of*)
NU	ordinal/cardinal number	ñettel (*third*), ñetti (*three*)
PDEM	demonstrative pron.	lii (*this*)
PDMAT	dem. pron. with article	googu (*those* g-class)
PERS	free personal pron.	man (*I*), yéen (*you* pl.)
PIART	attributive indef. pron.	bépp (*any/every*)
PIS	substituting indef. pron.	ñépp (*everybody*)
PREL	substituting relative pron.	bi (*that/which/who*)
PRO	pronominal, obj.	ko (*him/her/it*)
PRO V3SG	possessive pron., third person singular	domaam (*his/her/its son*)
PRS	pronominal, subj.	ma (*I*), nga (*you* 2nd sg.)
PVPS	possessive pron.	sama (*my*)
PWQ	substituting interr. pron.	lu (*what*), lan (*what*)
PWQN	substituting interr. pron. (persons)	ku (*who*), kan (*who*)
REFL	reflexive	boppam (*himself*)
RFW	foreign lang. material	ràbbuni (*"my God"*)
U	particle	de (*well*)
UN	negation particle	dul (*not*)
UPL	('i'-) plural particle	i
UPSF	sent. focus particle	(maa) **ngi** (dem) (*(I'm) (going) here and now*)
URP	past tense particle 'woon'	woon
UVL	verb linking particle	a
VERS	impersonal verb form	dees (*it does*)
VMBZ	modal verb, base form	war (*shall*)
VMCC	modal verb, 'circumstantial' form	ware, mane
VMCR	modal verb, conditional past	manoon (*could*)
VMIMPE	modal verb, imperative	jëkkal (*begin with*)
VMNEG	modal verb, negative	bëggul (*won't*)
VMPV	modal verb, perfect	tàmbalee (*begun*)
VMRP	modal verb, remote past	waroon (*should*)
VNEIMP	negative imperative	buleen (*don't*)
VVBP	full verb, base form	def (*make*), lekk (*eat*)
VVCC	full verb, 'circumstantial' form	yónnee
VVCR	full verb, conditional past	amoon (*had*)
VVFP	full verb with particle (preposition or article)	taseek (*to meet with sth.*)
VVFUT	full verb, future	seedeeli (*will attest/confirm*)
VVHR	full verb, habitual past	joxaan (*usually gave*)
VVIMPE	full verb, imperative	toppal (*follow!*)
VVNEG	full verb, negative	xamul (*don't/doesn't know*), nekkul
VVPV	full verb, perfect	lekkee (*as 123 Pers SG/PL eat(s)*)
VVRP	full verb, remote past	toppoon (*followed*)
VX	imperfective auxiliary (present tense)	di (*do*)
VXAV	semi auxiliary	daldi (*do sth. instantly*)
VXCP	auxiliary, conditional present tense	dee (*would do*)
VXNEG	auxiliary, negative	du (*doesn't/don't*), duñu (*don't*)
VXR	auxiliary, remote past	doon (*did*)