Design and Development of Part-of-Speech-Tagging Resources for Wolof

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)

Wolof

- Spoken in Senegal
- Lingua franca for 80% of Senegals population (9 million speakers)
- 4 million native speakers
- West-Atlantic language

Wolof Language

- Complex system of inflectional markers/pronouns (almost no verbal inflection)

Ex. Object vs. Subjec focus

(1)

Maa	**lekk**	mburu.
FOC-Subj.1SG	eat	bread.

It was me who ate bread.

(2)

Mburu	**laa**	**lekk**.
Bread	FOC-Obj.1SG	eat.

It was bread that I ate.

- Very productive derivation morphology

Ex. Applicative

(3)

Togg-al	naa	xale	bi	ceeb.
Cook-APPL	1SG	child	DET	rice.

I cooked rice for the child.

Wolof Resources

- No NLP tools or resources available for Wolof!
- Linguistically quite well documented
  (some descriptive grammars, recent work on specific aspects of the grammar)
- Some online resources
  - Wolof Wikipedia: 1065 articles
    (Problem: inconsistent orthography)
- We used the Wolof Bible
  - Consistent orthography
  - Available as a parallel corpus (e.g. English,French, Arabic translations)

Motivation

Low resource languages are ...

- investigated in theoretical linguistics, annotated corpora are missing
  - University of Potsdam: research programme on information structure, NLP resources support corpus-based, cross-lingual investigations of of information structure
- a test-bed for NLP techniques existing for well-resourced languages
- often simulated by using small sets from well-resourced languages (e.g. in research on bootstrapping, unsupervised learning techniques, ...)

Starting from Scratch: Tagset Design

- No established Part-of-Speech inventory for Wolof (not even on the level of coarse-grained lexical categories)
  - Debate about adjectives in Wolof
- Inconsistent glosses/categorisations in the theoretical literature
  - Inconsistencies for verb categories
- What is the appropriate level of tagset granularity?
  - Should the tagset capture e.g. nominal classes?

Tagset Design: General Strategy

- General desiderata for a tagset:
  - Capture interesting linguistic categories
  - Be predictable/learnable for automatic taggers
- EAGLES guidelines, Leech and Wilson [1996]
- Interleaving tagset design and annotation experiments
- Distinguishing various granularity levels

Establishing Tagset Granularity

- Started out with fairly detailed tagset (200 tags)
- Experiments with tagset reductions
- Final "standard tagset" includes theoretically interesting distinctions that can be reasonably made by automatic PoS taggers

Granularity levels

Definite Articles	**Detailed 200 tags**	**Medium 44 tags**	**General 14 tags**	**Standard 80 tags**
SG/b-class/proximal	ATDs.b.P	ATDs	AT	ARTD
PL/y-class/remote	ATDp.y.R	ATDp	AT	ARTD
SG/b-class/sent. focus	ATDs.b.SF	ATDSF	AT	ARTF
SG/w-class/sent. focus	ATDs.w.SF	ATDSF	AT	ARTF

Interleaving Tagset Design and Annotation

PoS categories for Wolof verbs

Problem:

- theoretical work on Wolof establishes 3 verb finiteness categories: VVFIN, VVINF, VVNFN (Zribi-Hertz and Diagne [2002])
- automatic PoS-Taggers do not learn the distinction

Ten most frequent errors on tagset with 3 verb finiteness categories

(incorr.) system tag	gold tag	error ratio wrt. gold tag	tokens affected
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%

Interleaving Tagset Design and Annotation

PoS categories for Wolof verbs

Solution:

- one tag for overtly non-inflected verbs (VV)
- several fine-grained tags for token-internally inflected verbs (e.g. VN for negated verbs)

Ten most frequent errors made on tagset with 1 verb category

(incorr.) system tag	gold tag	error ratio wrt. gold tag	tokens affected
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%

Capturing Linguistically Interesting Categories

PoS categories for focus markers

- Standard tagset captures different focus types
- It should allow for corpus-based investigations of information structure
- Evaluate focus identification based on automatic tagging

Quality of automatic POS-based focus identification on 100 sentences

Focus Type	Evaluation		Abs.Freq in Test set	Abs. Freq in Corpus
	Precision	Recall		
Subject (ISuF)	95.65%	100%	39	1119
Verb (IVF)	100%	90%	11	759
Object (ICF)	68.75%	90.90%	11	910
Sentence (ISF)	100%	87.5%	16	635
				3423 focus instances (predicted)

Creating Gold Standard Data

- Annotated data: ca. 27,000 tokens from the New Testament
- Annotation effort: 1 month for 1 person
- Automatic pre-annotation reduced the effort (by more than 50%)
- Implementation includes:
  - Tokeniser and sentence splitter (based on the GATE environment)
  - Heuristics for stemming and lemmatising

Automatic Pre-Annotation

- generation of a full form lexicon based on ...
  - closed-class lexemes (1700 entries)
  - suffix-guessing for open-class lexemes (25000 entries)
- pre-annotated each token with all options found in the full form lexicon

Suffix guessing on entire corpus

(4) ... **gis**-leen !
... look !

"-leen" is an imperative suffix indicates a verbal category

add "gis" as a verb to the lexicon

Pre-annotation

(5) man de ab kanaara la fi **gis**.
"I can only see a turkey here."

↓

(6) man_PERS|DWQ de_IJ ab_ARTI kanaara_NC la_PRO|ICF|ARTD fi_AV **gis_VVBP**

Comparing State-of-the-art PoS Taggers

Can our gold standard data be used for training reliable automatic taggers?

1. TnT tagger: Brants [2000]
   trigram Hidden Markov model
   96.7% accuracy on NEGRA
2. TreeTagger: Schmid [1994]
   decision tree model
   96.06% on NEGRA
3. SVMTool: Giménez and Màrquez [2004]
   support vector machine classifier (very rich, lexical feature model)
   97.1% on the Wall Street Journal

Comparing State-of-the-art PoS Taggers

- Results from ten-fold cross-validation
- 26,846 training tokens
- 2650 test tokens
- average number of ambiguities: 5.173 per word (on fine-grained tagset)

	Accuracy			
Tagset size	**200**	**44**	**15**	**80**
Baseline	85.7%	88.4%	89.5%	87.6%
TnT	**92.7%**	94.2%	94.8%	**94.5%**
TreeTagger	90.7%	93.6%	94.5%	93.8%
SVM Tool	**93.1%**	**95.3%**	**96.2%**	**95.2%**

Comparing State-of-the-art PoS Taggers

- Results are comparable to state-of-the art (given the size of the training data)
- Standard tagset seems to be appropriate for automatic tagging
- Even the fine-grained tagset allows for quite accurate automatic analysis
- Open question: do these results scale to other text types?

	Accuracy			
Tagset size	**200**	**44**	**15**	**80**
Baseline	85.7%	88.4%	89.5%	87.6%
TnT	**92.7%**	94.2%	94.8%	**94.5%**
TreeTagger	90.7%	93.6%	94.5%	93.8%
SVM Tool	**93.1%**	**95.3%**	**96.2%**	**95.2%**

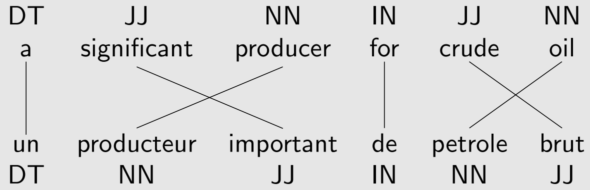
Conclusion

- Issues:
  - How to deal with under-studied, theoretically controversial phenomena?
  - How to satisfy theoretical and computational requirements on tagset design?
  - How to establish appropriate granularity of the tagset?
- Experience:
  - Even simple word lists are very useful for fast pre-annotation
  - Interleaving tagset design and annotation experiments
  - Automatic testing on different granularity levels

Towards Systematic Bootstrapping

- There is a lot of NLP research on bootstrapping resources for low resource languages (mostly "simulated")
- Classic: annotation projection paradigm, Yarowsky and Ngai [2001]
- Is it useful in a realistic scenario?

English-French projection example

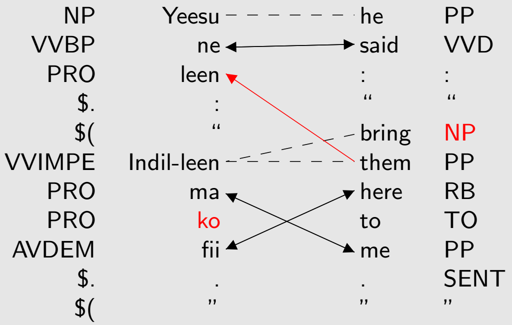


Crosslingual Projection Experiments

Added information from parallel corpus?

- Data seems very noisy for direction PoS projection
- English tagset cannot be directly adopted for Wolof, some manual annotation is required anyway
- "Light projection" scenario: use parallel PoS information as additional features in the training process

Wolof-English parallel example



Comparing Taggers with and without Parallel Information

- Results from HMM-Tagging, ten-fold cross-validation
- Parallel info based on GIZA word alignments
- English and French PoS annotation produced with TreeTagger

	Training data size (tokens)		
	418	1249	4968
no parallel information	59.7%	68.3%	82.7%
information from English	**62.6%**	70.2%	84.0%
information from English and French	**63.6%**	70.6%	84.1%

- Improvement only significant on smallest training set

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

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

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

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

David Yarowsky and Grace Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

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