A Morphological Analyzer for Wolof using Finite-State Techniques

Cheikh Bamba Dione

Department of Linguistic – Faculty of Humanities
University of Bergen

Colloquium on African Languages and Linguistics
November 22, 2011

CLARA

Outline

The Wolof Language

Salient typological features of Wolof:

- large number of phonemes (ca. 54 phonemes: 15 vowels, 39 consonants (stops,fricatives,nasal divided $\Rightarrow$ simple vs. strong)
- Wolof has ten canonical noun classes: 8 **singular** (***-b-, -g-, -j-, -k-, -l-, -m-, -s-, -w-***), 2 **plural** (***ñ, y***). Additionally, 3 special classes (***f-, n-, c-***) for locative, manner and prepositional indexes
- Focus in Wolof is marked morphologically by means of focus markers
- A complex system of verb suffixes coding valency changes
- Wolof is a clitic language: subject pronouns and agreement markers, object and locative clitics, imperfective markers are all clitics
- Absence of tone
- No passive: passive-similar phenomena are expressed as 1) medio-passive or 2) active constructions with impersonal 3PL subject
- No grammatical category for ADJ / no morphological case in Wolof / no gender-specific (pro)nouns

Wolof Morphology

A word in Wolof consists of stem + 1 or more affixes

- Affixes (prefix, suffix, infix, no circumfix) may contribute to the
  - syntactical meaning (tense, aspect, mood, subject, object)
  - lexical meaning (iterative, inversive,..)
- Nominal Morphology:
  1. Noun inflection: generally simple
  2. Noun derivation: complex $\Rightarrow$ compounding (noun-noun, verb-verb, noun-verb, verb-noun, etc. may use ideophones) and reduplication
- Verbal Morphology: agglutinative
  1. Verb inflection carried out by inflectional elements: Wolof verbs do not inflect (except few cases: past, conditional and negation)
  2. Verbal derivation: very complex $\Rightarrow$ uses a huge number of verbal suffixes

Wolof Nominal Morphology

Nominal Inflection

Nouns in Wolof are essentially inflected for Genitive (optionally followed by the nominal class), and Possessive 3rd person

(1) *Kër-**am** g-a mel ni kër-**u-g** buur.*
house-POSS.3SG CL-DIST liken COMPAR houseG-R-CLG kingB
'His house looks like a king's house.'

Wolof Nominal Morphology

Nominal derivation can occur in form of:

- Suffixation: 15 Suffixes, not always productive ⇒ seet "look" seet-u "mirror"
- Prefixation: ca. 7 prefixes *aji, al, ja, ma, maa, nja, waa*
- Consonant gradation/mutation (sometime predictable: f,s,d,g,j ⇒ p,c,nd, ng, nj)
- Combination of prefixation, suffixation and consonant mutation
- Compound:
  - Wolof has endocentric (*ndoxum taw* "rain water") and exocentric (*gaynde-géej* "shark") compounds as well.
  - Three forms of compounding: nominal, verbal or adverbial. The nominal compounding is the most frequently used form.
- Reduplication in Wolof is always total and
  - can be used for both noun and verb derivation
  - uses ideophonic stem (marginale derivation); *ñukk* ⇒ reduplicated ⇒ *ñukk-ñukk* "short steps run"
  - uses name of locations (i.e. *ndar* ⇒ *ndar-ndar*)

Wolof Verbal Morphology

Verbal Inflection

- verb base form (infinitive, PRES, FUT): Null morphem ⇒ *lekk-∅* "to eat"
- remote vs. habitual past: *-oon*, *-aan* ⇒ *lekk-aan* "used to eat"
- temporal conditional/perfect: ⇒ su lekk-*ee* "as/if s/he eats/has eaten"
- imperative:-*al*, *-leen* ⇒ *lekk-al* "eat!"
- Impersonal: presens and past *-ees*, *-eesoon*
- Negation: simple, not yet, anymore + subject agreement ⇒ *lekk-at-u-ma* "I don't eat anymore"

Wolof Verbal Morphology

Verbal Inflection

- verb base form (infinitive, PRES, FUT): Null morphem ⇒ *lekk-∅* "to eat"
- remote vs. habitual past: *-oon*, *-aan* ⇒ *lekk-aan* "used to eat"
- temporal conditional/perfect: ⇒ su lekk-*ee* "as/if s/he eats/has eaten"
- imperative:-*al*, *-leen* ⇒ *lekk-al* "eat!"
- Impersonal: presens and past *-ees*, *-eesoon*
- Negation: simple, not yet, anymore + subject agreement ⇒ *lekk-at-u-ma* "I don't eat anymore"

Verbal Derivation: regular or marginal

1. regular verb derivation
   - denominal verb derivation starts from a **nominal** root
   - deverbal verb derivation starts from a **verbal** root
   - ambivalente verb derivation starts from a **nominal/verbal** root
2. marginal verb derivation starts from an ideophonic root ⇒ ŋun-ideophonic-stem + i-verbalizer ⇒ ŋunŋun-i "to whisper"

Wolof Verbal Morphology

Linear ordering of Wolof verbal suffixes in verb derivation

I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
ar	e_1 i_1 ali anti andi at aan i_2	u oo	adi antu ante	andoo	aale	i_3 si	al_1	le lu	e_2	al_2	aat ati

Table: Template of verbal suffixes used in Wolof verb derivation

ar: effort, e_1: verbalizer, i_1: inversive , i_2: vebalizer , ali: completive, anti: corrective, andi: meanwhile at: intensive, aan: discontinuative, u:mediopassive, oo: together, adi: privative, antu: depreciative, ante: reciprocal, andoo: collective, aale: associative, i_3: go , si: come, al_1: causative, le: Participative, lu: causative, e_2: locative / instrumental, al_2: applicative, aat: iterative, ati: reiterative

Wolof Morphology: Pronouns

- **Strong pronouns** are full-fledged words: occur in positions otherwise open to lexical DP (dislocated, object focus, nonargument and P-governed), cannot occur as direct objects.
- Wolof **object and locative clitics (OLCs)** are used to mark 1) object, person, instrument or 2) locative, prepositional, distance.
- **Subject pronouns** have a syntactic distribution similar to their non-clitic counterparts (e.g. is predictable as they are specified for nominative case).

	Strong forms	Weak forms	
		subject pronouns	object pronouns & loc.
1sg	man	ma	ma
2sg	yow	nga	la
3sg	moom	mu	ko
1pl	**nun**	**nu**	**nu**
2pl	yeen	ngeen	leen
3pl	ñoom	ñu	leen
LOC - prep prox/dist	cii/caa		ci/ca
LOC distance - prox/dist	fii/faa		fi/fa

Table: Wolof Person and Locative Markers

Wolof Morphology: Inflectional Markers

- Depending on the construction, subject markers undergo morphological attachment to their left or to their right.
- Right-attached subject markers (i.e. ***nu-a***) are not proclitics. In sentence-initial position, they bear default initial stress, and undergo phonological coalescence with their rightward context.
- Left-attached subject markers (i.e. ***la-nu***) are unstressed.
- **Subject agreement:** subject markers are amalgam of PERS, NUM, MOOD, ASP, POL and FOCUS. All paradigms distinguish first, 2nd, 3rd person pl and sg.
- Verbal agreement is for **person** and **number** of the subject.

	1Sg	2Sg	3Sg	1Pl	2Pl	3Pl
SuF	*ma-a*	*ya-a*	*mu-a*	***nu**-a*	*yeen-a*	*ñu-a*
NSuF	*la-a*	*nga*	*la-∅*	*la-**nu***	*ngeen*	*la-ñu*
VF	*da-ma*	*da-nga*	*da-fa*	*da-**nu***	*da-ngeen*	*da-ñu*
Neut/Perf	*na-a*	*nga*	*na*	***na-nu***	*ngeen*	*na-ñu*
Neut/Impf	*di-naa*	*di-nga*	*di-na*	*di-na-**nu***	*di-ngeen*	*di-na-ñu*
Opt.	*na-a*	*na-nga*	*na-∅*	*na-**nu***	*na-ngeen*	*na-ñu*
Opt. Neg.	*bu-ma*	*bul*	*bu-mu*	*bu-**nu***	*bu-leen*	*bu-ñu*

Table: Subject agreement markers in Wolof

Wolof FST System

FST morphological analysis using Xerox finite state tool (fst)

1. two-level morphology: 1) a lower surface and 2) an upper or lexical level
2. Input: surface form is transformed into a lexical form (stem + morphosyntactic features)
3. Use of intermediate level
4. The tool handles the input in both directions: analysis and generation

Example

Task: Apply up ***fecceekuwaatoon*** "untied again" from ***fas***: "to tie"

Lexical: fas+V+Base+Inv+E+MPSV+Iter+PST

⇕ Lexicon + morphotactics

Intermediate: fas :i :e :u :aat :oon

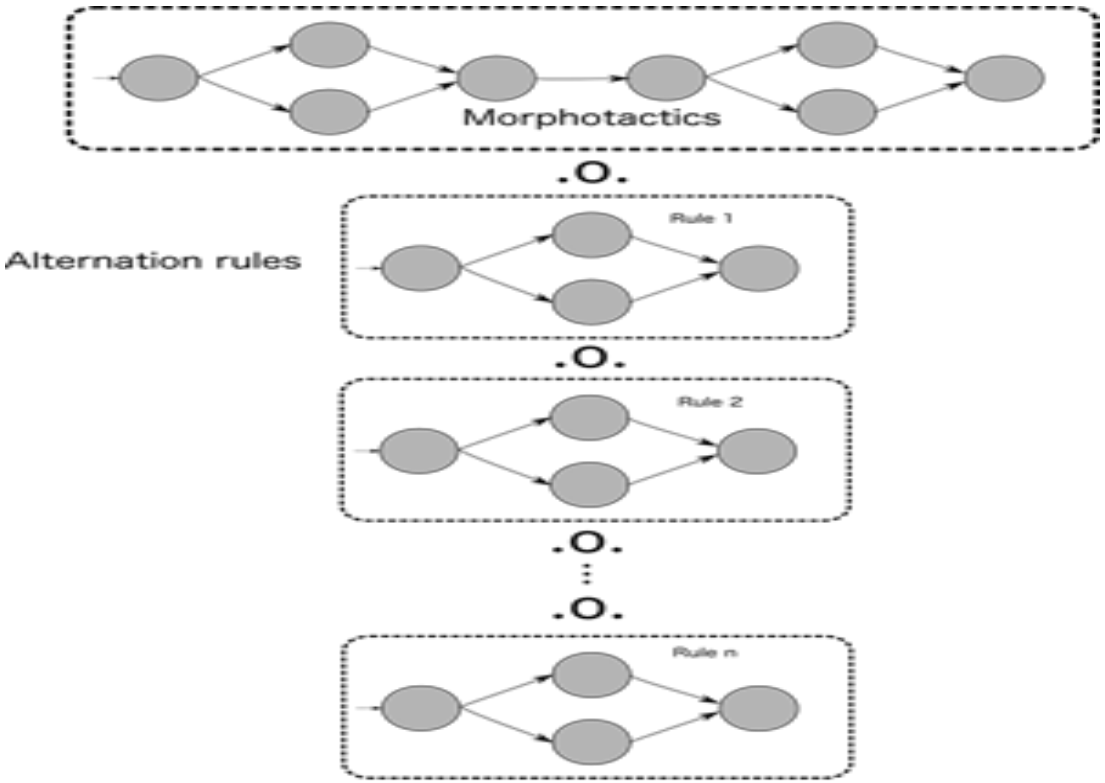
⇕ Orthographic rules

Surface: fecceekuwaatoon

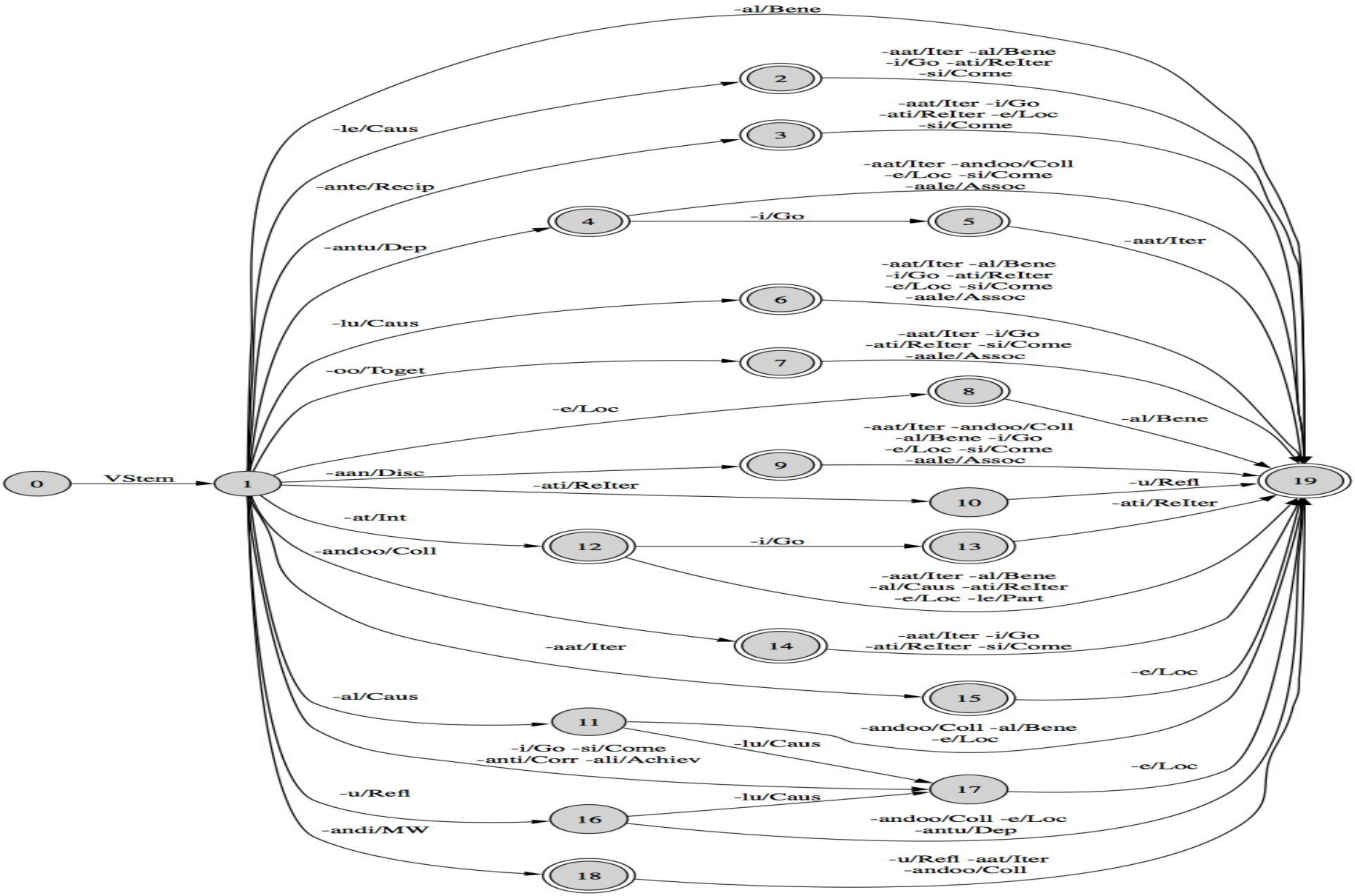
Morphological components

The components of the Wolof FST:

1. Lexicon: contains verbal and nominal stems, ideophone and closed classes
   - Statistics: common nouns (3800), proper nouns (1000), verbs (3500)
2. Morphotactics as **finite-state network** encoding the legal morphem. combination
3. Phonotactics as **finite-state transducers** describing the rules alternation
4. Composition of lexicon + phonotact. into a single network $\Rightarrow$ **lex. transducer**



Finite State Machine - Wolof Verb Derivation



Architecture Of The Verb Analyzer

- 'Verb': the lexical network that handles Wolof verbs is built up as a sequence of transducers

```
define Verb    [
                   VerbDerivationFSM
                   .o. vowelHarmony          # R10
                   .o. glideInsertion        # R1 Word Initial
                   .o. deleteRootVowel       # R11
                   .o. inversiveRule         # R5, R2, R4, R3
                   .o. consonEpenthesis1     # R14 'k' for medio-passive
                   .o. consonEpenthesis2     # R15 'j' for GO
                   .o. degemination          # R7
                   .o. deleteSuffixVowel     # R8
                   .o. glideInsertion2       # R6 stem final
                   .o. deleteVowel           # R12 stem final
                   .o. vowelCoalescence      # R9
               ]
```

Alternation Rules

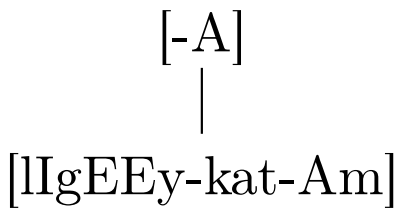
The harmony process is analyzed on the basis of four rules:

1. Morpheme Structure Constraint (MSC): a high vowel in **stem initial** position must be associated with the [+ATR] feature
2. Vowel Harmony Rule (VHR): the [+ATR] autosegment is spread from left to right to **all unassociated** vowel within a domain.
3. High Default Rule (HDR): all **non-linked high vowel** have to be specified as [+ATR]
4. Default Rule (DR): Every segment **left unassociated** must have the [-ATR] feature associated with it.

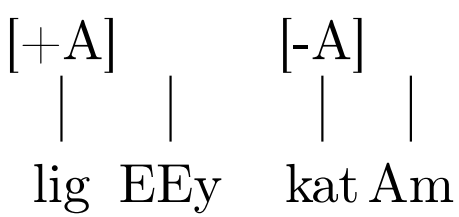
```
define vowelHarmony     [ MSC .o. harmonyRule .o. highDefaultRule .o. defaultRule];


define MSC              [I→i, U→u // .#. [SimpleCons|Prenasal|RootBoundary]* _ ];

define harmonyRule      [A→ë, E→é, O→ó, I→i, U→u // [i|u|ó|é|ë] [ConsSet]* _ ];

define highDefaultRule  [I→i, U→u // [O|A|E] ConsSet* _ ];

define defaultRule      [A→a, E→e, O→o // ConsSet* _ ];
```

(2) *lIgEEy-kat-Am* 'his/her worker'

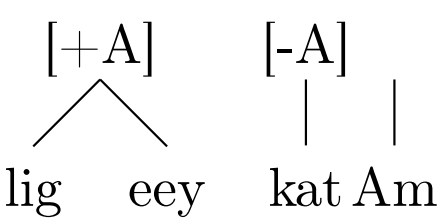
Underlying Rule



MSC

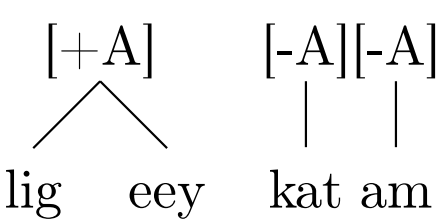


Harmony Rule



High Default N/A

Default Rule



ligeeykatam 'his/her worker'

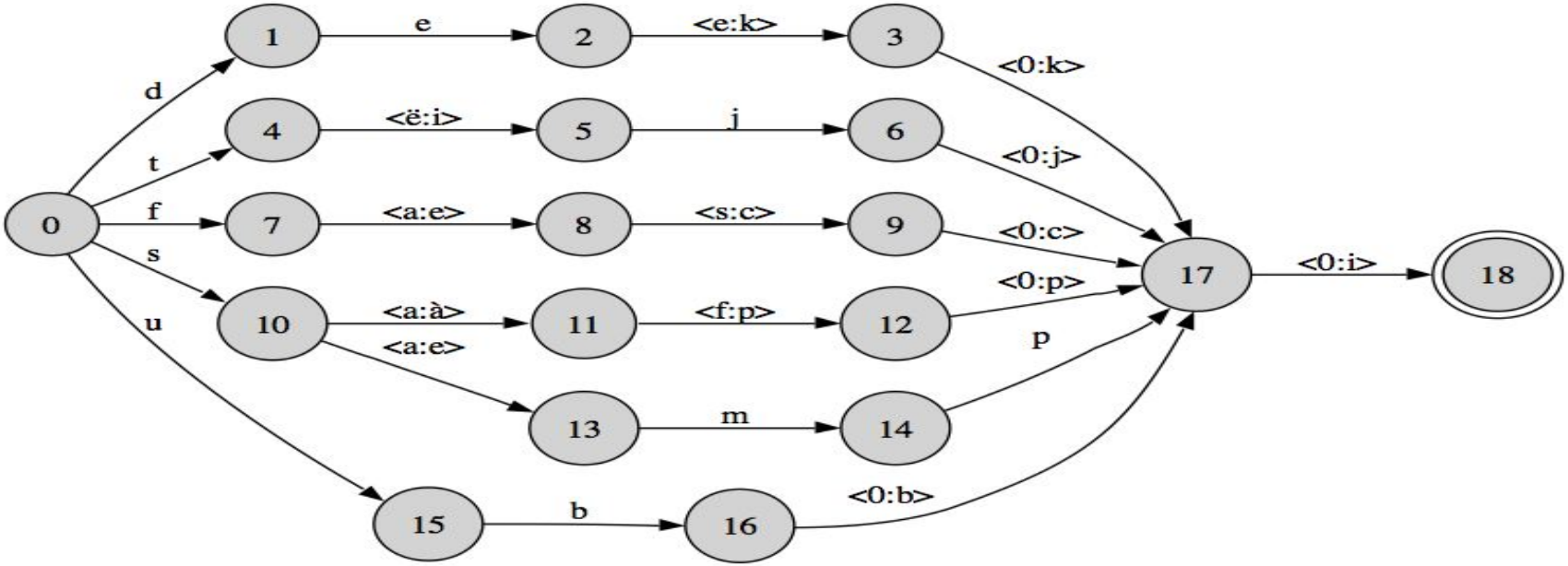
Inversive Formation

1. Gemination of the final consonant (GFC): *ub* ⇒ ***ubbi*** ' close/open'
2. Mutation and GFC: *saf* ⇒ *sa**ppi*** 'to be tasty/to lose taste'
3. Vowel gradation and GFC: *të**j*** ⇒ *t**ijji*** 'close/open'
4. Vowel gradation + mutation and GFC: *fas* ⇒ ***fecci*** 'tie/untie'
5. Vowel shortening + GFC: *suul* ⇒ ***sulli*** 'bury/exhume'
6. Vowel shortening + GFC: ***roof*** ⇒ ***roppi*** 'insert/extract'
7. Vowel gradation and shortening + mutation and GFC: *roof* ⇒ ***ruppi*** 'insert/extract'
8. Vowel shortening without GFC: *muur* ⇒ *m**u**ri* 'cover/uncover'
9. Vowel gradation and shortening without GFC: *yeew* ⇒ *y**i**wi* 'tie/untie'
10. Vowel gradation (word ending with geminate or prenasal): ***samp*** ⇒ ***sempi*** 'plant/plant out'
11. No vowel or consonant modification: *wekk* ⇒ *wekki* 'hang up/out'
12. Vowel shortening + consonant insertion: *dee* ⇒ *de**kki*** 'die/revive'

The Finite-State Transducer for the Inversive

Regularities for inversive, corrective and completive formation

- R2: mutation of the final consonant: the weak consonant becomes strong (e.g gemination process: -f→ -pp, -s→-cc, -r→-dd, 0→-kk)
- R3: inside of a root a long vowel should be short if it occurs before a strong consonant (e.g. geminate, prenasal)
- R4: vowel mutation ë→i , ó→u, a→à (strong consonant)
- R5: vowel mutation: a→e in case of inversive , a→o in case of corrective derivation



The Transducer for Inversive as FST Rule

The lexical network that handles the inversive derivation in Wolof

```
define Inversive    [
                        VerbRoot .o. vowelGradationHorizontale
                        .o. fortitionFinalCons .o. geminationFinalCons
                        .o. vowelGradationVerticale .o. vowelShortening
                    ]   InversiveSuffix;


define vowelGradHorizont    [   a→e || [ Cons | .#. ] _ StrongCons
                                .o. [ a→o || Nasal _ Cons .#.
                                .o. a→e || Fricatives _ Cons .#. ]
                            ];
define fortitionFinCons     [ f → pp , s→ cc, r→dd, s→q || _ .#. ];

define geminationFinCons    [ b→b b, c→c c, d→d d, ... || [ \SimpleCons | .#. ] _ .#. ];

define vowelGradVert        [ ë→i, ó→u, a→à || [ Cons | .#. ] _ StrongCons ];

define VowelShort           [ aa→à, ee→e, ... , uu→u || _ StrongCons ];
```

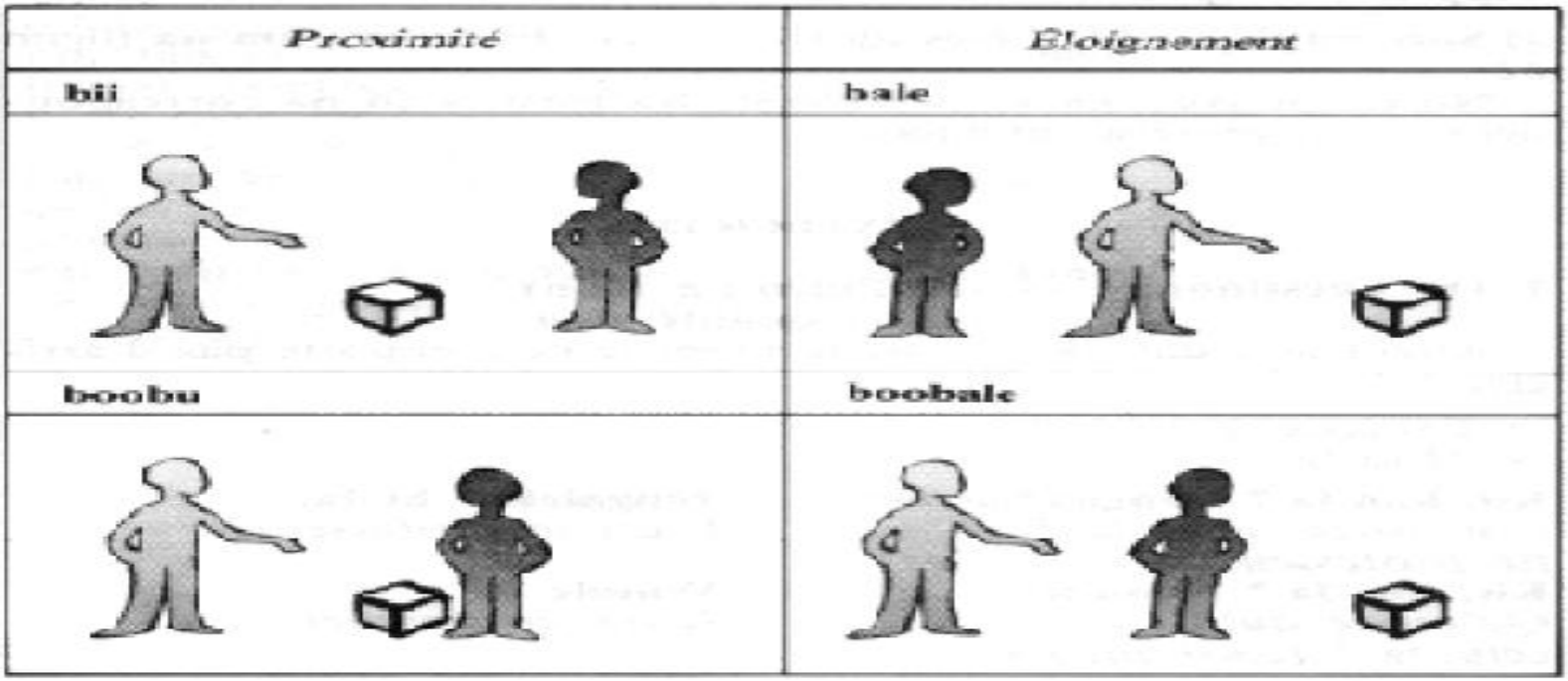
Long-distance Dependencies Using Flag Diacritics

- Flag diacritics are used to control long-distance dependencies (e.g. demonstratives) and constrain overgeneration in the network
- Recognized and applied at runtime ⇒ avoid size explosions
- Flag Diacritics are resolved before composing the lexical network

```
define ResolveVerbFD    [ "@R.POS.Noun@" "+N":0 | "@R.POS.Verb@" "+V":0 ]
                        [ "@R.VerbForm.Base@" "+Base":0 | "@D.VerbForm@" ]
                        [ "@R.Inversive.+@" "+Inv":0| "@D.Inversive@"]
                        [ "@R.CausativeDirect.+@" "+LU":0 | "@D.CausativeDirect@"]
                        [ "@R.CausativeIndirect.+@" "+LOO":0 | "@D.CausativeIndirect@"
                        [ "@R.PossessiveLE.+@" "+LE":0 | "@D.PossessiveLE@"]
                        [ "@R.CausativeE.+@" "+E":0 | "@D.CausativeE.+@"]
                        [ "@R.CausativeAl.+@" "+AL":0 | "@D.CausativeAl@"]
                        [ "@R.ApplicativeAL.+@" "+AL":0 | "@D.ApplicativeAL.+@"]
                        [ "@R.MedioPassive.+@" "+MPSV":0 | "@D.MedioPassive@"]
                        ...
                        ;
```

Wolof Demonstrative

Wolof demonstrative specifier encode deixis (proximal/distal) and reference [referent, distance to speaker/hearer])



Deixis				Reference					
proximal		distal		proximal			distal		
Cii	Cile	Cee	Cale	CooCii	CooCu	CooCule	CooCa	CooCale	CooCee

Table: Wolof Demonstratives: 130 forms

Grammatical Categories

Categories handled by the morphological analyzer

1. Nouns (N): common, proper, inflected and derived
2. Verbs (V): inflected and derived V (denominal, deverbal, marginal)
3. Auxiliar (AUX): inflected for aspect/tense
4. Inflectional markers (INFL): subject/verb/complement focus, optative, aspect, imperative
5. Clitics (Cl): Subject agreement markers (Cl_{subj}), tense (Cl_{tns}), object and locative (Cl_{obj})/(Cl_{loc})
6. Pronouns (PRON): personal, subject pronouns, relative, free relative, quantitative, locative, demonstrative, interrogatives, possessive
7. Specifiers: Determiners (def, indef, rel, int), Demons (including deixis and reference), quantifiers, numeral
8. Adverbs (ADV): temporal, locative, manner, standard.
9. Complementizers (COMP): standard, interrogative

Evaluating The Wolof Morphological Analyzer

- The evaluation is performed using the Xerox lookup utility, a runtime program that applies pre-compiled transducers to look up words.
- 4 different strategies: the single normal FST (strategy 0), capitalization (strategy1), word normalization (strategy2) and allowing the relaxation of accentuation/vowel harmony (strategy3)
- Achieved accuracy: **79.55%**

Not found: 805 words
Foreign words: 175; Proper nouns: 63; spelling errors: 60

Strategy	Frequency	Accuracy
strategy 0	8463 times	70.51%
strategy 1	1000 times	8.33%
strategy 2	6 times	0.05%
strategy 3	79 times	0.66%
not found:	2455 times	20.45%
corpus size:	12033 tokens	

Table: Lookup accuracy scores on the Wolof Wikipedia