Learning morphology and phonology

John Goldsmith
University of Chicago
MoDyCo/Paris X

Learning morphology and phonology

John Goldsmith
University of Chicago
MoDyCo/Paris X

All the particular properties that give a language its unique phonological character can be expressed in numbers.

-Nicolai Trubetzkoy, *Grundzüge der Phonologie*

Acknowledgments

My thanks for many conversations to Aris Xanthos, Yu Hu, Mark Johnson, Carl de Marcken, Bernard Laks, Partha Niyogi, Jason Riggle, Irina Matveeva, and others…

Roadmap

1. Unsupervised word segmentation
2. MDL: Minimum Description Length
3. Unsupervised morphological analysis
   Model; heuristics.
4. Elaborating the morphological model
5. Improving the phonological model: categories:
   - consonants/vowels
   - vowel harmony
6. What kind of linguistics is this?

0. Why mathematics? Why phonology?

One answer: mathematics provides an alternative to *cognitivism*, the view that linguistics is a cognitive science.

Cognitivism is the latest form, in linguistics, of *psychologism*, a view that has faded in and out of favor in all of the social sciences for the last 150 years: the view that the way to understand *x* is to understand how people analyze *x*.

1. *introduction*

- This work provides an answer to the challenge: if *linguistics* is not a science of what does on in a speaker’s head, then what is it a science *of*?

1. Word segmentation

The inventory of words in a language is a major component of the language, and very little of it (if any) can be attributed to universal grammar, or be viewed as part of the essence of language.

So how is it learned?

1. Word segmentation

Reporting work by Michael Brent and by Carl de Marcken at MIT in the mid 1990s.

Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!

Blah blah, Ginger! Blah blah blah blah blah blah Ginger blah blah blah blah blah blah blah…

1983
What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah…

1. Word segmentation

- **Strategy**: We *assume* that a speaker has a lexicon, with a probability distribution assigned to it; and that the parse assigned to a string is the parse with the greatest probability.
- That is already a (partial) hypothesis about word-parsing: given a lexicon, choose the parse with the greatest probability.
- It can also serve as part of a hypothesis about lexicon-selection.

1. Word segmentation

Assume an alphabet A.

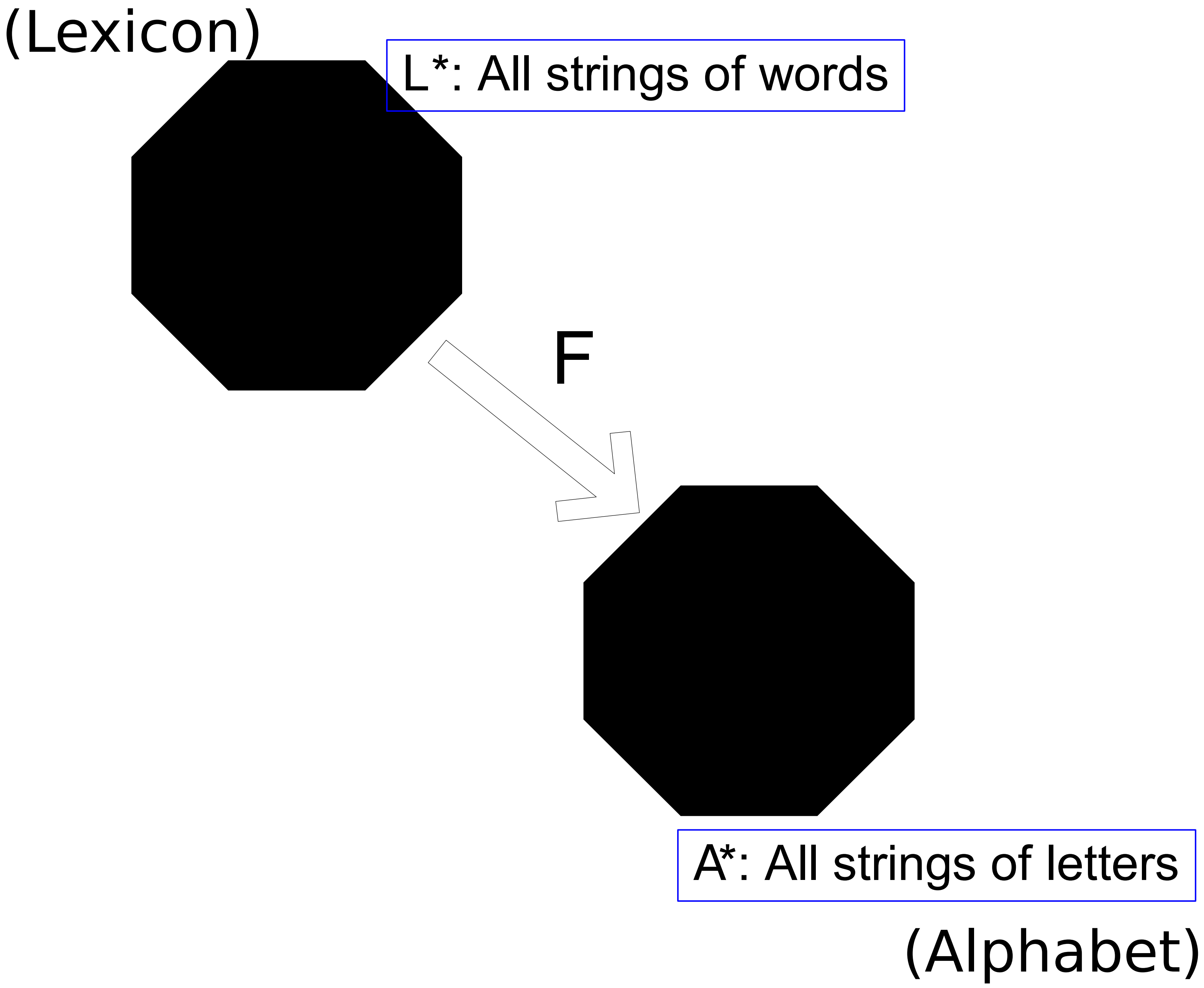
An *utterance* is a string of letters chosen from A *; a *corpus* is a set of utterances.

Language *model* used: multigram model (variable length words).

A *lexicon L* is a pair of objects (L, p_L): a set L ⊂ A *, and a probability distribution p_L that is defined on A* for which L is the support of p_L. We call L the *words.*

- **We insist that A ⊂ L: all individual letters are words.**
- **We define a *sentence* as a member of L*.**
- **Each sentence can be uniquely associated with an utterance (an element in A *) by a mapping F:**

1. Word segmentation



(Lexicon)

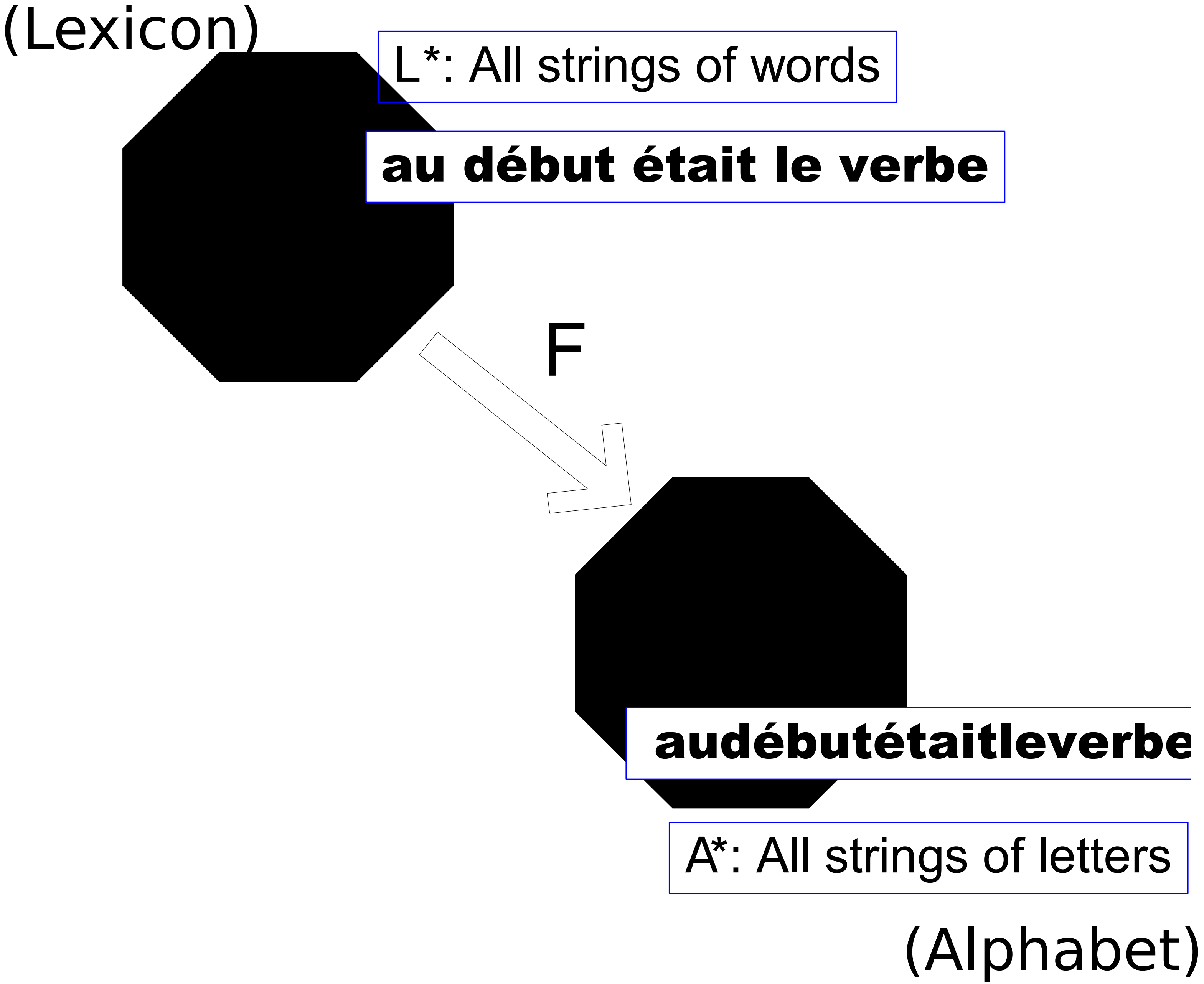
L*: All strings of words

F

A*: All strings of letters

(Alphabet)

1. Word segmentation



(Lexicon)

L*: All strings of words

au début était le verbe

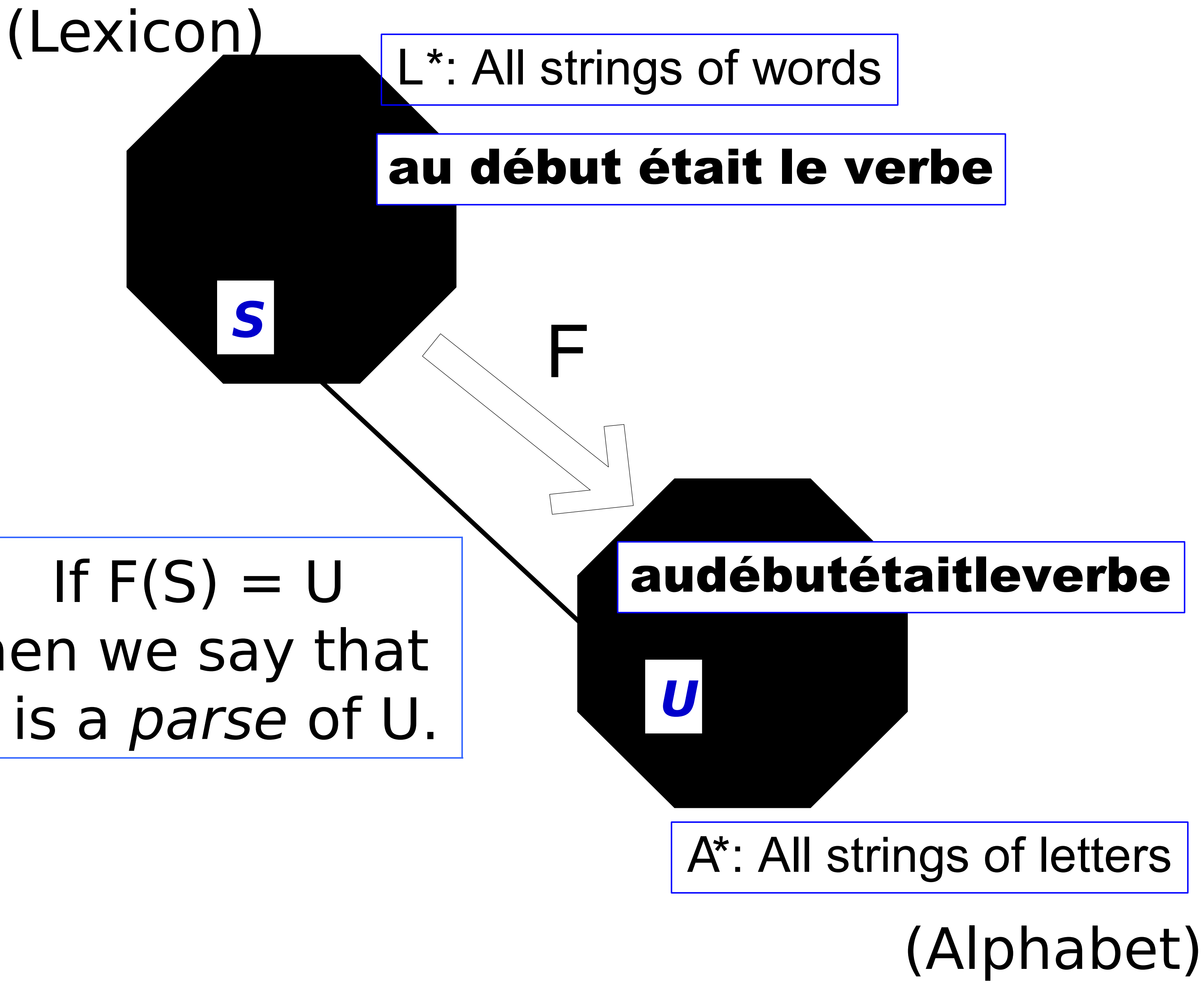
F

audébutétaitleverbe

A*: All strings of letters

(Alphabet)

1. Word segmentation



(Lexicon)

L*: All strings of words

au début était le verbe

S

F

audébutétaitleverbe

U

A*: All strings of letters

(Alphabet)

If F(S) = U
then we say that
S is a *parse* of U.

1. Word segmentation

- The distribution p over L is extended to a distribution p* over L* in the natural way:
  - We assume a probability distribution λ over sentence length *l*:

$$\sum_{i=1} \lambda(i) = 1$$

- If S is a sentence of length *l*=|*S*|, then

$$p^*(S) = \lambda(l) \prod_{i=1}^{l} p(S[i])$$

Now we can define the probability of a corpus, given a lexicon

- U is an utterance; L, a lexicon.

$$p(U \mid L) = \underset{q \in \{parses(U)\}}{\arg\max}\ pr(q)$$

You might think it should be the sum of the probabilities of the parses of U.

$$p(U \mid L) = \sum_{q \in \{parses(U)\}} pr(q)$$

That would be reasonable.

Calculating either argmax or sum requires dynamic programming techniques.

1. Word segmentation

Best lexicon for a corpus U?

You might expect that the *best lexicon* for a corpus would be the lexicon that assigns the highest probability to the joint object which is the corpus C:

$$\overset{\sqcup}{\mathrm{L}} = \underset{L \in \mathrm{A}^*,\, \mathrm{pr}}{\arg\max}\; pr_L(C \mid L)$$

But ***no***: such a lexicon would simply be all the members of the corpus. A sentence is its own best probability model.

2. Minimum Description Length (MDL) analysis

MDL is an approach to statistical analysis that assumes that prior to analyzing any data, we have a universe of possible models (= UG); each element $G \in UG$ is a probabilistic model for the set of possible corpora; and

A prior distribution $\pi(\ )$ has been defined over UG based on the length of the shortest binary encoding of each G, where the encoding method has the prefix property: $\pi(G) = 2^{-\mathrm{length}(En(G))}$

2.1 Bayes' rule

$$pr(G \mid C) = \frac{pr(C \mid G)\, pr(G)}{pr(C)}$$

$$= \frac{p_G^*(C)\pi(G)}{pr(C)}$$

$$= \frac{p_G^*(C)\pi(G)}{\int_{UG} p_g^*(C)\pi(g)dg}$$

2. MDL

$$pr(G \mid C) = \frac{pr(C \mid G)\,pr(G)}{pr(C)}$$

$$= \frac{p_G(C)\pi(G)}{pr(C)}$$

$$= \frac{p_G(C)\pi(G)}{\int p_g(C)\pi(g)dg}$$

$$\boxed{\begin{aligned}&\log pr(G \mid C)\\ &= \boxed{\log p_G(C)} - \boxed{H(G)} - K.\end{aligned}}$$

log prob of corpus, in grammar G

Length of G’s encoding

2. MDL

$$\boxed{\log pr(G \mid C) = \boxed{\log p_G(C)} - \boxed{H(G)} - K.}$$

log prob of corpus, in grammar G

Length of G’s encoding

We already figured out how to compute this, given G=(L,p)

$$\boxed{\|G\| \approx \sum_{w \in G} |w| * \log(26)}$$

How one talks in MDL...

It is sensible to call –log prob (X) = $\log(\frac{1}{prob\ x})$ the ***information content*** of an item X, and also to refer to that quantity as the ***optimal compressed length*** of X.

In light of that, we can call the following quantity the *description length of corpus C, given grammar G*:

$$\left[-\log prob\,(C\,|\,G)\right] + [length(Enc(G))]$$

= Compressed length of corpus

+ compressed length of grammar

= -log prob (G|C) + a constant

2. MDL

How one talks in MDL…

It is sensible to call –log prob (X) = $\log(\frac{1}{prob\ x})$ the ***information content*** of an item X, and also to refer to that quantity as the ***optimal compressed length*** of X.

In light of that, we can call the following quantity the description length of corpus C, given grammar G:

$$[-\log prob(C|G)] + [length(Enc(G))]$$

= Compressed length of corpus

+ compressed length of grammar

= -log prob (G|C) + a constant

= evaluation metric of early generative grammar

MDL dialect

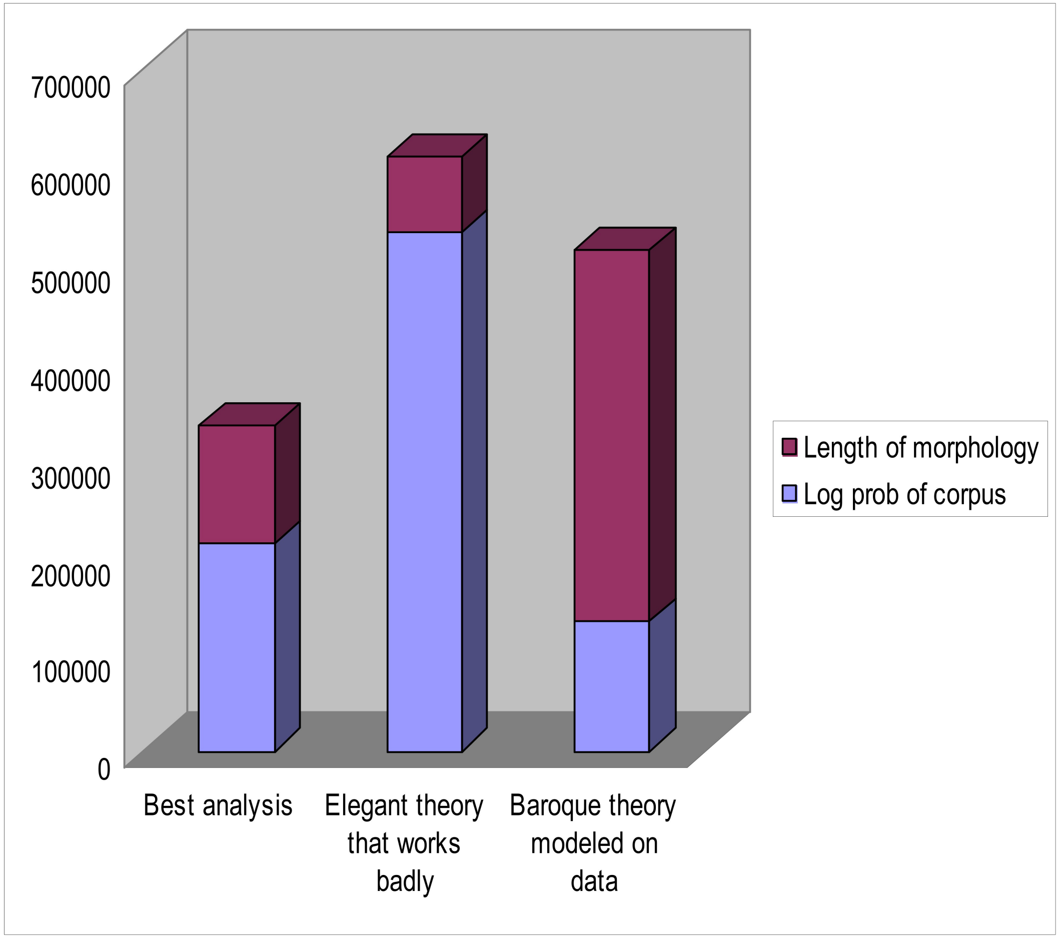
- MDL analysis: find the grammar G for which the total description length is the smallest:

Compressed length of data, given G +

Compressed length of G

2. MDL

Essence of MDL



2.2 Search heuristic

Easy!

start *small:* initial lexicon = A;

if l_1 and l_2 are in L, and $l_1.l_2$ occurs in the corpus, add $l_1.l_2$ to the lexicon if that modification decreases the description length.

Similarly, remove l_3 from the lexicon if that decreases the description length.

MDL: tells us when to stop growing the lexicon

If we search for words in a bottom-up fashion, we need a criterion for when to stop making bigger pieces.

MDL plays that role in this approach.

A little example to fix ideas…

How do these two multigram models of English compare? Why is Number 2 better?

Lexicon 1: {a,b,…s,t,u…z}

Lexicon 2: {a,b,…s,t,th,u…z}

2. MDL

A little example to fix ideas...

Notation:

[t] = count of *t*

[h] = count of *h*

[th] = count of *th*

Z = total number of words (tokens)

$$Z = \sum_{l \in lexicon} [l]$$

Log probability of corpus:

$$\sum_{m \text{ in } lexicon} [m] \log \frac{[m]}{Z}$$

2. MDL

$$\sum_{m \text{ in lexicon}} [m] \log \frac{[m]}{Z}$$

$$\text{where } Z = \sum_{l \in lexicon} [l]$$

Log prob of sentence C

$$[t]_1 \log \frac{[t]_1}{Z_1}$$

$$+[h]_1 \log \frac{[h]_1}{Z_1}$$

$$+\sum_{m \neq t,h} [m] \log \frac{[m]}{Z_1}$$

All letters are separate

$$[t]_2 \log \frac{[t]_2}{Z_2}$$

$$+[h]_2 \log \frac{[h]_2}{Z_2}$$

$$+\sum_{m \neq t,h} [m] \log \frac{[m]}{Z_2}$$

$$+[th]_2 \log \frac{[th]_2}{Z_2}$$

th is treated as a separate chunk

$$[t]_2 = [t]_1 - [th]$$

$$[h]_2 = [h]_1 - [th]$$

$$[Z]_2 = [Z]_1 - [th]$$

$$[t]_1 \log \frac{[t]_1}{Z_1} + [h]_1 \log \frac{[h]_1}{Z_1} + \sum_{m \neq t,h} [m] \log \frac{[m]}{Z_1}$$

All letters are separate

$$[t]_2 \log \frac{[t]_2}{Z_2} + [h]_2 \log \frac{[h]_2}{Z_2} + \sum_{m \neq t,h} [m] \log \frac{[m]}{Z_2} + [th]_2 \log \frac{[th]_2}{Z_2}$$

th is treated as a separate chunk

$$define\ \Delta f\ as\ \log \frac{f_2}{f_1}; then\ \Delta pr(C) =$$

$$\boxed{-Z_1 \Delta Z + [t]_1 \Delta t + [h]_1 \Delta h + [th] \log \frac{pr_2(th)}{pr_2(t)\, pr_2(h)}}$$

This is **positive** if Lexicon 2 is better

2. MDL

Effect of having fewer “words” altogether

$$define\ \Delta f\ as\ \log \frac{f_2}{f_1};\ then\ \Delta pr(C) =$$

$$\boxed{\boxed{-Z_1 \Delta Z} + [t]_1 \Delta t + [h]_1 \Delta h + [th] \log \frac{pr_2(th)}{pr_2(t)\, pr_2(h)}}$$

This is **positive** if Lexicon 2 is better

Effect of frequency of /t/ and /h/ decreasing

$$\textit{define } \Delta f \textit{ as } \log \frac{f_2}{f_1} ; \textit{then } \Delta pr(C) =$$

$$-Z_1 \Delta Z + [t]_1 \Delta t + [h]_1 \Delta h + [th] \log \frac{pr_2(th)}{pr_2(t)\, pr_2(h)}$$

This is **positive** if Lexicon 2 is better

2. MDL

Effect /th/ being
treated as a unit
rather than separate pieces

$$define\ \Delta f\ as\ \log\frac{f_2}{f_1}; then\ \Delta pr(C) =$$

$$\boxed{-Z_1\Delta Z + [t]_1\Delta t + [h]_1\Delta h + \boxed{[th]\log\frac{pr_2(th)}{pr_2(t)\,pr_2(h)}}}$$

This is **positive** if
Lexicon 2 is
better

2.3 Results

- The Fulton County Grand Ju ry s aid Friday an investi gation of At l anta 's recent prim ary e lection produc ed no e videnc e that any ir regul ar it i e s took place .
- Thejury further s aid in term - end present ment s thatthe City Ex ecutive Commit t e e ,which had over - all charg e ofthe e lection , d e serv e s the pra is e and than k softhe City of At l anta forthe man ner in whichthe e lection was conduc ted.

Chunks are too big Chunks are too small

Summary

1. Word segmentation is *possible*, using (1) variable length strings (*multigrams)*, (2) a probabilistic model of a corpus and (3) a search for maximum likelihood, if (4) we use MDL to tell us when to stop adding to the lexicon.
2. The results are *interesting*, but they suffer from being incapable of modeling real linguistic structure beyond simple chunks.

Summary

1. Word segmentation is *possible*, using **(1) variable length strings (*multigrams*), (2) a probabilistic model of a corpus** and (3) a search for maximum likelihood, if (4) we use MDL to tell us when to stop adding to the lexicon.
2. The results are *interesting*, but they suffer from being incapable of modeling real linguistic structure beyond simple chunks.

Question:

Will we find that *types* of linguistic structure correspond naturally to *ways* of improving our MDL model, either to *increase the probability of the data*, or to *decrease the size of the grammar*?

3. Morphology (*primo*)

Problem: *Given* a set of words, find the *best* morphological structure for the words – where “best” means it maximally agrees with linguists (where they agree with each other!).

Because we are going from *larger* units to *smaller* units (words to morphemes), the probability of the data is certain to *decrease*.

The improvement will come from drastically shortening the grammar = discover regularities.

Naïve MDL

Corpus:

jump, jumps, jumping

laugh, laughed, laughing

sing, sang, singing

the, dog, dogs

total: **62** letters

Analysis:

Stems: jump laugh sing sang dog (20 letters)

Suffixes: s ing ed (6 letters)

Unanalyzed: the (3 letters)

total: **29** letters.

Model/heuristic

1st approximation: a morphology is:

1. a list of *stems,*
2. a list of affixes (prefixes, suffixes), and
3. a list of *pointers* indicating which combinations are permissible.

Unlike the word segmentation problem, now we have *no obvious search heuristics.*

These are very important (for that reason)—and I will not talk about them.

Size of model

M[orphology] =

{ Stems T, Affixes F, Signatures Σ }

$$\|M\| = \|T\| + \|F\| + \|\Sigma\|$$

stems $$\|T\| = \sum_{t \in T} \|t\|$$

affixes $$\|F\| = \sum_{f \in F} \|f\|$$

sig's $$\|\Sigma\| = \sum_{\sigma \in T} \|\sigma\|$$

$$\|s\| = string\ \ length(s) * \log(26)$$

$$or = \sum_{i=1}^{|s|} \|s[i]\| = \sum_{i=1}^{|s|} -\log freq\ (s\,[i])$$

What is a signature, and what is its length?

extensivity

What is a signature?

$$\left\{\begin{matrix} account \\ appeal \\ attack \\ 40\ more... \end{matrix}\right\}\left\{\begin{matrix} NULL \\ ed \\ ing \end{matrix}\right\}$$

$$\left\{\begin{matrix} élevé \\ équipé \\ étonnant \\ 78\ more \end{matrix}\right\}\left\{\begin{matrix} NULL \\ e \\ s \\ es \end{matrix}\right\}$$

What is the *length* (=information content) of a signature?

A signature is an ordered pair of two sets of pointers: (i) a set of pointers to stems; and (ii) a set of pointers to affixes.

The length of a pointer *p* is –log freq (*p*):

So the total length of the signatures is:

$$\sum_{\sigma \in Sigs} \boxed{\left(\sum_{t \in Stems(\sigma)} \boxed{\log \frac{[W]}{[t]}} + \sum_{f \in Suffixes(\sigma)} \boxed{\log \frac{[\sigma]}{[f\, in\, \sigma]}} \right)}$$

Sum over signatures

Sum over stem ptrs

Generation 1 *Linguistica*

http://linguistica.uchicago.edu

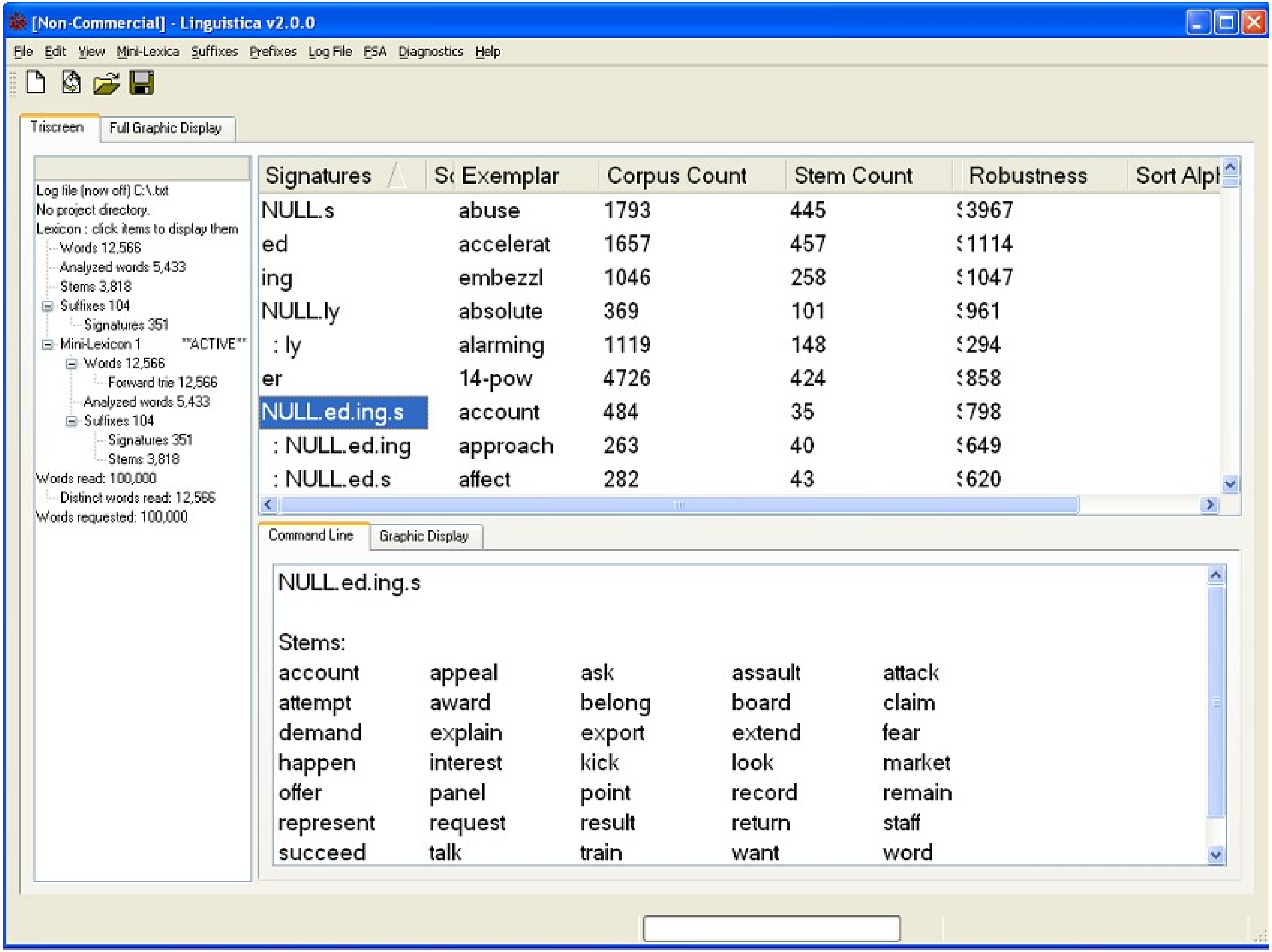
Initial pass:

- assumes that words are composed of 1 or 2 morphemes;
- finds all cases where signatures exist with at least 2 stems and 2 affixes:

$$\left\{\begin{matrix} jump \\ walk \end{matrix}\right\}\left\{\begin{matrix} NULL \\ ed \\ ing \end{matrix}\right\}$$

Generation 1

Then it refines this initial approximation in a large number of ways, always trying to decrease the description length of the initial corpus.



Log file (now off) C:\.txt
No project directory.
Lexicon : click items to display them
Words 12,566
Analyzed words 5,433
Stems 3,818
Suffixes 104
Signatures 351
Mini-Lexicon 1 **ACTIVE**
Words 12,566
Forward trie 12,566
Analyzed words 5,433
Suffixes 104
Signatures 351
Stems 3,818
Words read: 100,000
Distinct words read: 12,566
Words requested: 100,000

Signatures	S(	Exemplar	Corpus Count	Stem Count	Robustness	Sort Alpl
NULL.s		abuse	1793	445	3967	
ed		accelerat	1657	457	1114	
ing		embezzl	1046	258	1047	
NULL.ly		absolute	369	101	961	
: ly		alarming	1119	148	294	
er		14-pow	4726	424	858	
NULL.ed.ing.s		account	484	35	798	
: NULL.ed.ing		approach	263	40	649	
: NULL.ed.s		affect	282	43	620	

NULL.ed.ing.s

Stems:

```
account     appeal      ask         assault     attack
attempt     award       belong      board       claim
demand      explain     export      extend      fear
happen      interest    kick        look        market
offer       panel       point       record      remain
represent   request     result      return      staff
succeed     talk        train       want        word
```

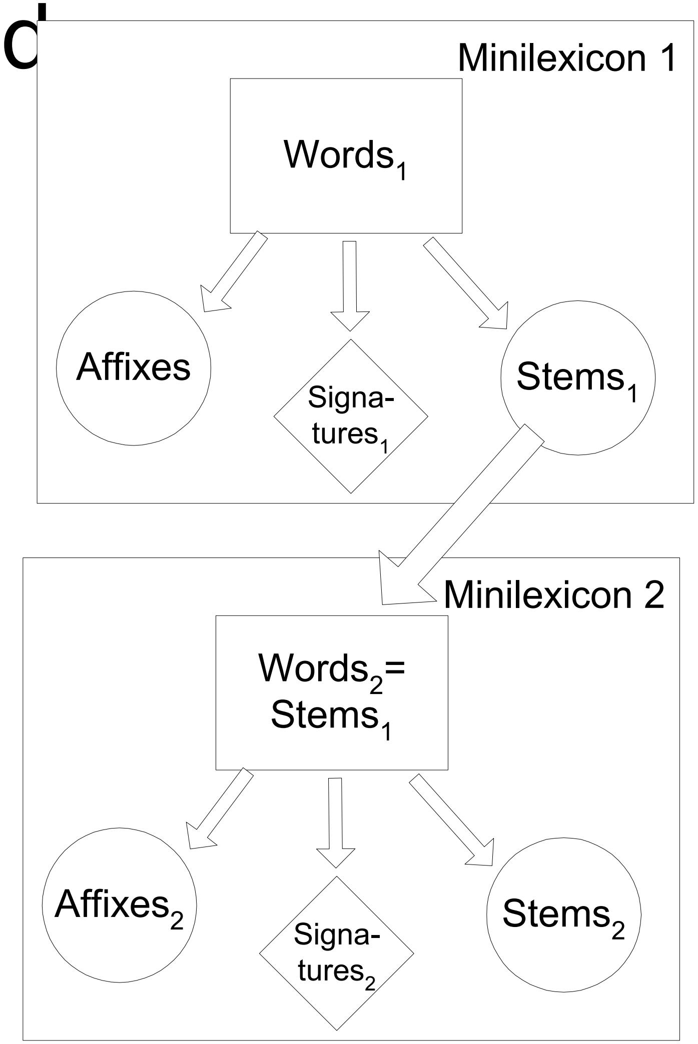
Refinements

1. Correct errors in segmentation

$$\begin{Bmatrix} \mathit{affirmati} \\ \mathit{aggressi} \\ \mathit{attenti} \\ 20\ \mathit{more} \end{Bmatrix}\begin{Bmatrix} \mathit{on} \\ \mathit{ve} \end{Bmatrix} \Rightarrow \begin{Bmatrix} \mathit{affirm} \\ \mathit{aggress} \\ \mathit{attent} \\ 20\ \mathit{more} \end{Bmatrix}\begin{Bmatrix} \mathit{ion} \\ \mathit{ive} \end{Bmatrix}$$

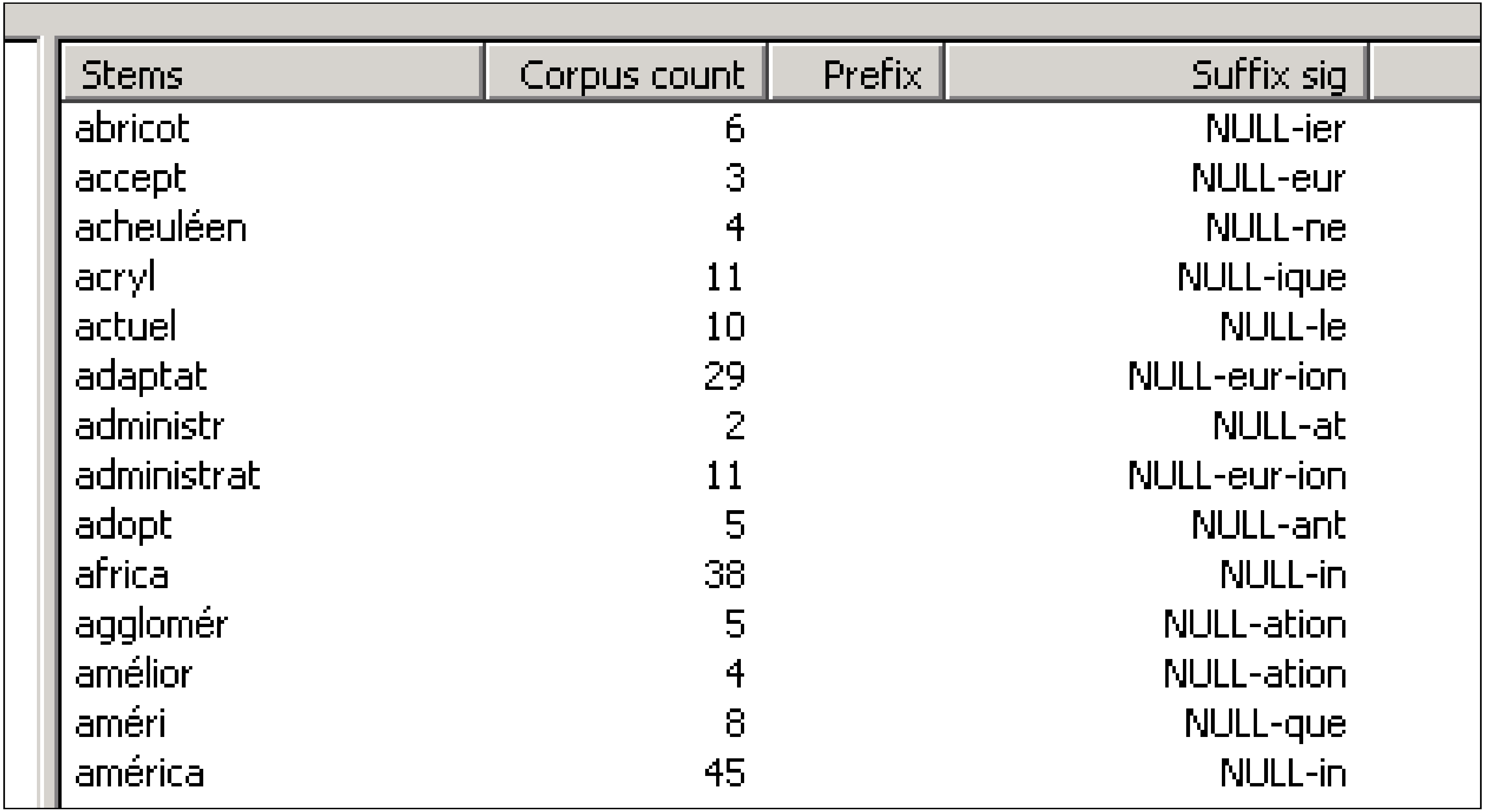
2. Create signatures with only one observed stem: we have *NULL*, *ed*, *ion*, *s* as suffixes, but only one stem (*act*) with exactly those suffixes.

3. Find recursive structure: allow stems to be analyzed

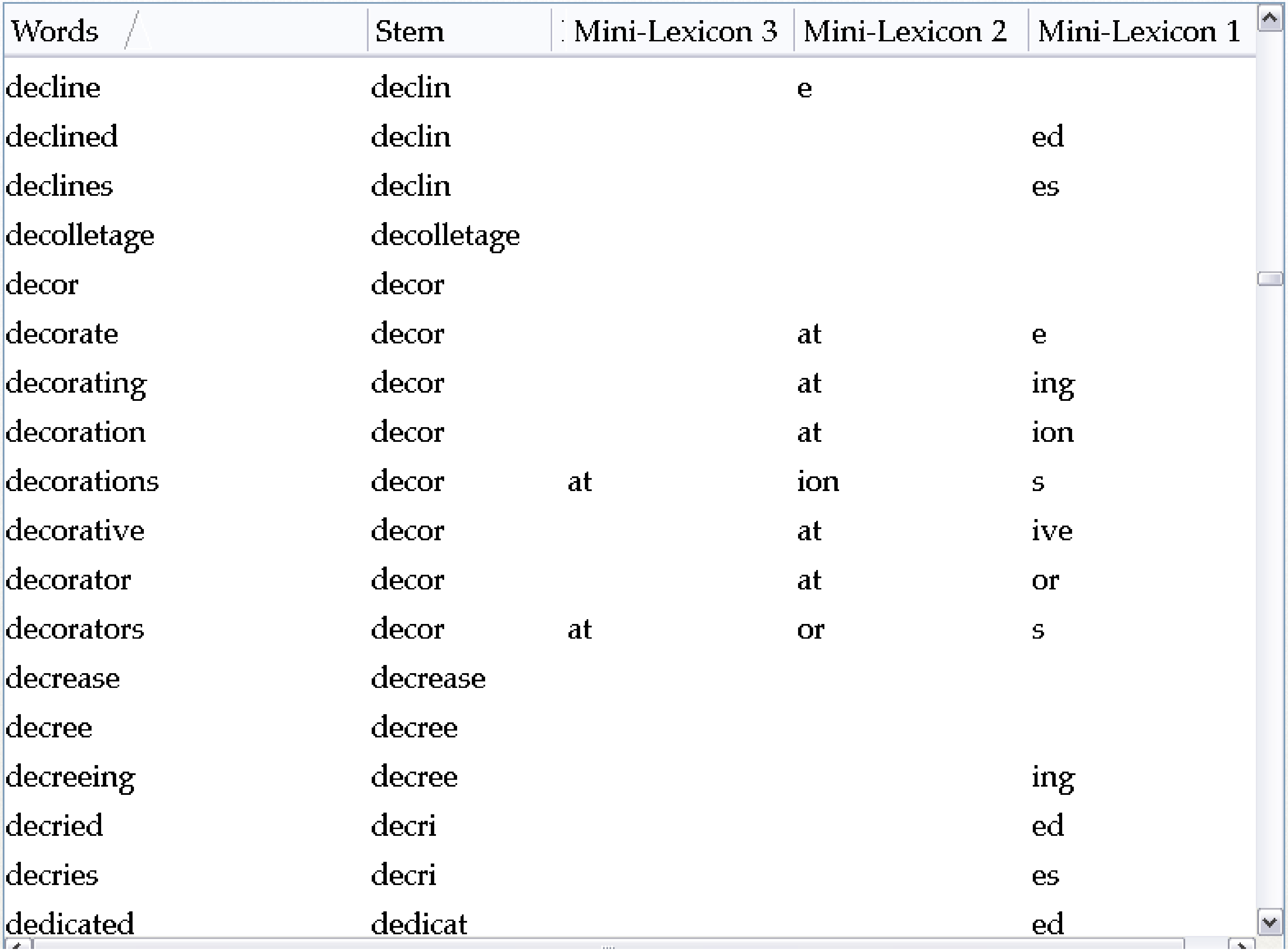


3. Morphology

French roots



Stems	Corpus count	Prefix	Suffix sig
abricot	6		NULL-ier
accept	3		NULL-eur
acheuléen	4		NULL-ne
acryl	11		NULL-ique
actuel	10		NULL-le
adaptat	29		NULL-eur-ion
administr	2		NULL-at
administrat	11		NULL-eur-ion
adopt	5		NULL-ant
africa	38		NULL-in
aggloméer	5		NULL-ation
amélior	4		NULL-ation
améri	8		NULL-que
américa	45		NULL-in



Words	Stem	Mini-Lexicon 3	Mini-Lexicon 2	Mini-Lexicon 1
decline	declin		e	
declined	declin			ed
declines	declin			es
decolletage	decolletage			
decor	decor			
decorate	decor		at	e
decorating	decor		at	ing
decoration	decor		at	ion
decorations	decor	at	ion	s
decorative	decor		at	ive
decorator	decor		at	or
decorators	decor	at	or	s
decrease	decrease			
decree	decree			
decreeing	decree			ing
decried	decri			ed
decries	decri			es
dedicated	dedicat			ed

4. Detect allomorphy

Signature: <e>ion . NULL

composite concentrate corporate détente
discriminate evacuate inflate opposite
participate probate prosecute tense

What is this?

composite and **composition**

composite → **composit** → **composit** + **ion**

It infers that **ion** deletes a stem-final ‘e’ before attaching.

3. Summary

Works very well on European languages.

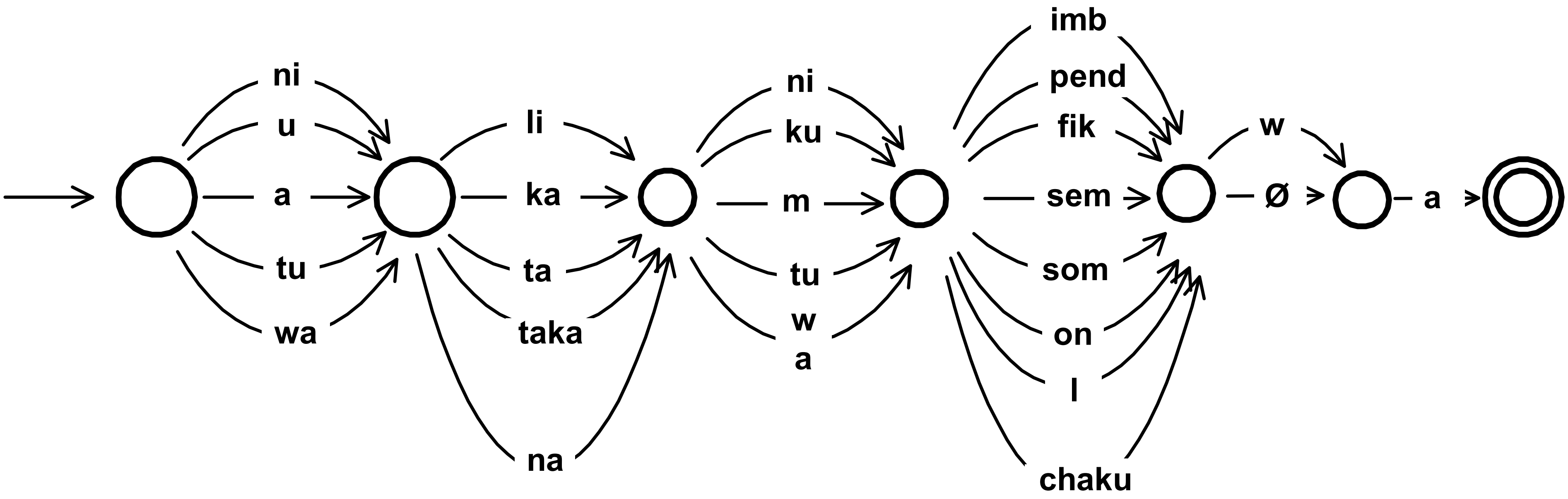
Challenges:

1. Works very poorly on languages with *richer morphologies* (average # morphemes/word >> 2). (Most languages have rich morphologies.)
2. Various other deficiencies.

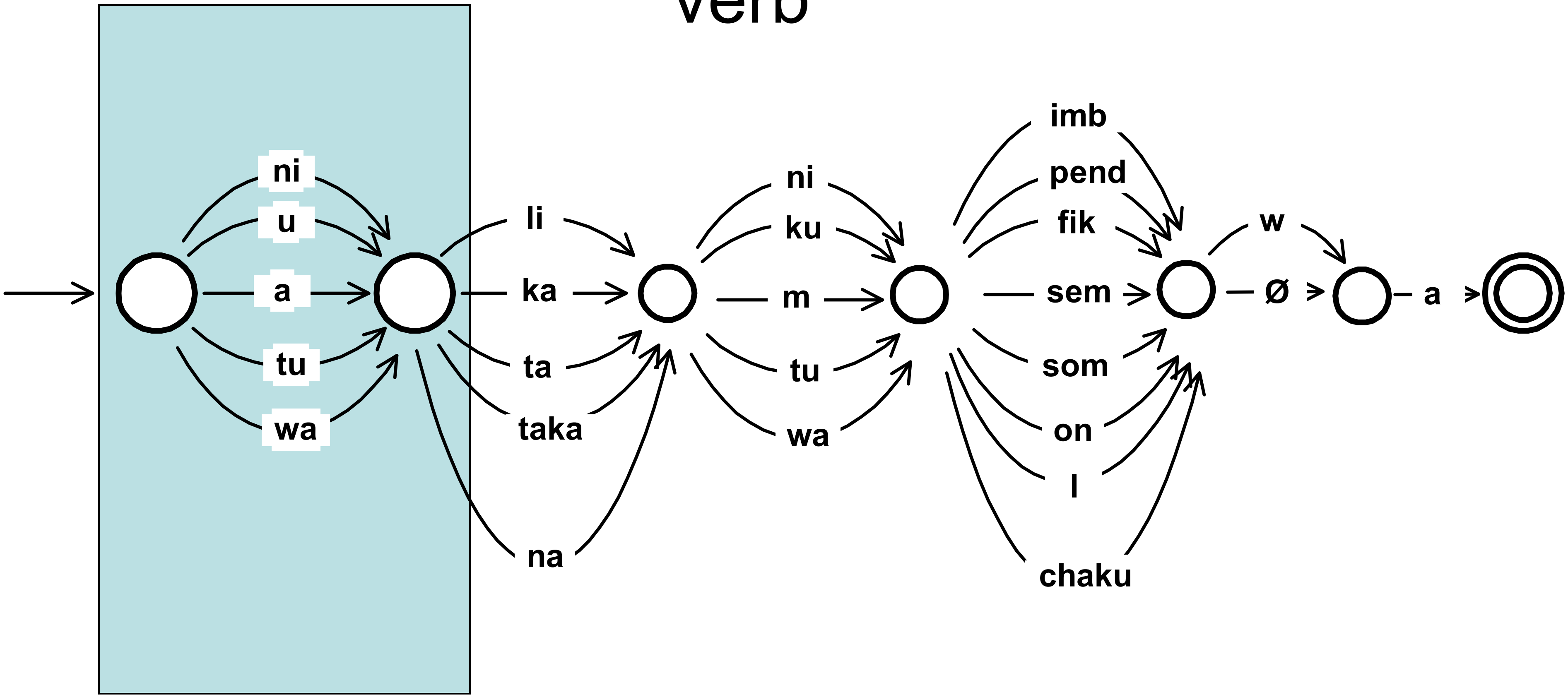
4. Morphology (*secundo*)

The initial bootstrap in the previous version does not even work on most languages, where the expected morphology contains sequences of 5 or more morphemes.

Swahili verb

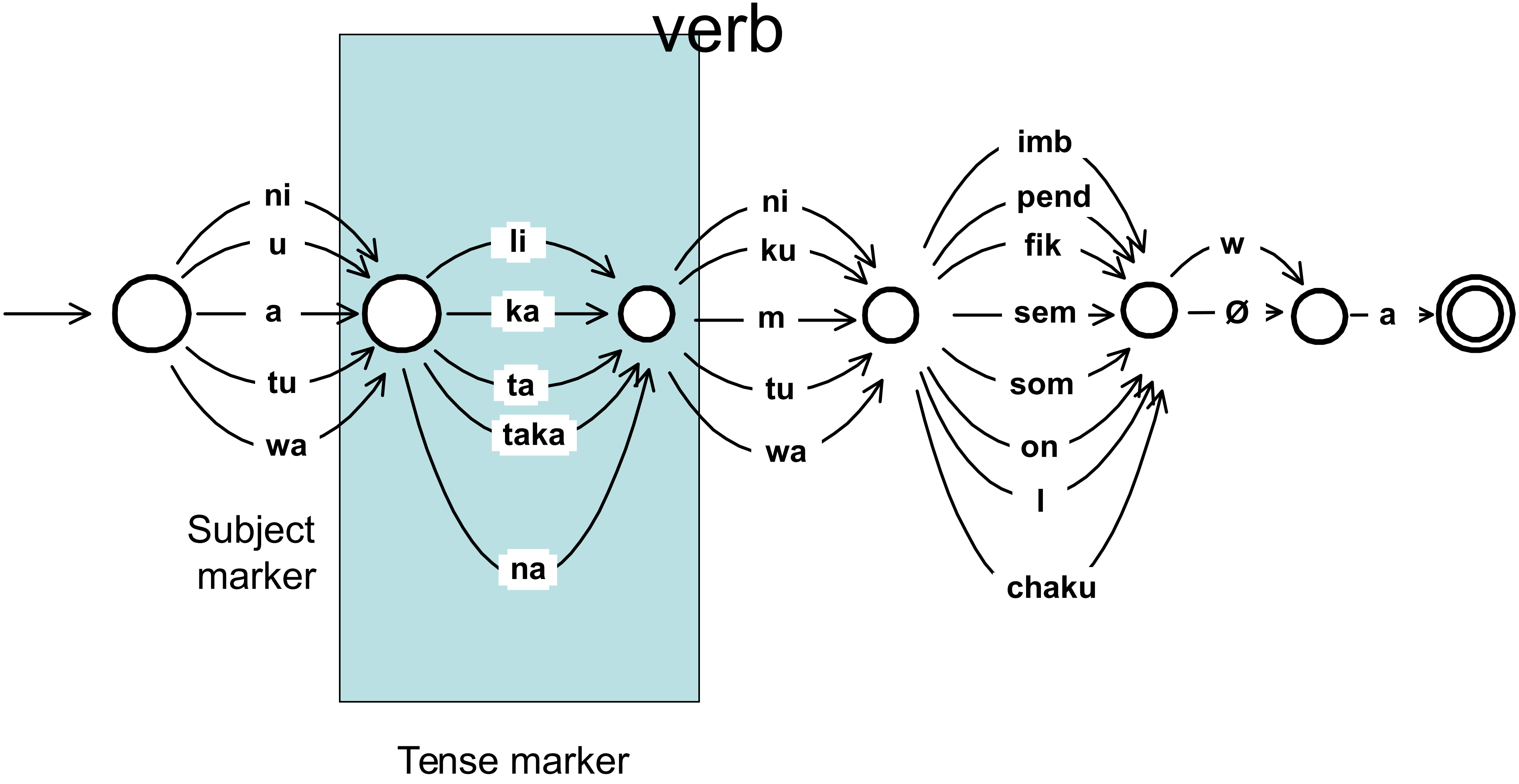


Swahili verb



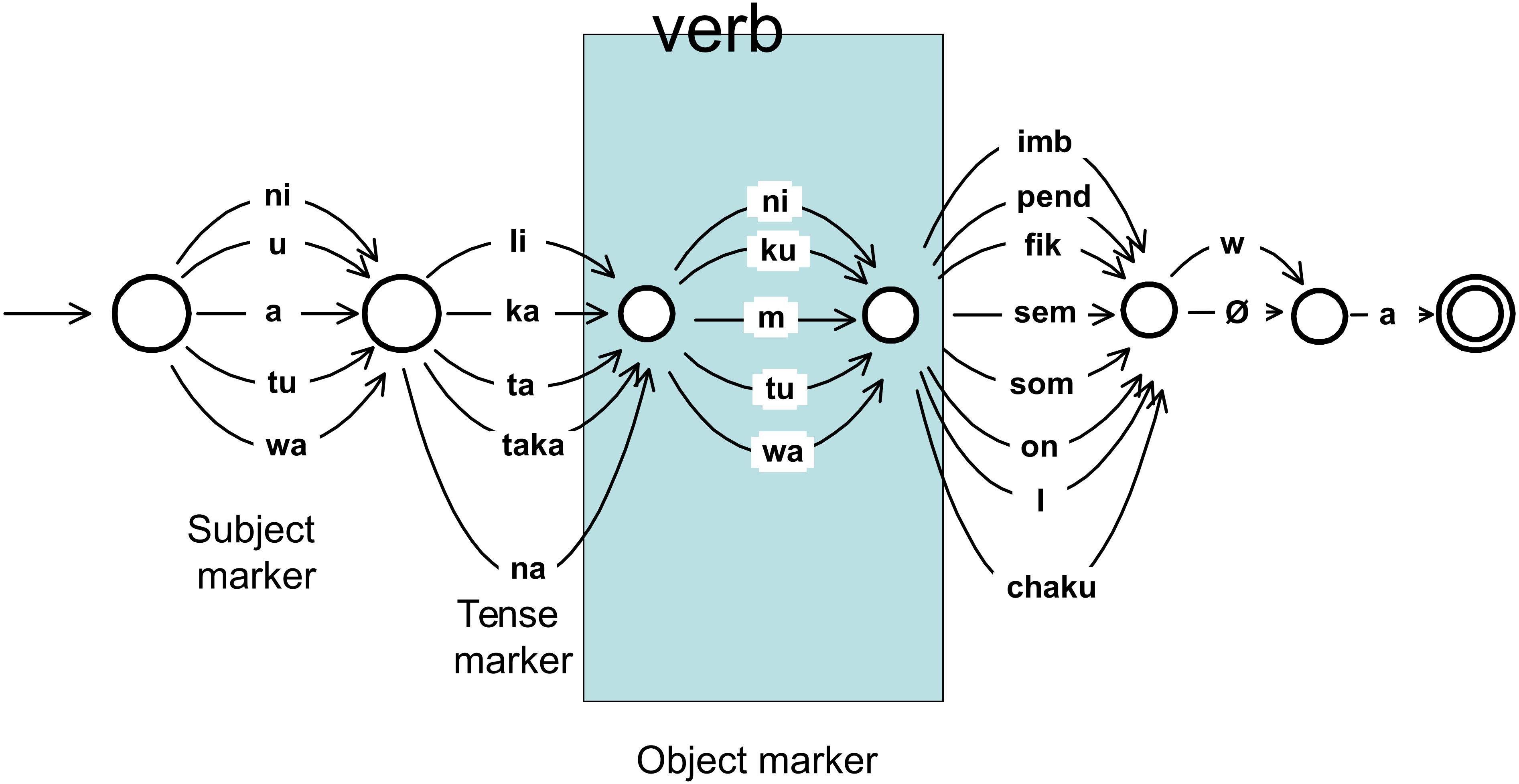
Subject marker

Swahili verb



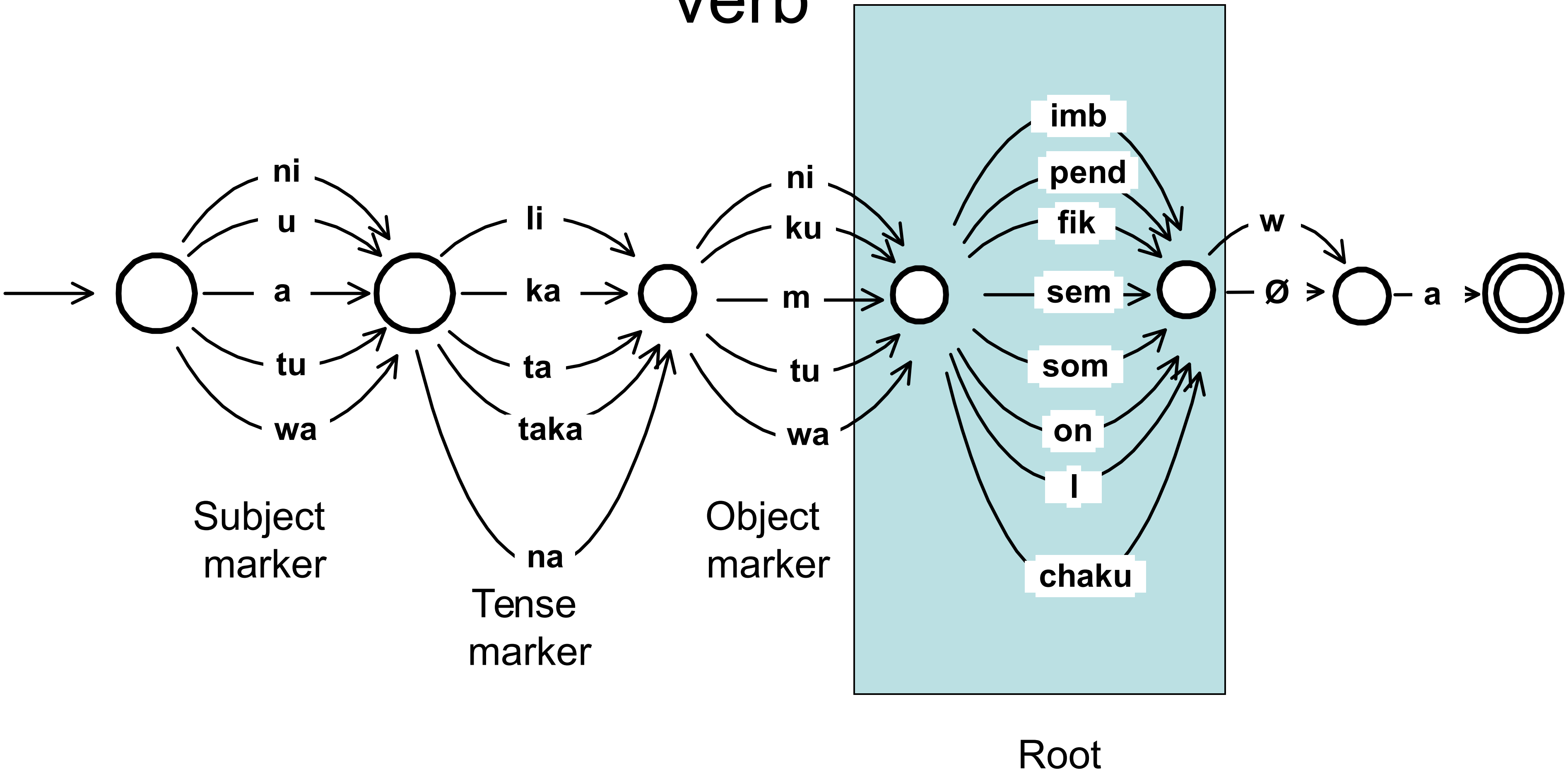
4. Morphology

Swahili verb

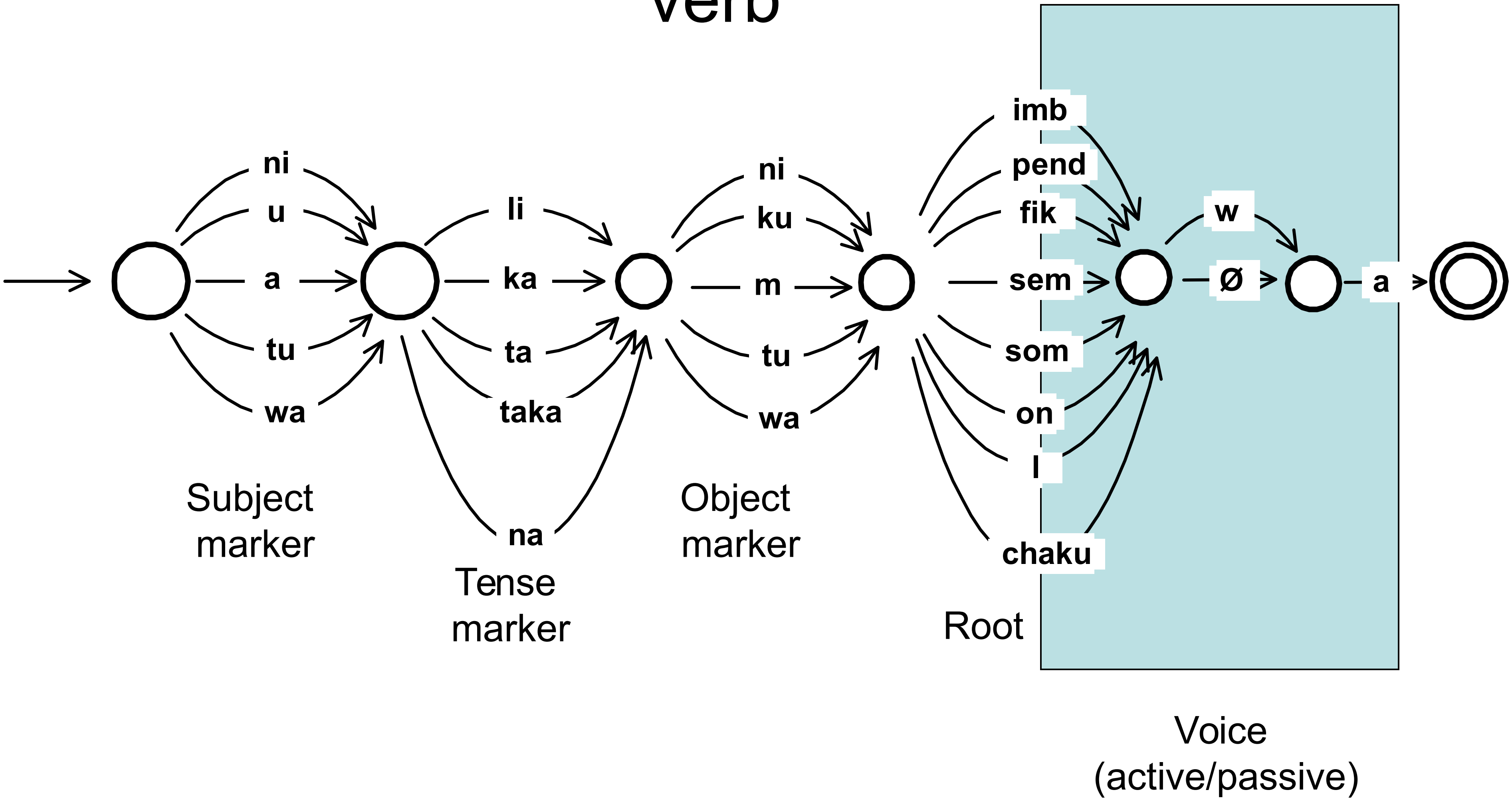


4. Morphology

Swahili verb



Swahili verb



ni
u
a
tu
wa

li
ka
ta
taka
na

ni
ku
m
tu
wa

imb
pend
fik
sem
som
on
l
chaku

w
Ø
a

Subject marker

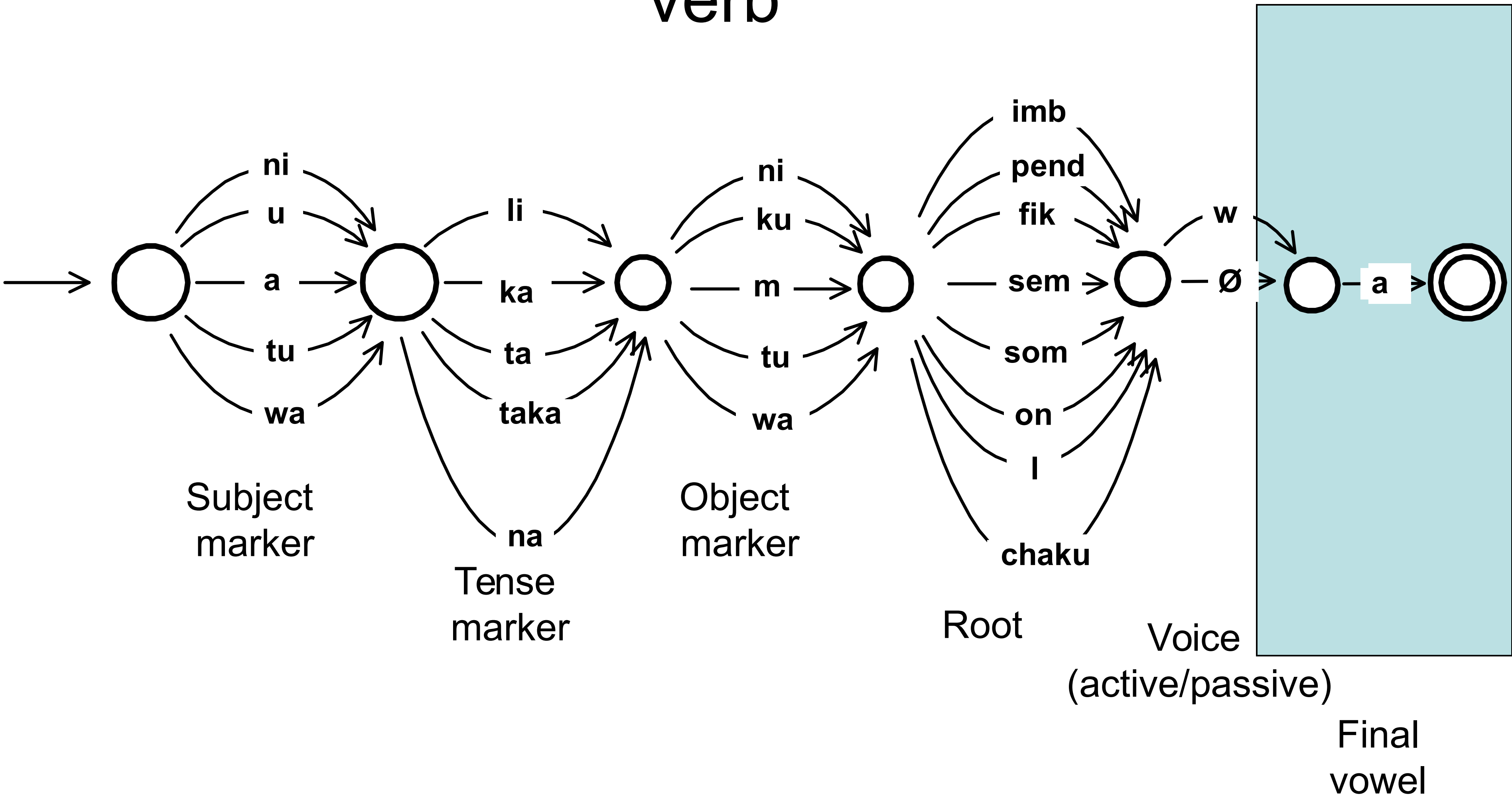
Tense marker

Object marker

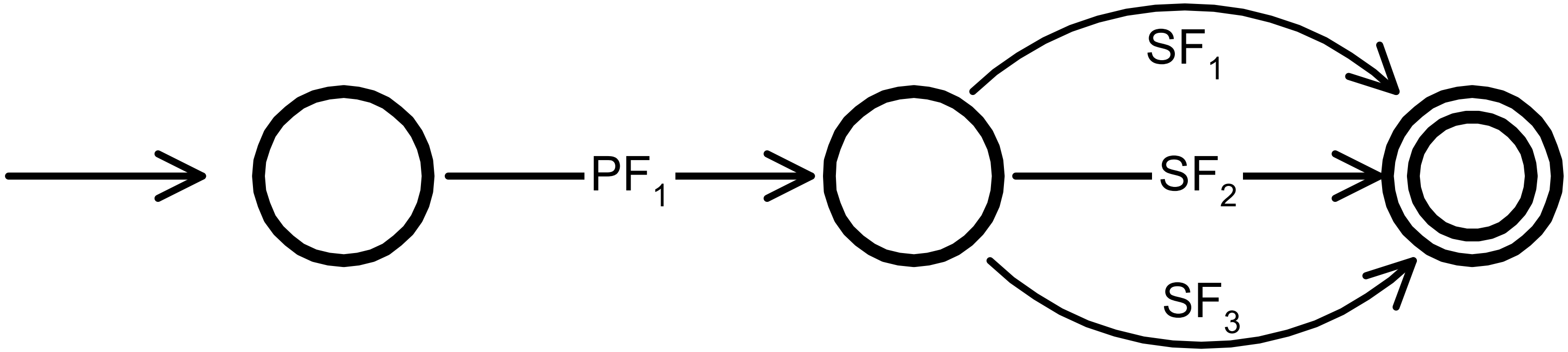
Root

Voice (active/passive)

Swahili verb



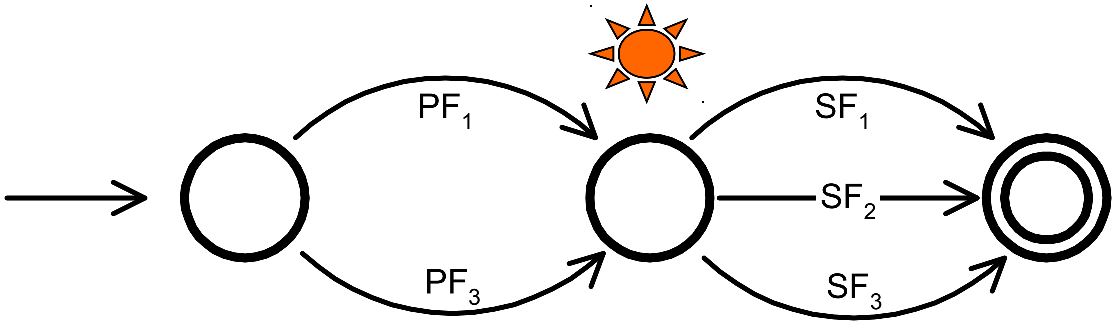
Finite state automaton (FSA)



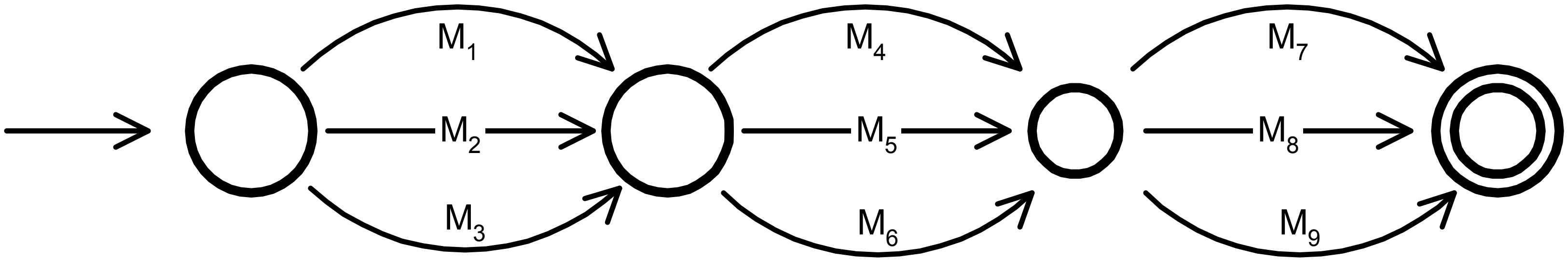
Signature:

reduces false positives

$$\left\{ \begin{matrix} \mathit{jump} \\ \mathit{walk} \end{matrix} \right\} \left\{ \begin{matrix} \mathit{NULL} \\ \mathit{ed} \\ \mathit{ing} \end{matrix} \right\}$$

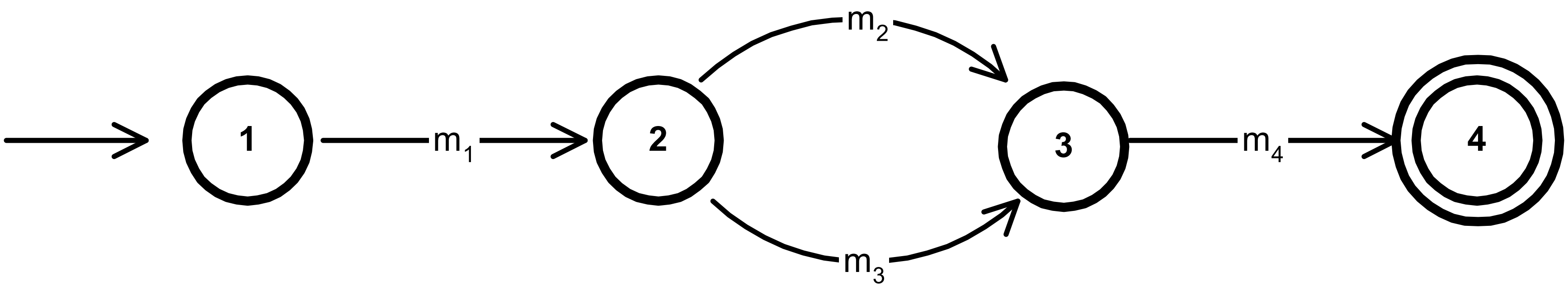


Generalize the signature…

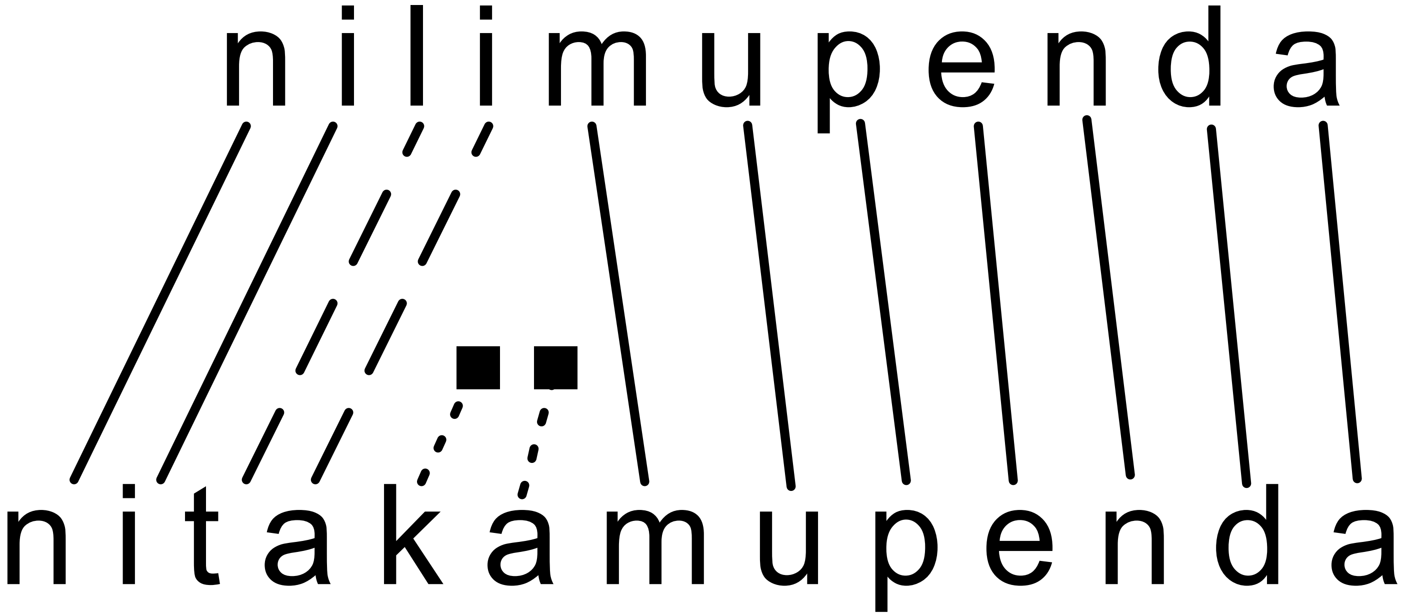


Sequential FSA: each state has a unique successor.

Alignments



Alignments: String edit distance algorithm

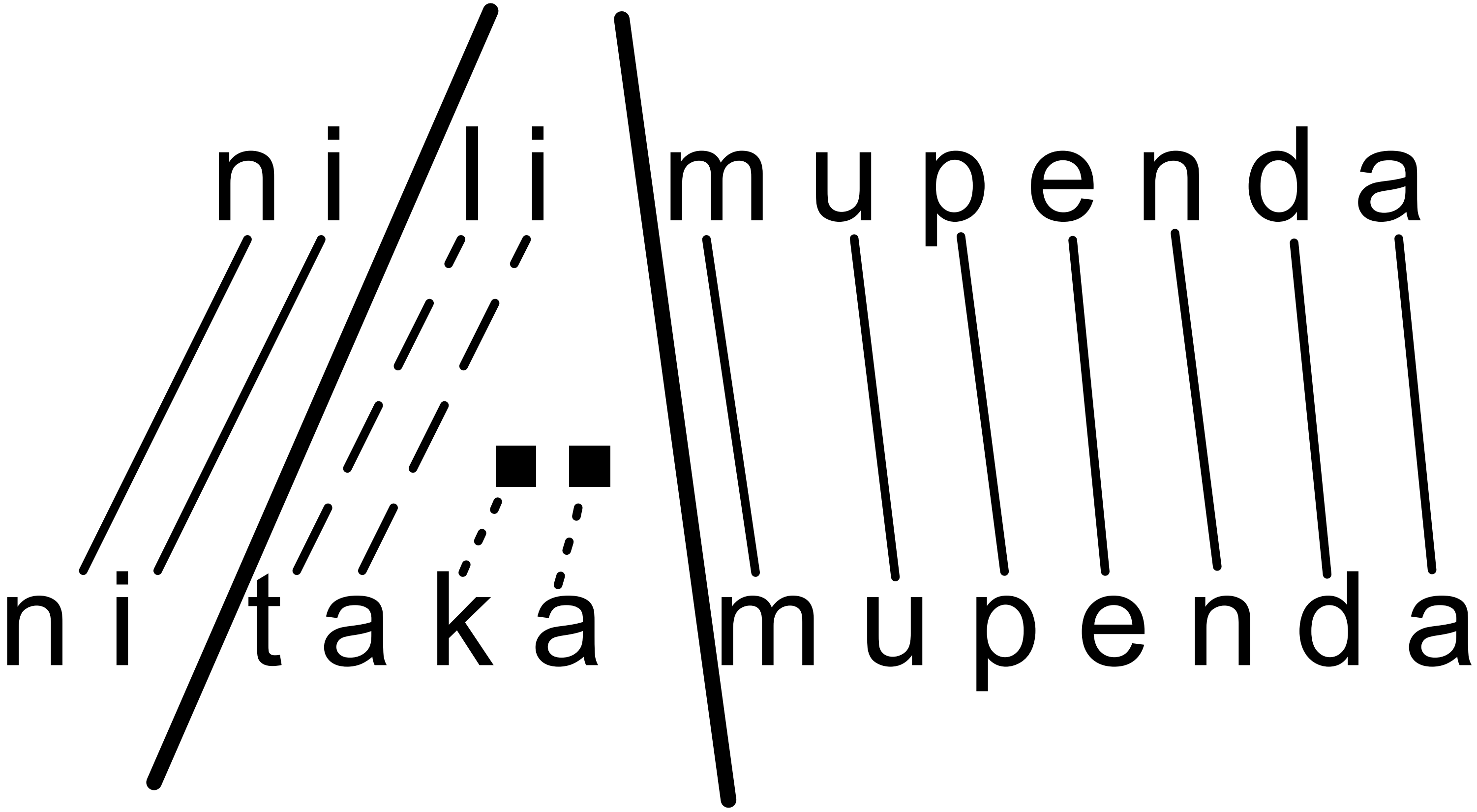


n i l i m u p e n d a

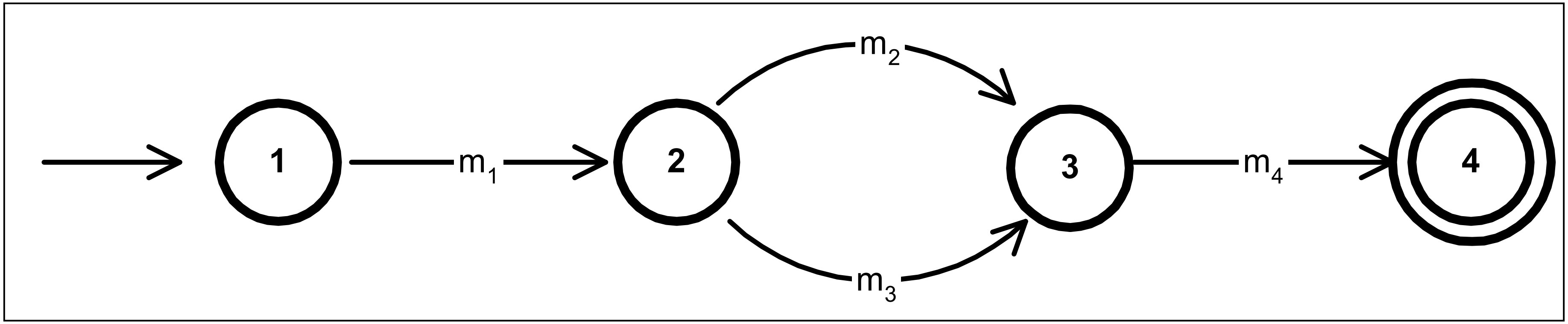
n i t a k a m u p e n d a

4. Morphology

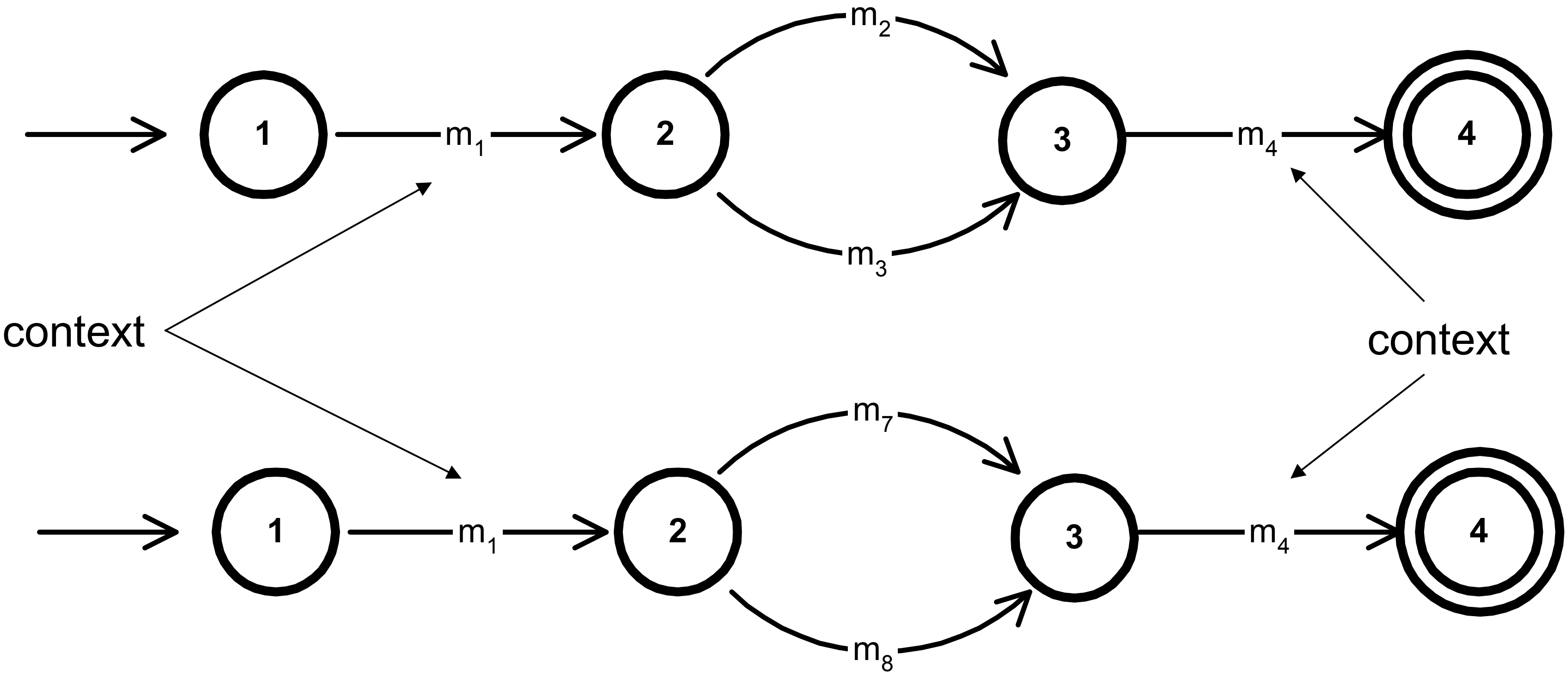
Alignments: make cuts



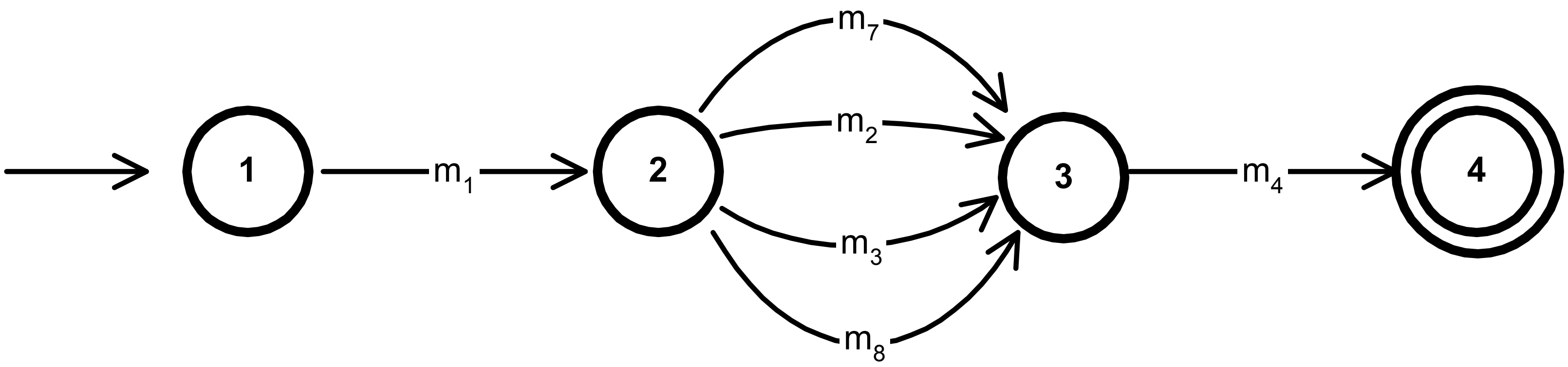
Elementary alignment



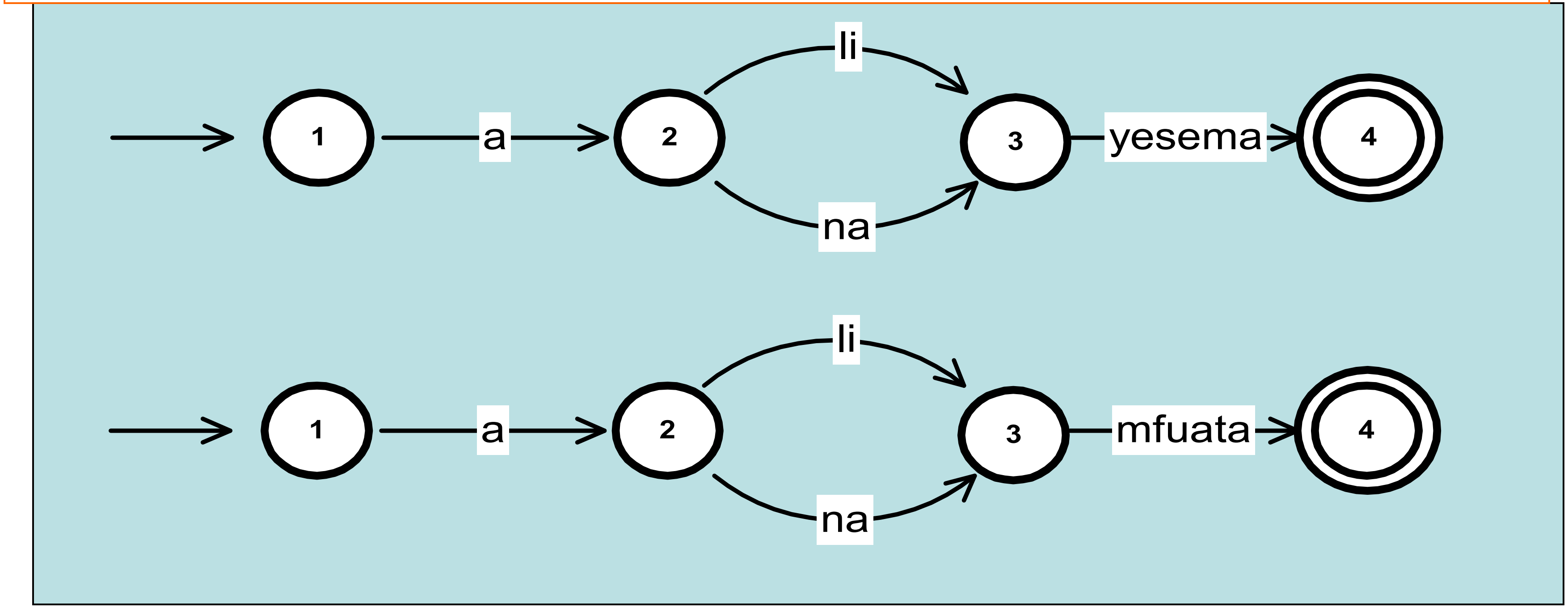
Collapsing elementary alignments

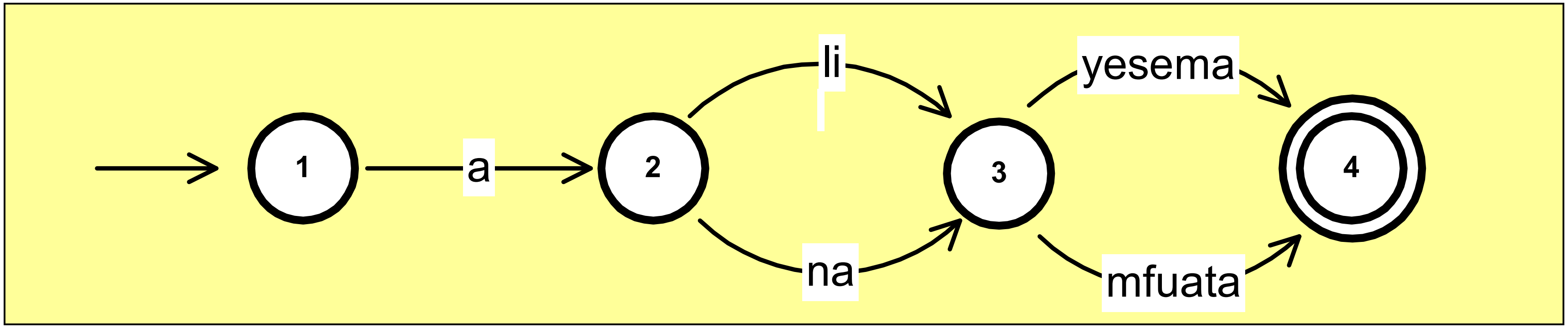


Two or more sequential FSAs with identical contexts are collapsed:



3. Further collapsing FSAs





4.3 Top templates: 8,200 Swahili words

State 1	State 2	State 3
a, wa (sg., pl. human subject markers)	246 stems	
ku, hu (infinitive, habitual markers)	51 stems	
wa (pl. subject marker)	*ka, li* (tense markers)	25 stems
a (sg. subject marker)	*ka, li* (tense markers)	29 stems
a (sg. subject marker)	*ka, na* (tense markers)	28 stems
37 strings	*w* (passive marker) / Ø	*a*

Precision and recall

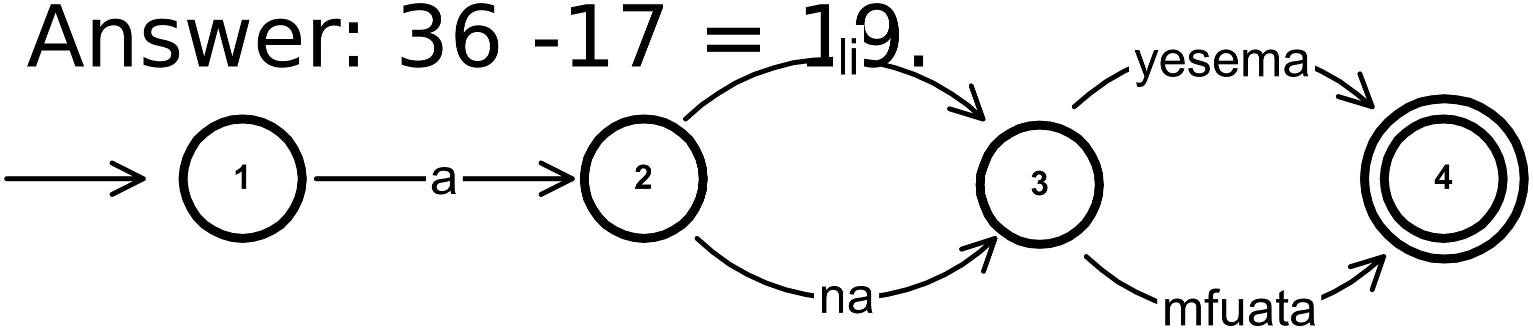
	Precision	Recall	F-score
String edit distance	0.77	0.57	0.65
Stem-affix	0.54	0.14	0.22
Affix-stem	0.68	0.20	0.31

Collapsed templates

		One Template	The other template	Collapsed Template	% found on Yahoo search
	1	{a}-{ka,na}-{stems}	{a}-{ka,ki}-{stems}	{a}-{ka,ki,na}-{stems}	86 (37/43)
	2	{wa}-{ka,na}-{stems}	{wa}-{ka,ki}-{stems}	{wa}-{ka,ki,na}-{stems}	95 (21/22)
	3	{a}-{ka,ki,na}-{stems}	{wa}-{ka,ki,na}-{stems}	{a,wa}-{ka,ki,na}-{stems}	84 (154/183)
	4	{a}-{liye,me}-{stems}	{a}-{liye,li}-{stems}	{a}-{liye,li,me}-{stems}	100 (21/21)
	5	{a}-{ki,li}-{stems}	{wa}-{ki,li}-{stems}	{a,wa}-{ki,li}-{stems}	90 (36/40)
	6	{a}-{lipo,li}-{stems}	{wa}-{lipo,li}-{stems}	{a,wa}-{lipo,li}-{stems}	90 (27/30)
7		{a,wa}-{ki,li}-{stems}	{a,wa}-{lipo,li}-{stems}	{a,wa}-{ki,lipo,li}-{stems}	74 (52/70)
8		{a}-{na,naye}-{stems}	{a}-{na,ta}-{stems}	{a}-{na,ta,naye}-{stems}	80 (12/15)

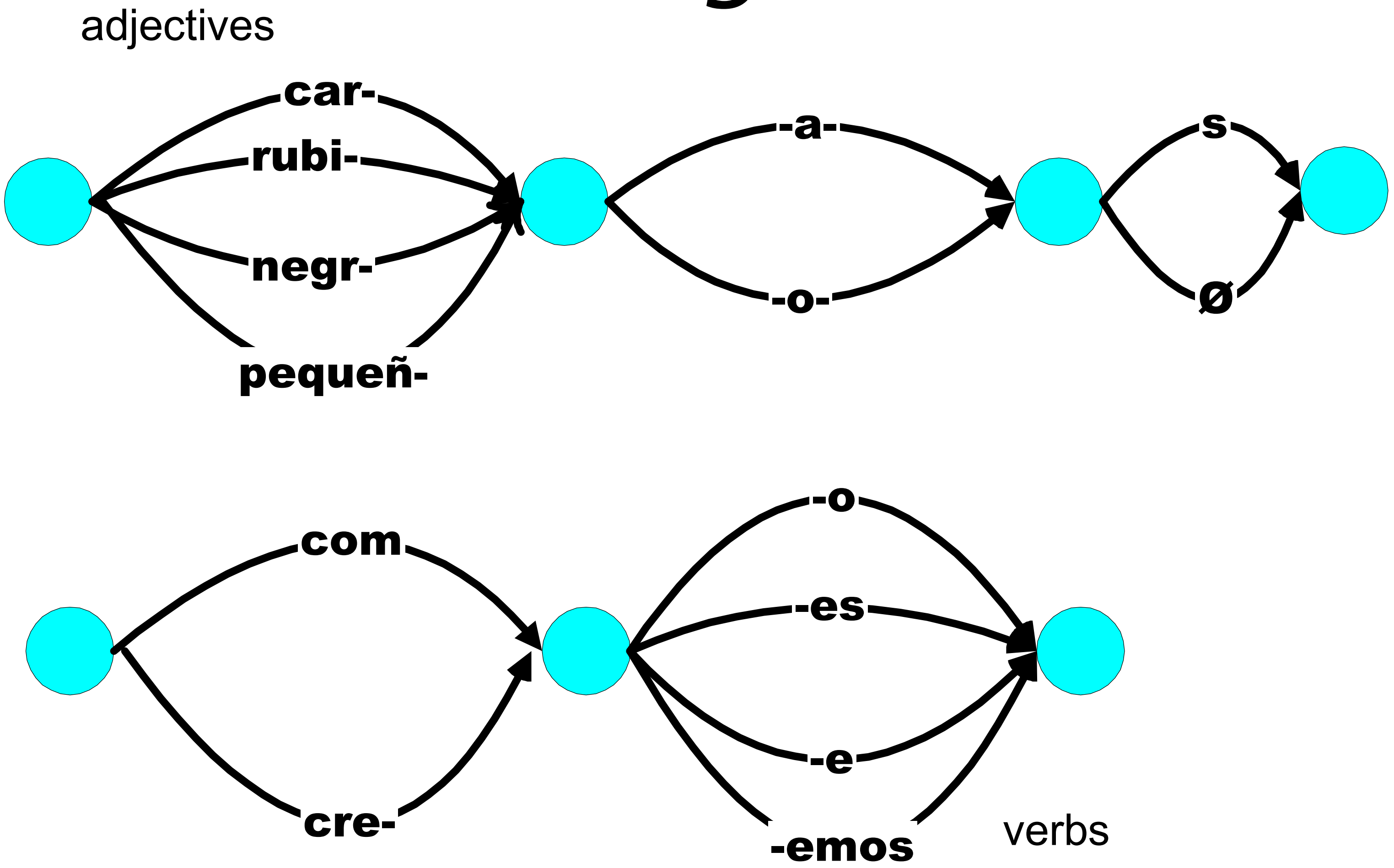
4. 1 Evaluating the robustness of these templates (sequential FSAs)

- Measure: How many letters do we *save* by expressing words in a template rather than by writing each one out individually?
  Answer: 36 -17 = 19.

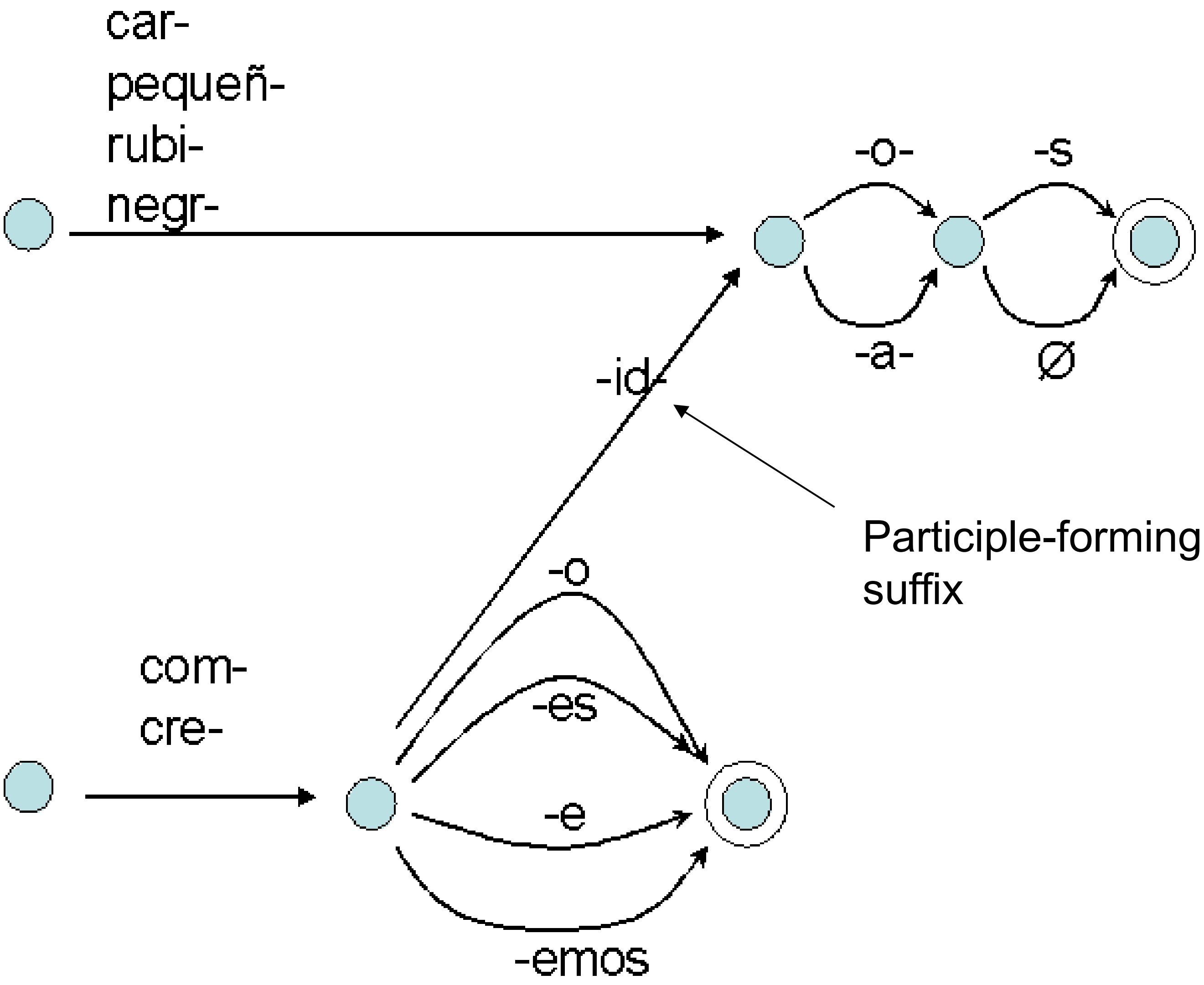


4. Morphology

Most edges are *convergent…*



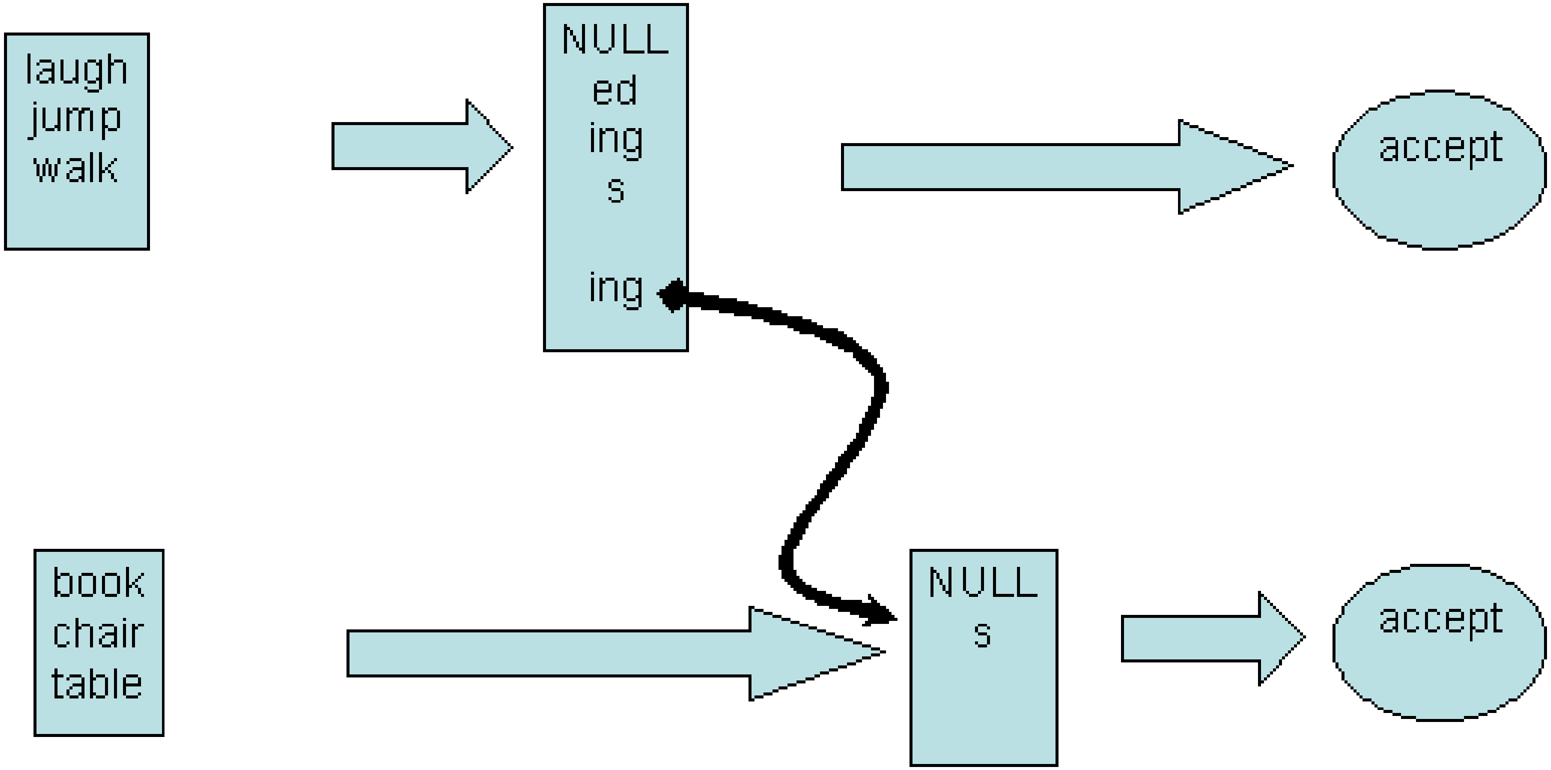
But some diverge (Spanish):



Participle-forming suffix

4. Morphology

English has much the same:



laugh
jump
walk

NULL
ed
ing
s

ing

accept

book
chair
table

NULL
s

accept

4. Summary

We need to enrich the heuristics and consider a broader set of possible grammars.

With that, improvements seem to be unlimited at this point in time.

Focus: Decrease the length of the analysis, especially in the length of the *substance* (morphemes) described.

5. Phonology

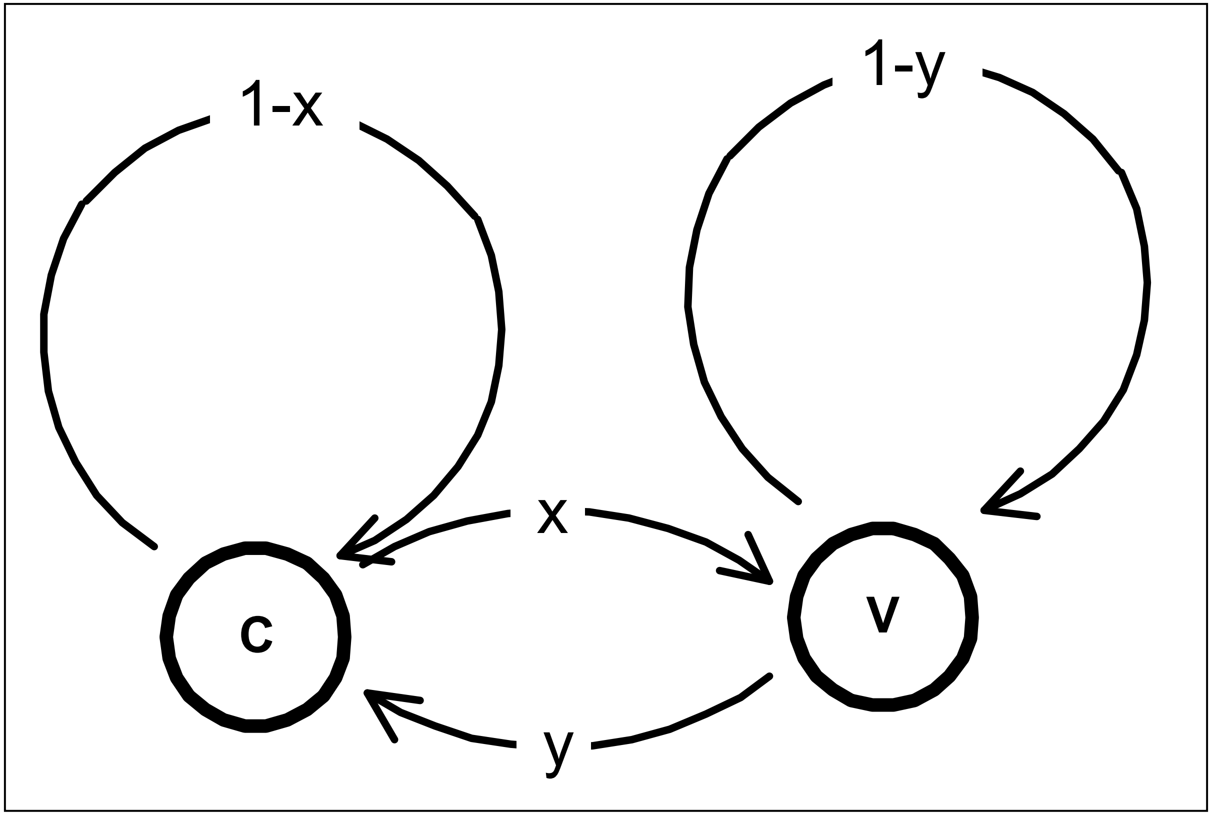
So far we have said little about phonology.

We have assumed no interesting probabilistic model of segment (=phoneme) placement. (0th or 1st order Markov model).

But we can shorten the length of the grammar by taking this into consideration.

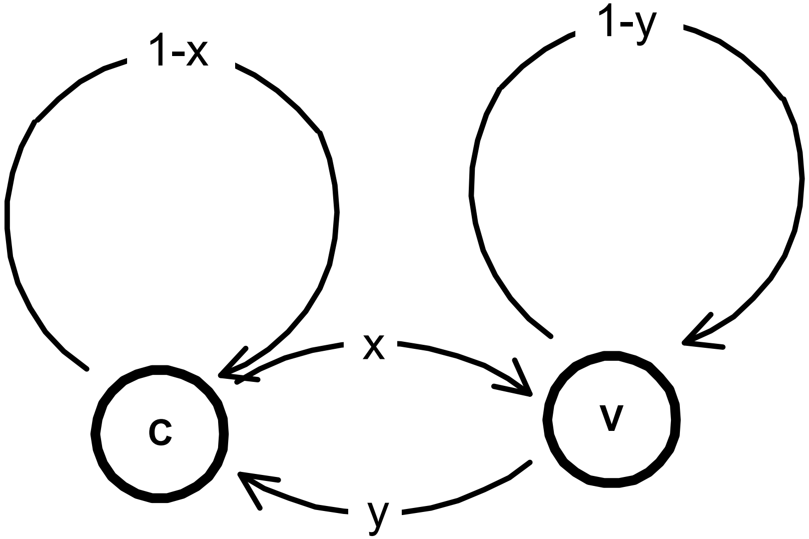
These slides present material done jointly with Aris Xanthos and with Jason Riggle.

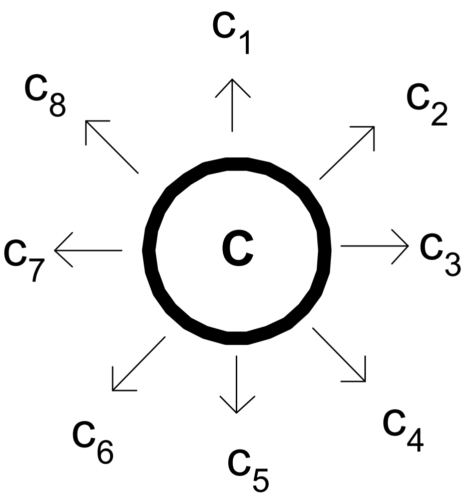
Much more interesting model:



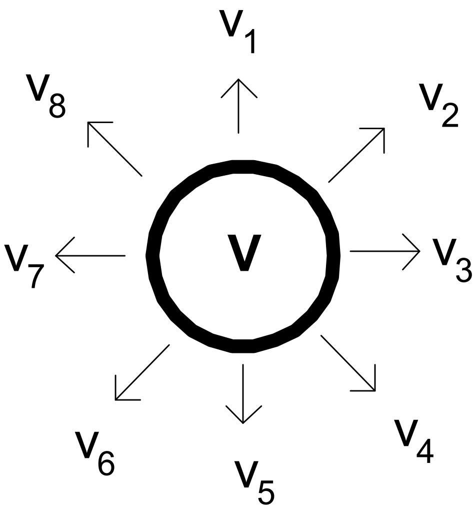
For state transitions; and the same model for emissions: both states emit all of the symbols, but with different probabilities....

5. Phonology





$$\sum_i c_i = 1$$



$$\sum_i v_i = 1$$

The question is…

- How could we obtain the *best* probabilities for *x* and *y* (transition probabilities), and all of the emission probabilities for the two states?
- Bear in mind: each state generates *all* of the symbols. The only way to ensure that a state does *not* generate a symbol *s* is to assign a zero probability for the emission of the symbol *s* in that state.

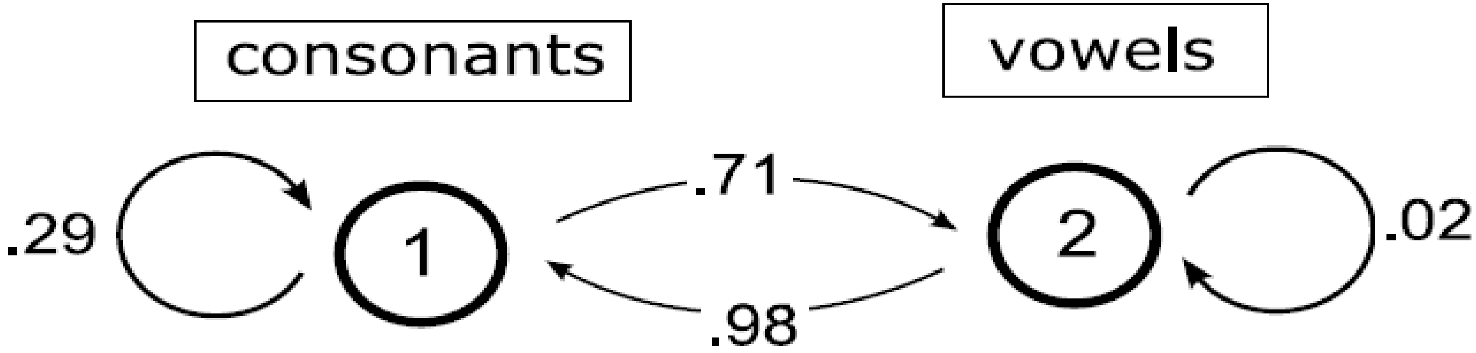
Hidden Markov model

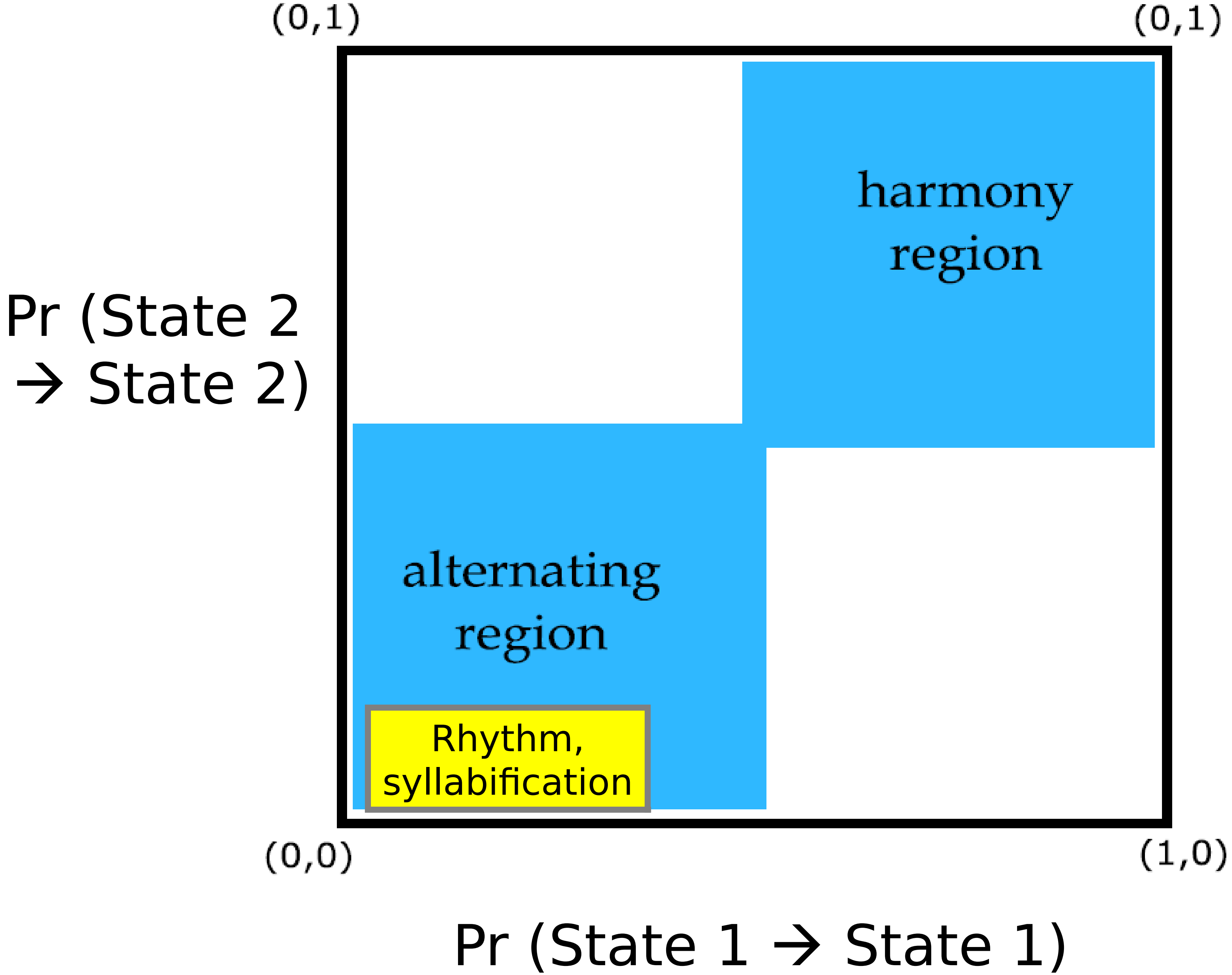
With a well-understood training algorithm, an HMM will find the optimal parameters to generate the data so as to assign it the highest probability.

How does it organize the phonological data?

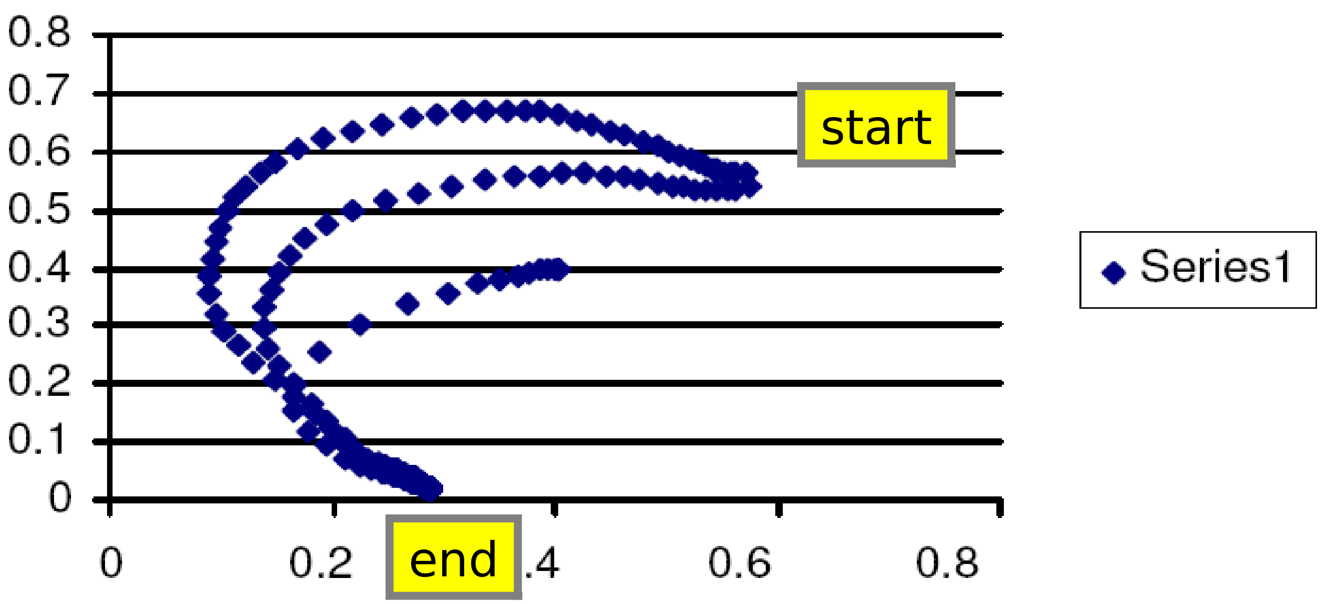
5. Phonology

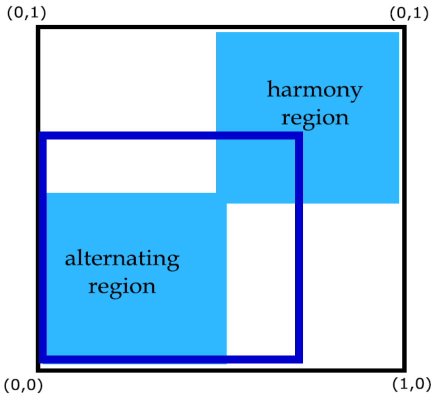
English FSA




(0,1)
(0,1)
harmony
region
Pr (State 2
→ State 2)
alternating
region
Rhythm,
syllabification
(0,0)
(1,0)
Pr (State 1 → State 1)

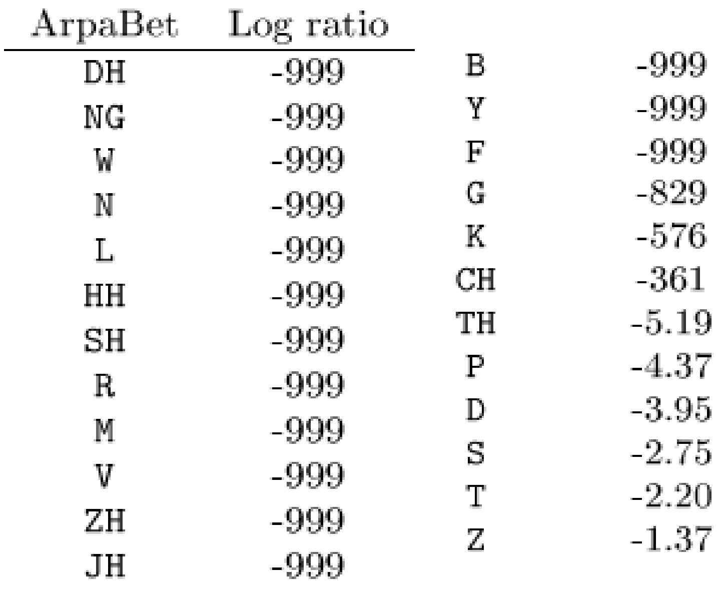
Dynamics of English 2 state transitions (learning)





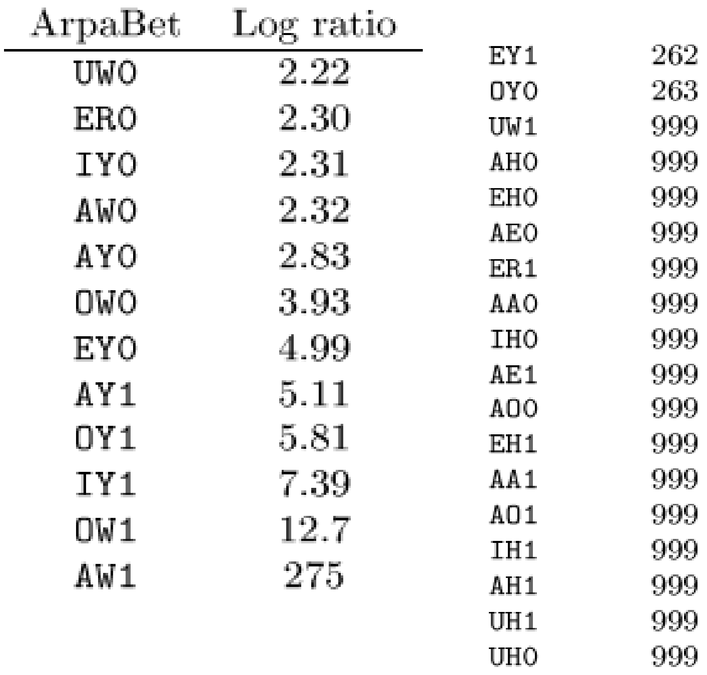
English: Log ratios of the emission probabilities of the 2 states:

$$\log \frac{p_1(\phi)}{p_2(\phi)}$$



ArpaBet	Log ratio		
DH	-999	B	-999
NG	-999	Y	-999
W	-999	F	-999
N	-999	G	-829
L	-999	K	-576
HH	-999	CH	-361
SH	-999	TH	-5.19
R	-999	P	-4.37
M	-999	D	-3.95
V	-999	S	-2.75
ZH	-999	T	-2.20
JH	-999	Z	-1.37

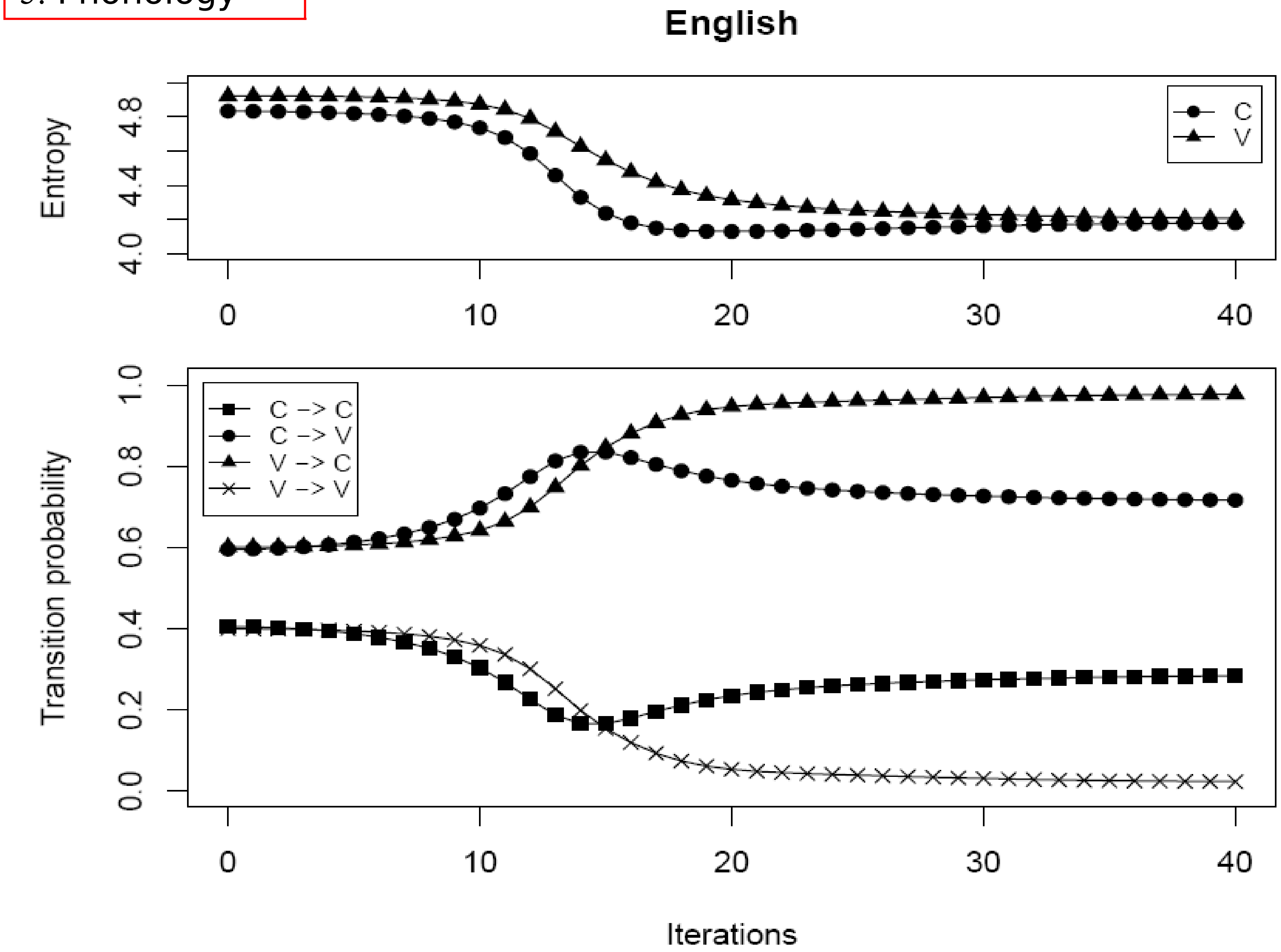
negative



ArpaBet	Log ratio		
UW0	2.22	EY1	262
ER0	2.30	OY0	263
IY0	2.31	UW1	999
AW0	2.32	AH0	999
AY0	2.83	EH0	999
OW0	3.93	AE0	999
EY0	4.99	ER1	999
AY1	5.11	AA0	999
OY1	5.81	IH0	999
IY1	7.39	AE1	999
OW1	12.7	AO0	999
AW1	275	EH1	999
		AA1	999
		AO1	999
		IH1	999
		AH1	999
		UH1	999
		UH0	999

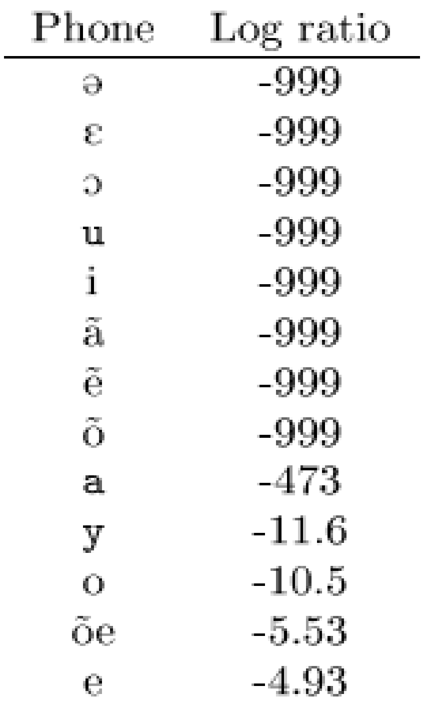
positive

5. Phonology


English
Entropy
4.0
4.4
4.8
● C
▲ V
0
10
20
30
40
Transition probability
0.0
0.2
0.4
0.6
0.8
1.0
■ C -> C
● C -> V
▲ V -> C
× V -> V
0
10
20
30
40
Iterations

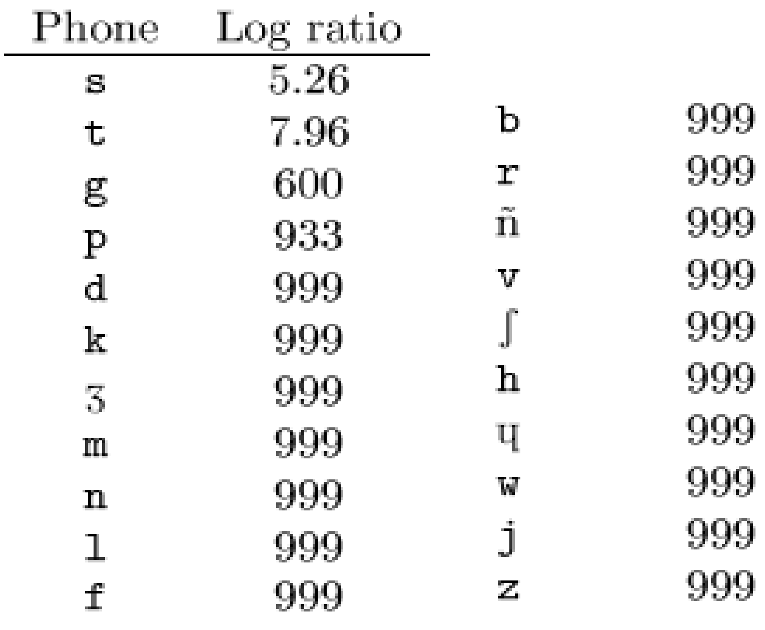
French: Log ratios of the emission probabilities of the 2 states:

$$\log \frac{p_1(\phi)}{p_2(\phi)}$$



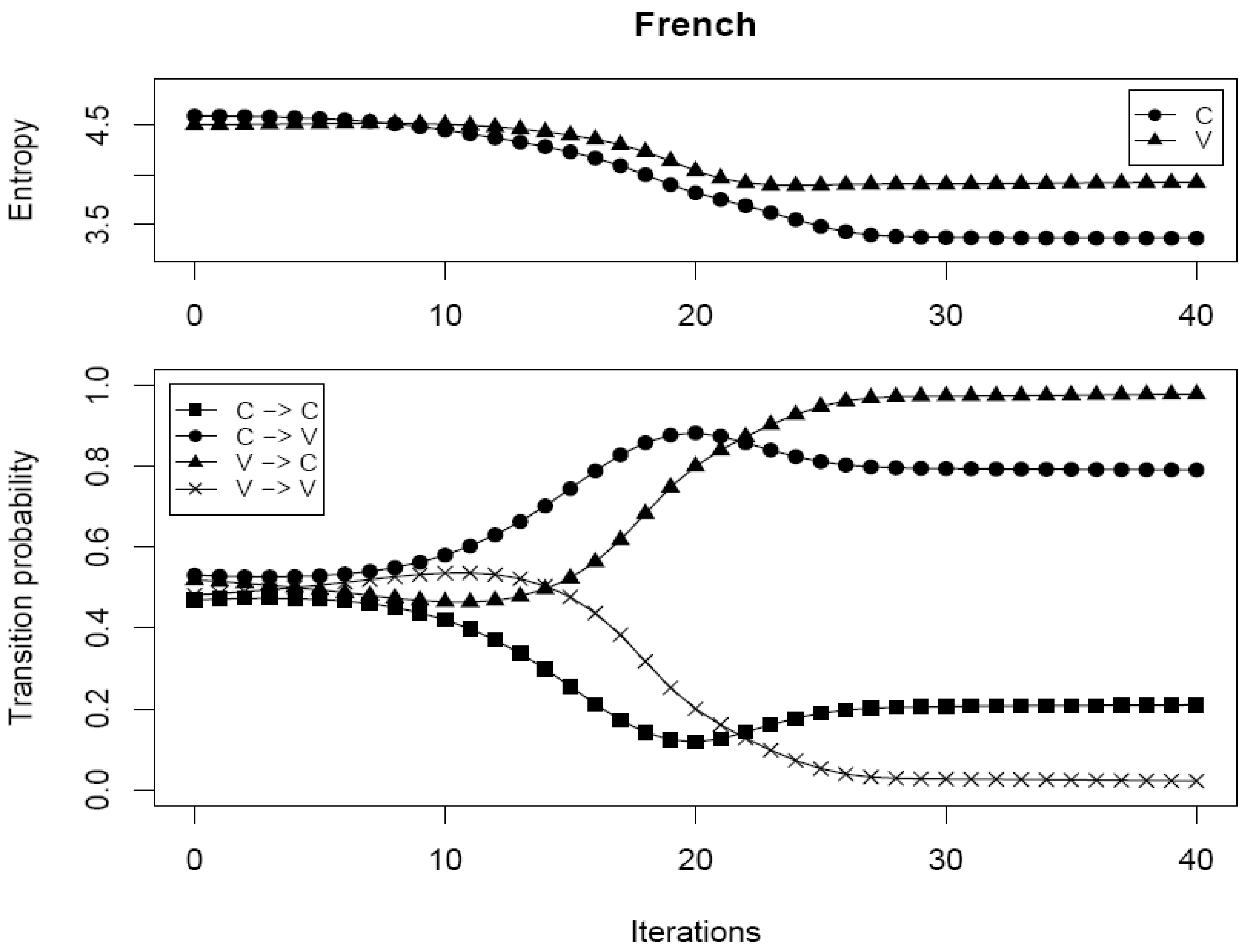
Phone	Log ratio
ə	-999
ɛ	-999
ɔ	-999
u	-999
i	-999
ã	-999
ẽ	-999
õ	-999
a	-473
y	-11.6
o	-10.5
õe	-5.53
e	-4.93

negative



Phone	Log ratio
s	5.26
t	7.96
g	600
p	933
d	999
k	999
ʒ	999
m	999
n	999
l	999
f	999
b	999
r	999
ñ	999
v	999
ʃ	999
h	999
ɥ	999
w	999
j	999
z	999

positive


French
Entropy
3.5
4.5
C
V
0
10
20
30
40
Transition probability
0.0
0.2
0.4
0.6
0.8
1.0
C -> C
C -> V
V -> C
V -> V
0
10
20
30
40
Iterations

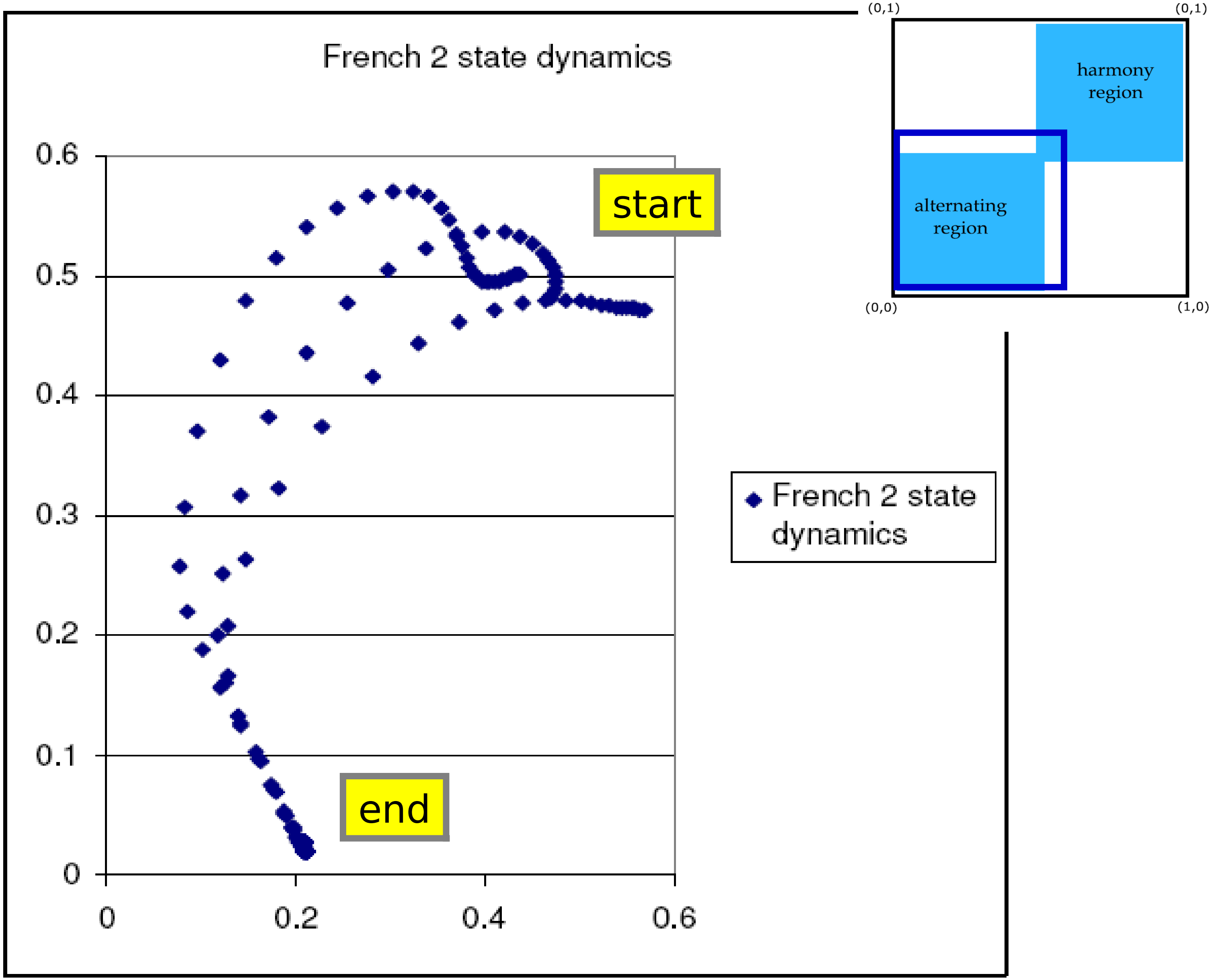
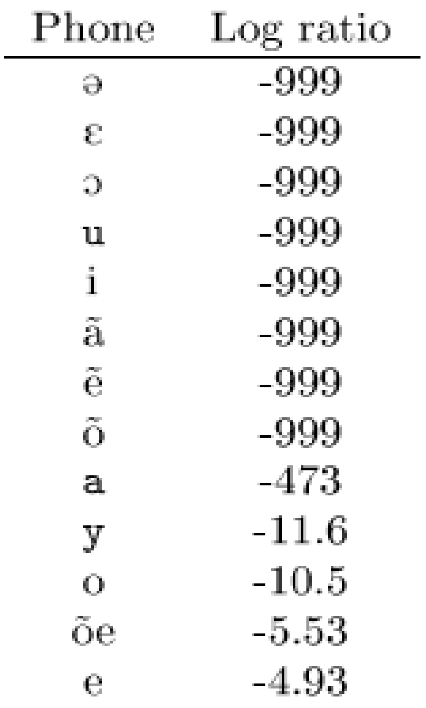


Figure 11: Dynamics of learning French c/v

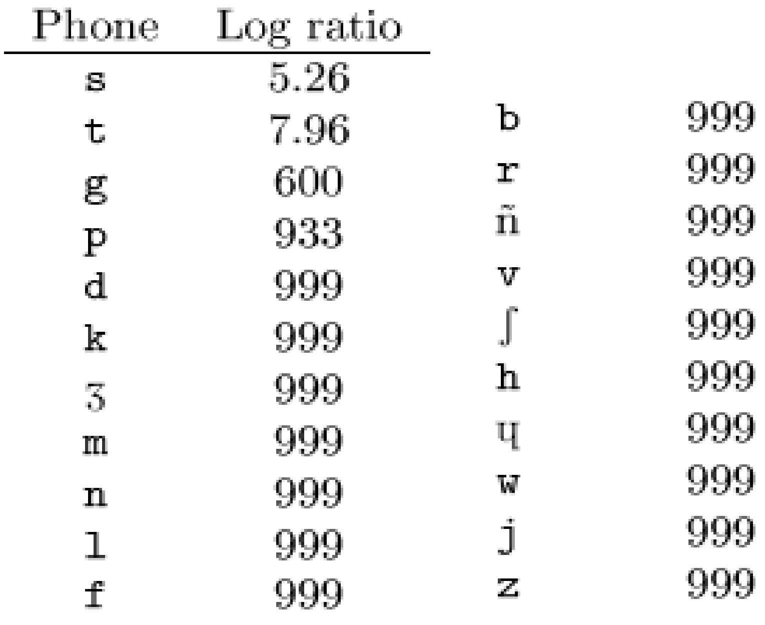
Finnish: Log ratios of the emission probabilities of the 2 states:

$$\log \frac{p_1(\phi)}{p_2(\phi)}$$



Phone	Log ratio
ə	-999
ɛ	-999
ɔ	-999
u	-999
i	-999
ã	-999
ẽ	-999
õ	-999
a	-473
y	-11.6
o	-10.5
õe	-5.53
e	-4.93

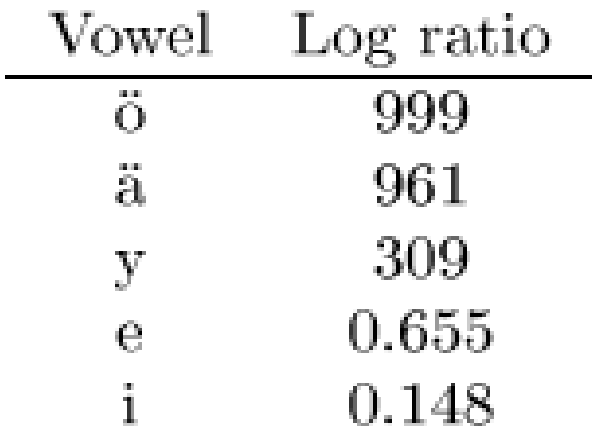
negative



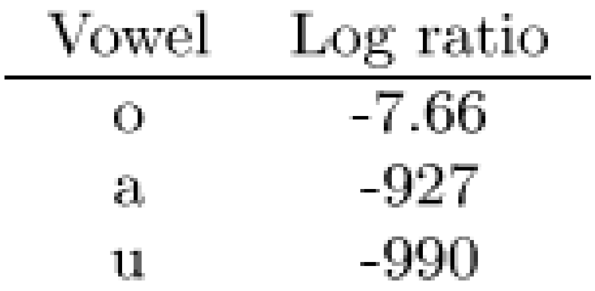
Phone	Log ratio		
s	5.26		
t	7.96	b	999
g	600	r	999
p	933	ñ	999
d	999	v	999
k	999	ʃ	999
ʒ	999	h	999
m	999	ɥ	999
n	999	w	999
l	999	j	999
f	999	z	999

positive

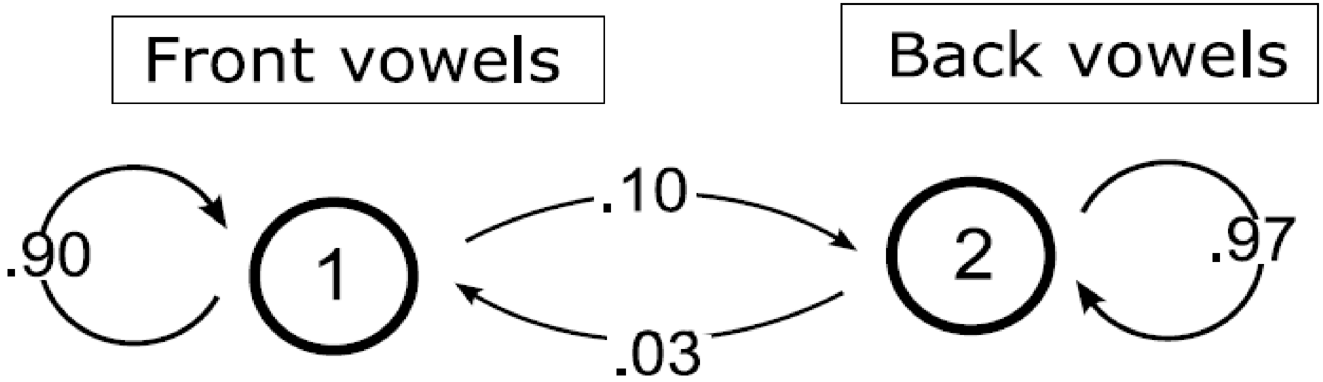
Finnish vowels and their harmony

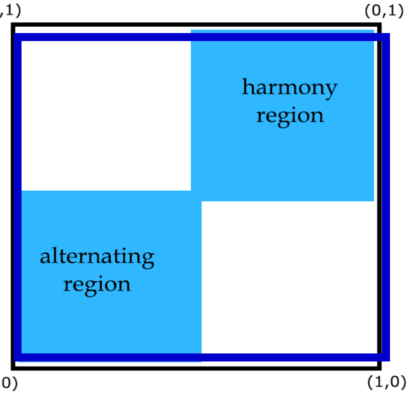


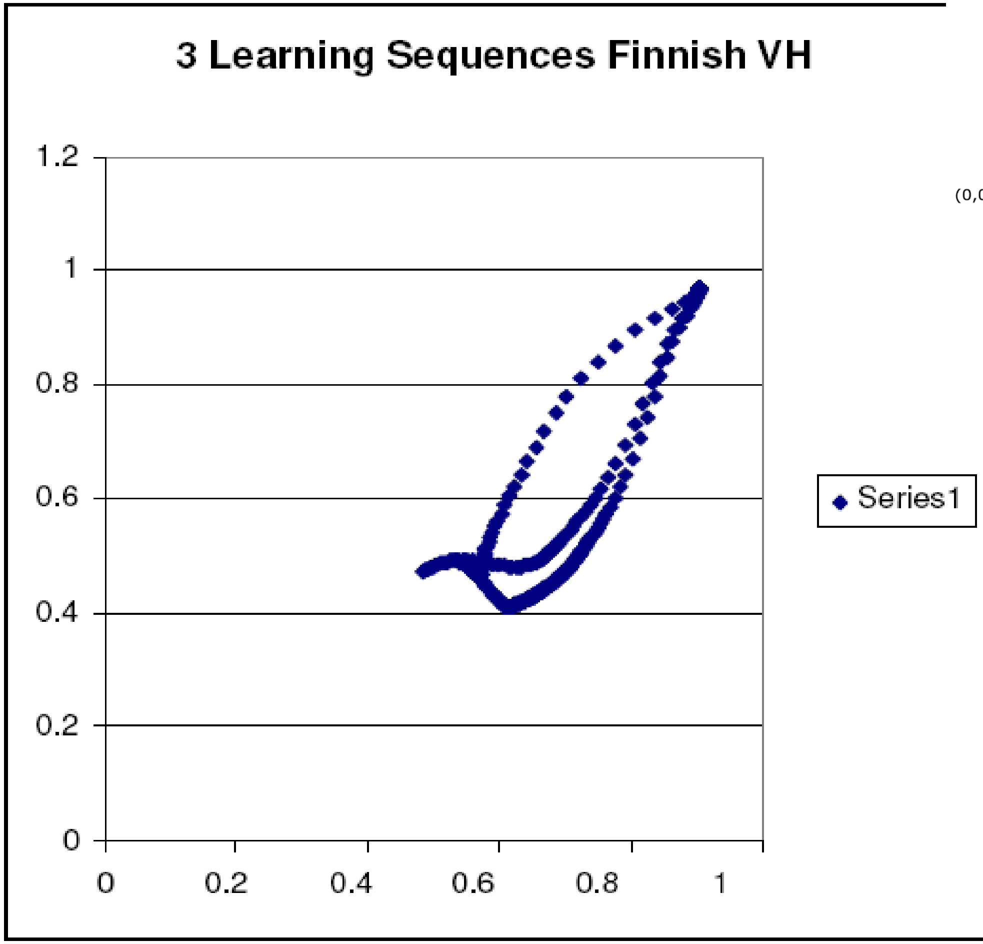
Vowel	Log ratio
ö	999
ä	961
y	309
e	0.655
i	0.148

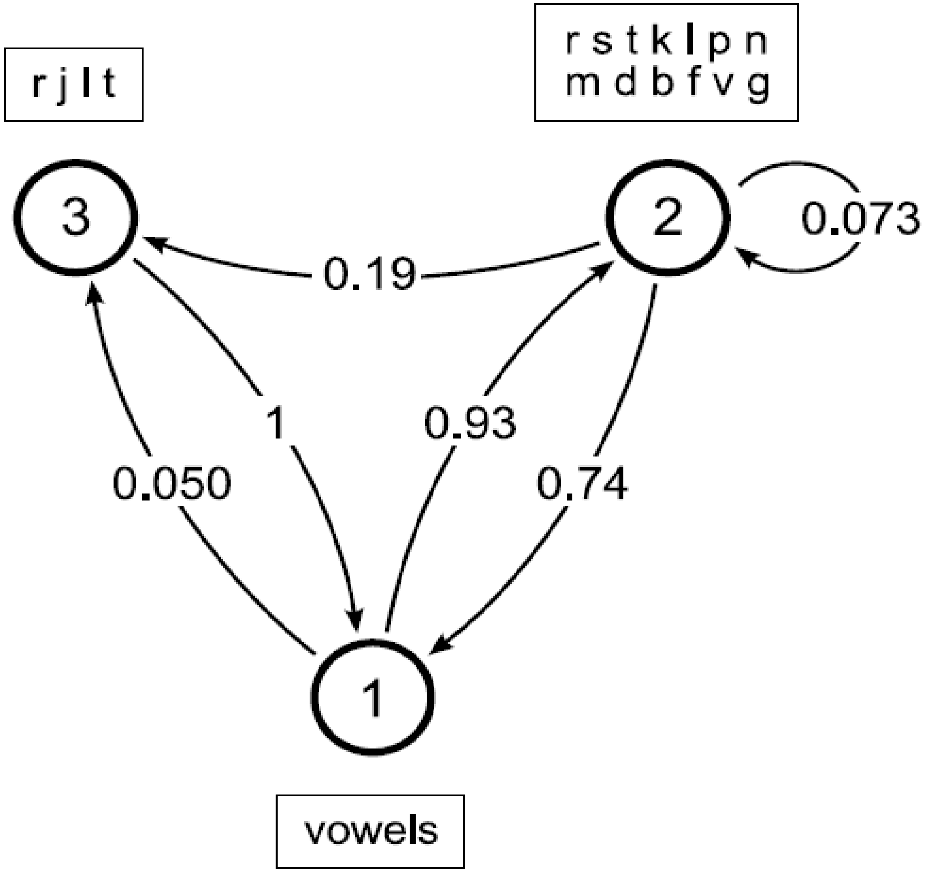


Vowel	Log ratio
o	-7.66
a	-927
u	-990


Front vowels
Back vowels
.90
1
.10
.03
2
.97


(0,1)
(0,1)
harmony region
alternating region
(0,0)
(1,0)


3 Learning Sequences Finnish VH
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
◆ Series1


r j l t
r s t k l p n
m d b f v g
3
2
0.073
0.19
1
0.93
0.050
0.74
1
vowels

From State 1	Prob
a	0.17
e	0.15
i	0.15
ə	0.15
o	0.087
ɛ	0.058
2	0.056
y	0.043
4	.036
I	.027
u	.026
ɔ	.026

From State 2	Prob
r	.14
s	.11
t	.10
k	.096
l	.078
p	.072
n	.062
m	.059
d	.059
b	.047
f	.037
v	.031
g	0.029
z	0.026
ʒ	0.021

From State 3	Prob
r	.28
j	.21
l	.13
t	.12
w	.059
e	.051
m	.033

Table 12: Emission probabilities, 3 state HMM for French

Emit:	while in state:	prob	transition	prob
a	3	0.6	$3 \rightarrow 2$	0.62
b	2	0.06	$2 \rightarrow 1$	0.24
r	1	0.34	$1 \rightarrow 3$	0.77
a	3	0.6		

probability: 0.0023

Emit:	while in state:	prob	transition	prob
a	3	0.6	$3 \rightarrow 1$	0.37
b	1	$3 \cdot 10^{-}35$	$1 \rightarrow 2$	0.22
r	2	0.06	$2 \rightarrow 3$	0.75
a	3	0.6		

probability: ≈ 0

Emit:	while in state:	prob	transition	prob
a	3	0.6	$3 \rightarrow 2$	0.62
r	2	0.06	$2 \rightarrow 1$	0.24
b	1	$3 \cdot 10^{-}35$	$1 \rightarrow 3$	0.77
a	3	0.6		

probability: ≈ 0

Emit:	while in state:	prob	transition	prob
a	3	0.6	$3 \rightarrow 1$	0.37
r	1	0.34	$1 \rightarrow 2$	0.22
b	2	0.06	$2 \rightarrow 3$	0.75
a	3	0.6		

probability: 0.0012

6. What kind of linguistics is this?

It is an approach to linguistic analysis which is non-cognitivist:

It makes no claims about hidden or occult properties of the human system (for which linguistic tools are not designed to provide answers).

It welcomes psychologists, without claiming to replace them, or to do their job.

It asks linguists to study language as a natural phenomenon, and to evaluate their success like any other natural science.

I have not addressed two important areas of phonology: automatic morphophonology, and the geometry of phonological representations.

That will have to wait à la prochaine.

6. What kind of linguistics is this?

Facts about a language L may be divided into (type 1) those facts that are particular to L, and

(type 2) those that are shared by *all* languages.

In all likelihood, *type 1* information is vastly larger than *type 2* information.

6. What is this?

Type 1 information is:

universal;

in all likelihood, not learned, and not even learnable in a short time period;

innate;

not influenced by historical or cultural concerns.

6. What is this?

It seems clear to me that linguistics is the study of both Type 1 and Type 2 information. Much of the focus in linguistic theory has focused on Type 1 information (what is common to all acquisition paths).

This work

6. What is this?

Linguistics seeks the *essence* common to all languages. This essence can exist nowhere other than in the biological nature of the human being. This essence does not need to be learned. This essence can probably not be learned (in a reasonable time). This essence is UG.

6. What is this?

- Linguistics seeks to analyze each human language. Languages vary, due to their history, to their speakers' history, and to the ends to which they are put. Finding ways to characterize each language adequately is the primary goal of linguistics; it is best accomplished by analyzing linguistic data in the same way that other sciences proceed, ceteris paribus.