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...
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?
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 \( x \).
• 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 **Ginger** blah blah blah 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.
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 $L \subseteq A^*$, 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 $A \subseteq L$: 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
1. Word segmentation

(Lexicon)

L*: All strings of words

A*: All strings of letters

(Alphabet)
au début était le verbe

L*: All strings of words

A*: All strings of letters
1. Word segmentation

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

$L^*$: All strings of words

$A^*$: All strings of letters

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

(Alphabet)

(Lexicon)
The distribution $p$ over $L$ is extended to a distribution $p^*$ over $L^*$ in the natural way:

- We assume a probability distribution $\lambda$ over sentence length $l$:
  $$\sum_{i=1}^{\lambda(l)+1} \lambda(i) = 1$$

- 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.

$$p(U \mid L) = \arg \max_{q \in \{ \text{parses}(U) \}} \, pr(q)$$

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

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

That would be reasonable.
Calculating either argmax or sum requires dynamic programming techniques.
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:

\[
\hat{L} = \arg \max_{L \in A^*, \Pr} \Pr_{\hat{L}}(C | L)
\]

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 $\pi(\cdot)$ 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: $\pi(G) = 2^{-\text{length}(\text{En}(G))}$
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**

\[
pr(G \mid C) = \frac{pr(C \mid G)pr(G)}{pr(C)} = \frac{p_G(C)\pi(G)}{pr(C)} = \frac{p_G(C)\pi(G)}{\int p_g(C)\pi(g)dg}
\]

\[
\log pr(G \mid C) = \log p_G(C) - H(G) - K.
\]

- **log prob of corpus, in grammar G**
- **Length of G’s encoding**
\[ \log \text{pr}(G \mid C) = \log p_G(C) - H(G) - K. \]

log prob of corpus, in grammar G

Length of G’s encoding

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

\[ \|G\| \approx \sum_{w \in G} |w| \ast \log(26) \]
It is sensible to call $-\log \text{prob} (X) \equiv \frac{1}{\text{prob } x}$ 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 $G$*:

$$[- \log \text{prob} (C \mid G)] + [\text{length}(\text{Enc}(G))]$$

= Compressed length of corpus
+ compressed length of grammar
= $-\log \text{prob} (G\mid C) + \text{a constant}$
How one talks in MDL...

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

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

\[
- \log \text{prob } (C \mid G) + \text{length}(\text{Enc}(G))
\]

\(\text{= Compressed length of corpus} + \text{compressed length of grammar} = -\log \text{prob } (G \mid C) + \text{a constant}\)

\(\text{= evaluation metric of early generative grammar}\)
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
Essence of MDL

![Bar chart showing the essence of MDL with different categories and their corresponding values.]

- Best analysis
- Elegant theory that works badly
- Baroque theory modeled on data

Legend:
- Length of morphology
- Log prob of corpus
2.2 Search heuristic

Easy!

start \textit{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.
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}
Notation:

- \([t]\) = count of \(t\)
- \([h]\) = count of \(h\)
- \([\text{th}]\) = count of \(\text{th}\)
- \(Z\) = total number of words (tokens)

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

Log probability of corpus:

\[\sum_{m \in \text{lexicon}} [m] \log \frac{[m]}{Z}\]
Log prob of sentence C

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

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

All letters are separate

\[ [t]_2 = [t]_1 - [th] \]
\[ [h]_2 = [h]_1 - [th] \]
\[ [Z]_2 = [Z]_1 - [th] \]

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

th is treated as a separate chunk
\[
[t]_1 \log \frac{[t]_1}{Z_1} + \sum_{m \neq t, h} [m] \log \frac{[m]}{Z_1} \\
+[h]_1 \log \frac{[h]_1}{Z_1} \\
+ \sum_{m \neq t, h} [m] \log \frac{[m]}{Z_1} \\
+[th]_2 \log \frac{[th]_2}{Z_2} \\
+[h]_2 \log \frac{[h]_2}{Z_2} \\
+[t]_2 \log \frac{[t]_2}{Z_2}
\]

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

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

This is \textbf{positive} if \text{Lexicon 2 is better}.
Effect of having fewer “words” altogether

\[ \Delta f = \log \frac{f_2}{f_1} \]

\[ \Delta pr(C) = -Z_1 \Delta Z + [t] \Delta t + [h] \Delta h + [th] \log \frac{pr_2(th)}{pr_2(t) pr_2(h)} \]

This is \textbf{positive} if Lexicon 2 is better.
Effect of frequency of /t/ and /h/ decreasing

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

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

This is **positive** if Lexicon 2 is better.
Effect /th/ being treated as a unit rather than separate pieces

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

\[ -Z_1 \Delta Z + [t] \Delta t + [h] \Delta h + [th] \log \frac{\text{pr}_2(\text{th})}{\text{pr}_2(t) \text{pr}_2(h)} \]

This is **positive** if Lexicon 2 is better.
2.3 Results

• The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place.

• The jury further said in term-end presentations that the City Executive Committee, which had overall charge of the election, deserved praise and thanks of the City of Atlanta for the manner in which the election was conducted.

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: sing 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

\[ M[\text{morphology}] = \{ \text{Stems } T, \text{Affixes } F, \text{Signatures } \Sigma \} \]

\[ |M| = |T| + |F| + |\Sigma| \]

**Stems** \[ |T| = \sum_{t \in T} |t| \]

**Affixes** \[ |F| = \sum_{f \in F} |f| \]

**Sig’s** \[ |\Sigma| = \sum_{\sigma \in T} |\sigma| \]

\[ |s| = \text{string length}(s) \times \log(26) \]

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

What is a signature, and what is its length?
What is a signature?

- account
- appeal
- attack
- 40 more...

- NULL
- ed
- ing

- NULL
- NULL
- es
- es

- élevé
- équipé
- étonnant
- 78 more
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 $-\log \text{freq}(p)$:

So the total length of the signatures is:

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

Sum over signatures

Sum over stem ptrs
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:

\[
\begin{align*}
\{ \text{jump} \} & \quad \{ \text{NULL} \} \\
\{ \text{walk} \} & \quad \{ \text{ed} \} \\
& \quad \{ \text{ing} \}
\end{align*}
\]
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.
### Signature Counts and Exemplars

<table>
<thead>
<tr>
<th>Signature</th>
<th>Exemplar</th>
<th>Corpus Count</th>
<th>Stem Count</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL.s</td>
<td>abuse</td>
<td>1793</td>
<td>445</td>
<td>3967</td>
</tr>
<tr>
<td>ed</td>
<td>accelerat</td>
<td>1657</td>
<td>457</td>
<td>1114</td>
</tr>
<tr>
<td>ing</td>
<td>embezzl</td>
<td>1046</td>
<td>258</td>
<td>1047</td>
</tr>
<tr>
<td>NULL.ly</td>
<td>absolute</td>
<td>369</td>
<td>101</td>
<td>961</td>
</tr>
<tr>
<td>: ly</td>
<td>alarming</td>
<td>1119</td>
<td>148</td>
<td>294</td>
</tr>
<tr>
<td>er</td>
<td>14-pow</td>
<td>4726</td>
<td>424</td>
<td>858</td>
</tr>
<tr>
<td>NULL.ed.ing.s</td>
<td>account</td>
<td>484</td>
<td>35</td>
<td>798</td>
</tr>
<tr>
<td>: NULL.ed.ing</td>
<td>approach</td>
<td>263</td>
<td>40</td>
<td>649</td>
</tr>
<tr>
<td>: NULL.ed.s</td>
<td>affect</td>
<td>282</td>
<td>43</td>
<td>620</td>
</tr>
</tbody>
</table>

### 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
3. Morphology

Refinements

1. Correct errors in segmentation

\[
\begin{align*}
\text{affirmati} & \quad \text{on} \\
\text{agressi} & \quad \text{ve} \\
\text{attenti} & \quad \text{ive} \\
20 \text{ more} & \quad \text{20 more}
\end{align*}
\]

\[
\Rightarrow
\begin{align*}
\text{affirm} & \\
\text{agress} & \\
\text{attent} & \\
\text{ive}
\end{align*}
\]

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.

Diagram:

- Minilexicon 1
  - Words$_1$
  - Affixes
  - Stems$_1$
  - Signatures$_1$

- Minilexicon 2
  - Words$_2 = \text{Stems}_1$
  - Affixes$_2$
  - Stems$_2$
  - Signatures$_2$
### 3. Morphology

## French roots

<table>
<thead>
<tr>
<th>Stems</th>
<th>Corpus count</th>
<th>Prefix</th>
<th>Suffix sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>abricot</td>
<td>6</td>
<td></td>
<td>NULL-ier</td>
</tr>
<tr>
<td>accept</td>
<td>3</td>
<td></td>
<td>NULL-eur</td>
</tr>
<tr>
<td>acheuléen</td>
<td>4</td>
<td></td>
<td>NULL-ne</td>
</tr>
<tr>
<td>acryl</td>
<td>11</td>
<td></td>
<td>NULL-i que</td>
</tr>
<tr>
<td>actuel</td>
<td>10</td>
<td></td>
<td>NULL-le</td>
</tr>
<tr>
<td>adaptat</td>
<td>29</td>
<td></td>
<td>NULL-eur-ion</td>
</tr>
<tr>
<td>administr</td>
<td>2</td>
<td></td>
<td>NULL-at</td>
</tr>
<tr>
<td>administrat</td>
<td>11</td>
<td></td>
<td>NULL-eur-ion</td>
</tr>
<tr>
<td>adopt</td>
<td>5</td>
<td></td>
<td>NULL-ant</td>
</tr>
<tr>
<td>africa</td>
<td>38</td>
<td></td>
<td>NULL-in</td>
</tr>
<tr>
<td>agglomér</td>
<td>5</td>
<td></td>
<td>NULL-ation</td>
</tr>
<tr>
<td>amélior</td>
<td>4</td>
<td></td>
<td>NULL-ation</td>
</tr>
<tr>
<td>améri</td>
<td>8</td>
<td></td>
<td>NULL-que</td>
</tr>
<tr>
<td>améric a</td>
<td>45</td>
<td></td>
<td>NULL-in</td>
</tr>
<tr>
<td>Words</td>
<td>Stem</td>
<td>Mini-Lexicon 3</td>
<td>Mini-Lexicon 2</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>decline</td>
<td>declin</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>declined</td>
<td>declin</td>
<td></td>
<td>ed</td>
</tr>
<tr>
<td>declines</td>
<td>declin</td>
<td></td>
<td>es</td>
</tr>
<tr>
<td>decolletage</td>
<td>decolletage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decor</td>
<td>decor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decorate</td>
<td>decor</td>
<td>at</td>
<td>e</td>
</tr>
<tr>
<td>decorating</td>
<td>decor</td>
<td>at</td>
<td>ing</td>
</tr>
<tr>
<td>decoration</td>
<td>decor</td>
<td>at</td>
<td>ion</td>
</tr>
<tr>
<td>decorations</td>
<td>decor</td>
<td>at</td>
<td>ion</td>
</tr>
<tr>
<td>decorative</td>
<td>decor</td>
<td>at</td>
<td>ive</td>
</tr>
<tr>
<td>decorator</td>
<td>decor</td>
<td>at</td>
<td>or</td>
</tr>
<tr>
<td>decorators</td>
<td>decor</td>
<td>at</td>
<td>or</td>
</tr>
<tr>
<td>decrease</td>
<td>decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decree</td>
<td>decree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de decreeing</td>
<td>decree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decried</td>
<td>decri</td>
<td></td>
<td>ed</td>
</tr>
<tr>
<td>decries</td>
<td>decri</td>
<td></td>
<td>es</td>
</tr>
<tr>
<td>dedicated</td>
<td>dedicat</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Detect allomorphy

Signature: \(<e>\text{ion} \ . \ 	ext{NULL}\)

<table>
<thead>
<tr>
<th>composite</th>
<th>concentrate</th>
<th>corporate</th>
<th>détente</th>
</tr>
</thead>
<tbody>
<tr>
<td>discriminate</td>
<td>evacuate</td>
<td>inflate</td>
<td>opposite</td>
</tr>
<tr>
<td>participate</td>
<td>probate</td>
<td>prosecute</td>
<td>tense</td>
</tr>
</tbody>
</table>

What is this?

**composite** and **composition**

**composite \rightarrow composit \rightarrow 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.
4. Morphology

Swahili verb

ni          u          a          tu          wa
li          ka          m          tu          wa
ni          ku
sem
imb
pend
fik
som
on
l
chaku
4. Morphology

Swahili verb

Subject marker

ni
u
a
tu
wa
li
ka
m
ni
ku
tu
wa
sem
imb
pend
fik
som
on
chaku
Ø
w
a
4. Morphology

Swahili verb

Subject marker

Tense marker

ni
u
a
tu
wa
li
ka
ta
taka
na
ni
ku
m
tu
wa
ni
ku
m
tu
wa

imb
pend
fik
sem
ø
som
on
l
chaku

w

Ø

a
4. Morphology

Swahili verb

Subject marker:
- ni
- u
- a
- tu
- wa

Tense marker:
- li
- ka
- ta
- taka
- na

Object marker:
- ni
- ku
- m
- tu
- wa

Imb:
- pend
- fik
- sem
- som
- on
- l
- chaku

w
Ø
a
4. Morphology

Swahili verb

Subject marker

Tense marker

Object marker

Root

ni
u
a
wa
li
ka
ta
taka
na
ni
ku
m
tu
wa
w
Ø

imb
pend
fik
sem
som
on
l
chaku
4. Morphology

Swahili verb

Subject marker

Tense marker

Object marker

Root

Voice (active/passive)
4. Morphology

Swahili verb

Subject marker

Tense marker

Object marker

Root

Voice (active/passive)

Final vowel
Finite state automaton (FSA)
Signature: reduces false positives

\{
\textit{jump}, \textit{walk}
\}\quad\{\quad \textit{ed}, \textit{ing} \quad \}

\{ \textit{NULL} \}

4. Morphology
Generalize the signature...

Sequential FSA: each state has a unique successor.
4. Morphology

Alignments

1 \rightarrow m_1 \rightarrow 2 \rightarrow m_2 \rightarrow 3 \rightarrow m_4 \rightarrow 4
Alignments: String edit distance algorithm
4. Morphology

Alignments: make cuts

\[ \text{nilili} \quad \text{mupenda} \]

\[ \text{nitakamura} \quad \text{mupenda} \]
4. Morphology

Elementary alignment

1 \rightarrow m_1 \rightarrow 2 \rightarrow m_2 \rightarrow 3 \rightarrow m_3 \rightarrow 3 \rightarrow m_4 \rightarrow 4
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

<table>
<thead>
<tr>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>a, wa</em> (sg., pl. human subject markers)</td>
<td>246 stems</td>
<td></td>
</tr>
<tr>
<td><em>ku, hu</em> (infinitive, habitual markers)</td>
<td>51 stems</td>
<td></td>
</tr>
<tr>
<td><em>wa</em> (pl. subject marker)</td>
<td><em>ka, li</em> (tense markers)</td>
<td>25 stems</td>
</tr>
<tr>
<td><em>a</em> (sg. subject marker)</td>
<td><em>ka, li</em> (tense markers)</td>
<td>29 stems</td>
</tr>
<tr>
<td><em>a</em> (sg. subject marker)</td>
<td><em>ka, na</em> (tense markers)</td>
<td>28 stems</td>
</tr>
<tr>
<td>37 strings</td>
<td><em>w</em> (passive marker) / Ø</td>
<td><em>a</em></td>
</tr>
<tr>
<td>Method</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>String edit distance</td>
<td>0.77</td>
<td>0.57</td>
</tr>
<tr>
<td>Stem-affix</td>
<td>0.54</td>
<td>0.14</td>
</tr>
<tr>
<td>Affix-stem</td>
<td>0.68</td>
<td>0.20</td>
</tr>
</tbody>
</table>
## Collapsed templates

<table>
<thead>
<tr>
<th></th>
<th>One Template</th>
<th>The other template</th>
<th>Collapsed Template</th>
<th>% found on Yahoo search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{a}-{ka,na}-{stems}</td>
<td>{a}-{ka,ki}-{stems}</td>
<td>{a}-{ka,ki,na}-{stems}</td>
<td>86 (37/43)</td>
</tr>
<tr>
<td>2</td>
<td>{wa}-{ka,na}-{stems}</td>
<td>{wa}-{ka,ki}-{stems}</td>
<td>{wa}-{ka,ki,na}-{stems}</td>
<td>95 (21/22)</td>
</tr>
<tr>
<td>3</td>
<td>{a}-{ka,ki,na}-{stems}</td>
<td>{wa,ki,li}-{stems}</td>
<td>{a,wa}-{ka,ki,na}-{stems}</td>
<td>84 (154/183)</td>
</tr>
<tr>
<td>4</td>
<td>{a}-{liye,me}-{stems}</td>
<td>{a}-{liye,li}-{stems}</td>
<td>{a}-{liye,li,me}-{stems}</td>
<td>100 (21/21)</td>
</tr>
<tr>
<td>5</td>
<td>{a}-{ki,li}-{stems}</td>
<td>{wa}-{ki,li}-{stems}</td>
<td>{a,wa}-{ki,li}-{stems}</td>
<td>90 (36/40)</td>
</tr>
<tr>
<td>6</td>
<td>{a}-{lipo,li}-{stems}</td>
<td>{wa}-{lipo,li}-{stems}</td>
<td>{a,wa}-{lipo,li}-{stems}</td>
<td>90 (27/30)</td>
</tr>
<tr>
<td>7</td>
<td>{a,wa}-{ki,li}-{stems}</td>
<td>{a,wa}-{lipo,li}-{stems}</td>
<td>{a,wa}-{ki,lipo,li}-{stems}</td>
<td>74 (52/70)</td>
</tr>
<tr>
<td>8</td>
<td>{a}-{na,naye}-{stems}</td>
<td>{a}-{na,ta}-{stems}</td>
<td>{a}-{na,ta,naye}-{stems}</td>
<td>80 (12/15)</td>
</tr>
</tbody>
</table>
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: \(36 - 17 = 19\).
Most edges are convergent...

adjectives

car-
rubi-
negr-
pequeñ-

verbs

com-
cre-

-emos

-tems
But some diverge (Spanish):

car-
pequeñ-
rubi-
negr-

Participle-forming suffix

com-
cre-

O
4. Morphology

English has much the same:

- laugh
- jump
- walk
- book
- chair
- table

NULLs

NULLing

NULLed

accept

accept
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.
So far we have said little about phonology.

We have assumed no interesting probabilistic model of segment (=phoneme) placement. (0\textsuperscript{th} or 1\textsuperscript{st} 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.
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

consonants

1

.29

.71

.98

2

vowels

.02
Pr (State 1 → State 1)

Pr (State 2 → State 2)

alternating region

Rhythm, syllabification

harmony region
Dynamics of English 2 state transitions (learning)
**English**: Log ratios of the emission probabilities of the 2 states:

\[
\log \frac{p_1(\phi)}{p_2(\phi)}
\]

Here is a table showing the log ratios for various ArpaBet symbols:

<table>
<thead>
<tr>
<th>ArpaBet</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH</td>
<td>-999</td>
</tr>
<tr>
<td>NG</td>
<td>-999</td>
</tr>
<tr>
<td>W</td>
<td>-999</td>
</tr>
<tr>
<td>N</td>
<td>-999</td>
</tr>
<tr>
<td>L</td>
<td>-999</td>
</tr>
<tr>
<td>HH</td>
<td>-999</td>
</tr>
<tr>
<td>SH</td>
<td>-999</td>
</tr>
<tr>
<td>R</td>
<td>-999</td>
</tr>
<tr>
<td>M</td>
<td>-999</td>
</tr>
<tr>
<td>V</td>
<td>-999</td>
</tr>
<tr>
<td>ZH</td>
<td>-999</td>
</tr>
<tr>
<td>JH</td>
<td>-999</td>
</tr>
<tr>
<td>B</td>
<td>-999</td>
</tr>
<tr>
<td>Y</td>
<td>-999</td>
</tr>
<tr>
<td>F</td>
<td>-999</td>
</tr>
<tr>
<td>G</td>
<td>-829</td>
</tr>
<tr>
<td>K</td>
<td>-576</td>
</tr>
<tr>
<td>CH</td>
<td>-361</td>
</tr>
<tr>
<td>TH</td>
<td>-5.19</td>
</tr>
<tr>
<td>P</td>
<td>-4.37</td>
</tr>
<tr>
<td>D</td>
<td>-3.95</td>
</tr>
<tr>
<td>S</td>
<td>-2.75</td>
</tr>
<tr>
<td>T</td>
<td>-2.20</td>
</tr>
<tr>
<td>Z</td>
<td>-1.37</td>
</tr>
<tr>
<td>UW0</td>
<td>2.22</td>
</tr>
<tr>
<td>ERO</td>
<td>2.30</td>
</tr>
<tr>
<td>IY0</td>
<td>2.31</td>
</tr>
<tr>
<td>AWO</td>
<td>2.32</td>
</tr>
<tr>
<td>AYO</td>
<td>2.83</td>
</tr>
<tr>
<td>OWO</td>
<td>3.93</td>
</tr>
<tr>
<td>EY0</td>
<td>4.99</td>
</tr>
<tr>
<td>AY1</td>
<td>5.11</td>
</tr>
<tr>
<td>OY1</td>
<td>5.81</td>
</tr>
<tr>
<td>IY1</td>
<td>7.39</td>
</tr>
<tr>
<td>OW1</td>
<td>12.7</td>
</tr>
<tr>
<td>AW1</td>
<td>275</td>
</tr>
</tbody>
</table>

The log ratios for EY1, UYO, UW1, AH0, EHO, AEO, ER1, AAO, IHO, AE1, A00, EH1, AA1, A01, IH1, AH1, UH1, UHO are 262, 263, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, respectively.

In general, log ratios close to 0 indicate neutral emission probabilities, while positive log ratios indicate a positive correlation, and negative log ratios indicate a negative correlation.
5. Phonology

English

Entropy

Transition probability

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

\[
\log \frac{p_1(\phi)}{p_2(\phi)}
\]

<table>
<thead>
<tr>
<th>Phone</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ə</td>
<td>-999</td>
</tr>
<tr>
<td>e</td>
<td>-999</td>
</tr>
<tr>
<td>ð</td>
<td>-999</td>
</tr>
<tr>
<td>u</td>
<td>-999</td>
</tr>
<tr>
<td>i</td>
<td>-999</td>
</tr>
<tr>
<td>å</td>
<td>-999</td>
</tr>
<tr>
<td>ê</td>
<td>-999</td>
</tr>
<tr>
<td>ò</td>
<td>-999</td>
</tr>
<tr>
<td>a</td>
<td>-473</td>
</tr>
<tr>
<td>y</td>
<td>-11.6</td>
</tr>
<tr>
<td>o</td>
<td>-10.5</td>
</tr>
<tr>
<td>ōe</td>
<td>-5.53</td>
</tr>
<tr>
<td>e</td>
<td>-4.93</td>
</tr>
</tbody>
</table>

**Negative**

<table>
<thead>
<tr>
<th>Phone</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>5.26</td>
</tr>
<tr>
<td>t</td>
<td>7.96</td>
</tr>
<tr>
<td>g</td>
<td>600</td>
</tr>
<tr>
<td>d</td>
<td>933</td>
</tr>
<tr>
<td>k</td>
<td>999</td>
</tr>
<tr>
<td>ə</td>
<td>999</td>
</tr>
<tr>
<td>n</td>
<td>999</td>
</tr>
<tr>
<td>m</td>
<td>999</td>
</tr>
<tr>
<td>l</td>
<td>999</td>
</tr>
<tr>
<td>f</td>
<td>999</td>
</tr>
</tbody>
</table>

**Positive**
Figure 11: Dynamics of learning French c/v
Finnish: Log ratios of the emission probabilities of the 2 states:
\[
\log \frac{p_1(\phi)}{p_2(\phi)}
\]

<table>
<thead>
<tr>
<th>Phone</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ə</td>
<td>-999</td>
</tr>
<tr>
<td>ɛ</td>
<td>-999</td>
</tr>
<tr>
<td>õ</td>
<td>-999</td>
</tr>
<tr>
<td>u</td>
<td>-999</td>
</tr>
<tr>
<td>i</td>
<td>-999</td>
</tr>
<tr>
<td>â</td>
<td>-999</td>
</tr>
<tr>
<td>ê</td>
<td>-999</td>
</tr>
<tr>
<td>ō</td>
<td>-999</td>
</tr>
<tr>
<td>a</td>
<td>-473</td>
</tr>
<tr>
<td>y</td>
<td>-11.6</td>
</tr>
<tr>
<td>o</td>
<td>-10.5</td>
</tr>
<tr>
<td>ōe</td>
<td>-5.53</td>
</tr>
<tr>
<td>e</td>
<td>-4.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phone</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>5.26</td>
</tr>
<tr>
<td>t</td>
<td>7.96</td>
</tr>
<tr>
<td>g</td>
<td>600</td>
</tr>
<tr>
<td>p</td>
<td>933</td>
</tr>
<tr>
<td>d</td>
<td>999</td>
</tr>
<tr>
<td>k</td>
<td>999</td>
</tr>
<tr>
<td>s</td>
<td>999</td>
</tr>
<tr>
<td>m</td>
<td>999</td>
</tr>
<tr>
<td>n</td>
<td>999</td>
</tr>
<tr>
<td>l</td>
<td>999</td>
</tr>
<tr>
<td>f</td>
<td>999</td>
</tr>
<tr>
<td>b</td>
<td>999</td>
</tr>
<tr>
<td>r</td>
<td>999</td>
</tr>
<tr>
<td>ň</td>
<td>999</td>
</tr>
<tr>
<td>ñ</td>
<td>999</td>
</tr>
<tr>
<td>v</td>
<td>999</td>
</tr>
<tr>
<td>j</td>
<td>999</td>
</tr>
<tr>
<td>h</td>
<td>999</td>
</tr>
<tr>
<td>ū</td>
<td>999</td>
</tr>
<tr>
<td>w</td>
<td>999</td>
</tr>
<tr>
<td>j</td>
<td>999</td>
</tr>
<tr>
<td>z</td>
<td>999</td>
</tr>
</tbody>
</table>

negative       positive
5. Phonology

Finnish vowels and their harmony

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ö</td>
<td>999</td>
</tr>
<tr>
<td>ä</td>
<td>961</td>
</tr>
<tr>
<td>y</td>
<td>309</td>
</tr>
<tr>
<td>e</td>
<td>0.655</td>
</tr>
<tr>
<td>i</td>
<td>0.148</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>-7.66</td>
</tr>
<tr>
<td>a</td>
<td>-927</td>
</tr>
<tr>
<td>u</td>
<td>-990</td>
</tr>
<tr>
<td>From State 1</td>
<td>Prob</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>a</td>
<td>0.17</td>
</tr>
<tr>
<td>e</td>
<td>0.15</td>
</tr>
<tr>
<td>i</td>
<td>0.15</td>
</tr>
<tr>
<td>œ</td>
<td>0.15</td>
</tr>
<tr>
<td>o</td>
<td>0.087</td>
</tr>
<tr>
<td>ε</td>
<td>0.058</td>
</tr>
<tr>
<td>2</td>
<td>0.056</td>
</tr>
<tr>
<td>y</td>
<td>0.043</td>
</tr>
<tr>
<td>4</td>
<td>0.036</td>
</tr>
<tr>
<td>I</td>
<td>0.027</td>
</tr>
<tr>
<td>u</td>
<td>0.026</td>
</tr>
<tr>
<td>œ</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Emission probabilities, 3 state HMM for French
<table>
<thead>
<tr>
<th>Emit:</th>
<th>while in state:</th>
<th>prob</th>
<th>transition</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td>3 → 2</td>
<td>0.62</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>0.06</td>
<td>2 → 1</td>
<td>0.24</td>
</tr>
<tr>
<td>r</td>
<td>1</td>
<td>0.34</td>
<td>1 → 3</td>
<td>0.77</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emit:</td>
<td>while in state:</td>
<td>prob</td>
<td>transition</td>
<td>prob</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td>3 → 1</td>
<td>0.37</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3 \cdot 10^{-35}</td>
<td>1 → 2</td>
<td>0.22</td>
</tr>
<tr>
<td>r</td>
<td>2</td>
<td>0.06</td>
<td>2 → 3</td>
<td>0.75</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emit:</td>
<td>while in state:</td>
<td>prob</td>
<td>transition</td>
<td>prob</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td>3 → 2</td>
<td>0.62</td>
</tr>
<tr>
<td>r</td>
<td>2</td>
<td>0.06</td>
<td>2 → 1</td>
<td>0.24</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3 \cdot 10^{-35}</td>
<td>1 → 3</td>
<td>0.77</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emit:</td>
<td>while in state:</td>
<td>prob</td>
<td>transition</td>
<td>prob</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td>3 → 1</td>
<td>0.37</td>
</tr>
<tr>
<td>r</td>
<td>1</td>
<td>0.34</td>
<td>1 → 2</td>
<td>0.22</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>0.06</td>
<td>2 → 3</td>
<td>0.75</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.
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.
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.
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
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.
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.