Linguistics 106, lecture notes
Introduction to distributional analysis

30 May 2002

1 What is a morpheme?

- As a first approximation, **morphemes** can be defined as the minimal pairings of sound and meaning.
  Or, as Langacker puts it in your reading, the smallest structurally significant units that recur with constant meaning.

- By “minimal” and “smallest” we mean:
  No subpart of the sound of the morpheme denotes a subpart of the meaning of the morpheme.

**Example**

The sequence of phonemes [kæts] means ‘cats’, i.e. the plural of ‘cat’.

We can isolate two parts of [kæts] which have meanings that are part of the meaning ‘cats’: (1) [kæt], which means ‘cat’; and (2) [s], which means plural.

There are no subparts of [s] that have meanings that are part of the meaning ‘plural’.
Hence the [s] in [kæts] is a **minimal** pair of sound and meaning, a morpheme.

There are subparts of [kæt] that have meanings **outside this context**. For example, [æt] means ‘at’. But ‘at’ is not part of the meaning of ‘cat’.

Clearly, there are no subparts of [kæt] that are part of the meaning ‘cat’. Hence the [kæt] in [kæts] is a **minimal** pair of sound and meaning, a morpheme.

**Question**: How many morphemes are there in *catastrophe*?

2 Matching sounds to meanings

**Question**: Given a text in an unknown language, how can we isolate the morphemes with which it is composed?

- Well, if we have a translation of that text into a known language, we can try to match parts of the text to parts of the translation.

- To put it another way: If we have an analysis of the text’s meaning, we can try to match parts of the text’s sound to parts of its meaning.
Example

(1) Kayah Li (a language of Burma and Thailand, somewhat related to Burmese)

a. vēchūthwi
   ‘I stabbed the dog.’

b. vēphēthwi
   ‘I hugged the dog.’

c. ?açuÕathwi
   ‘He stabbed his (own) dog.’

d. ?açužlhthwi
   ‘He stabbed his dog.’ (i.e. some other guy’s dog)

• Notice, to do the deciphering, we need several sentences which differ minimally from one another.

• We also need extremely close translations.

   Ideally, there should be a translation for each individual morpheme in the original text. But such translations are not always available, or easy to give.

Example

Consider the previous example again.

The English translations in (1) are good translations; they are not approximate or metaphorical.

Still, notice that, at least in (a–c), they have fewer morphemes than the original Kayah Li sentences. For example, in (1a), the Kayah Li sentence has three morphemes. How many does the English have?

Also, consider how exactly you would translate “?a” into English.

• Let’s do another example to reinforce the point.

Example

(2) Deciphering Igbo (a language of Nigeria, somewhat related to Yoruba)

a. ìwàáŋwook’ògáraàhya
   ‘My son went to market.’

b. ìwàghìjoòok’ògáraàhya
   ‘Your son went to market.’

c. ìwàghìjoàanyjìgáraàhya
   ‘Your daughter went to market.’

d. ìwàghìjoàanyjìzuraàhya
   ‘Your daughter shopped at the market.’

• Another difficulty of the meaning-matching approach is that not all morphemes have meanings. Some are just grammatical markers. We will see such examples in future classes.
3 Distributional analysis: Introduction

Question: How can we isolate the morphemes of a text without having a translation?

- This question is interesting for at least two reasons:
  - Its pursuit turns up significant facts about the structure of human languages; and
  - Infants apparently do exactly this: while having only a very partial grasp of the meaning of what is said to them, they find the words of the language they are acquiring.

- Let’s try to find the answer for ourselves. Try to isolate morphemes in the following example.

  **Example**

  (3) Serbo-Croatian (a Slavic language of Serbia and Croatia)
  
  a. yáchitam
  
  b. ýápíyem
  
  c. tícítaš
  
  d. vícítate
  
  e. óncíta
  
  f. ónipíyu
  
  g. típyeš
  
  h. ónpiye
  
  i. mipíyemo
  
  j. ónicítayu
  
  k. típisíš
• In trying to segment these sentences without the aid of translations, we naturally arrive at (at least) the following strategies:

1. Find sequences that repeat in different contexts.
   
   **Example**
   
   ya repeats in the following two contexts: ______ $citam$ and ______ $p\acute{y}em$.
   
   $pi\acute{y}$ repeats in: $ya\underline{em}$; $\acute{o}ni\underline{u}$; $\acute{on}\underline{e}$; and $ti\underline{es}$.
   
   $cit$ repeats in: $ya\underline{am}$; $ti\underline{as}$; $\acute{vi}\underline{ate}$; $\acute{on}\underline{a}$; and $\acute{on}\underline{ayu}$.
   
   **Question:** What other such sequences do we find?

   We infer that such patterns of repetition are not accidental, and diagnose significant units of the language.

2. In general, ignore sequences that repeat, but are very often or always right next to some other sequence.
   
   **Example**
   
   $c$ repeats frequently, but always with $\textit{-it}$: $\textit{cit}$.
   
   $\acute{y}g$ repeats frequently, but always right next to $p\textit{-}p\acute{y}$

   We infer that these patterns of cooccurrence are not accidental; a sequence that can never be separated from its immediate context is not likely to be a significant unit of the language.

3. Find contexts that repeat with different things in them—that is ‘slots’ into which a number of different sequences can be substituted.
   
   **Example**
   
   The context $\acute{on}\underline{i\underline{u}}$ repeats twice: $\acute{on}\acute{p}\acute{y}\acute{u}$; $\acute{on}\acute{i\acute{c}\acute{i\acute{t}\acute{a}}\acute{y}}u$
   
   The context $ti\underline{\underline{s}}$ repeats three times: $ti\acute{c}\acute{t}\acute{a}s$; $ti\acute{p}\acute{y}\acute{e}\acute{s}$; $ti\acute{p}\acute{\acute{a}}\acute{s}\acute{i}$.

   We infer that what goes into a repeating context must be an isolable and significant unit of the language.

   **Example**
   
   The sequence $p\acute{u}\acute{\acute{s}}\acute{i}$ occurs in the repeating context $ti\underline{\underline{s}}$. So we infer that $p\acute{u}\acute{\acute{s}}\acute{i}$ is isolable, and significant—even though it itself never repeats in the sentences above.

4. When several repeating sequences repeat in the same several contexts, this is very strong evidence that each of these repeating sequences is an isolable and significant unit:
   
   **Example**
   
   The two sequences $p\acute{y}\acute{e}$ and $\acute{c}ita$, repeat in the three contexts $ya\underline{m}$; $ti\underline{s}$; and $\acute{on}\underline{a}$.
   
   They also ‘almost’ repeat in the context $\acute{on}\underline{i\underline{u}}$: $\acute{on}\acute{p}\acute{y}\acute{u}$, $\acute{on}\acute{i\acute{c}\acute{i\acute{t}\acute{a}}\acute{y}}u$.

   We infer that such regular patterns of occurrence cannot be accidental, and diagnose isolable and significant units of the language.
• It will be useful to have the following terms at our disposal:

**Environment**  The environment of \(X\) is what above we called the context of \(X\). It is just the sequence in which \(X\) occurs, minus \(X\) itself.

**Example**
In the sequence \(yáč\textit{it}\z\textit{am}\), the environment of \(it\) is: \(yáč\textit{am}\).

**Total Independence**  A sequence \(X\) is totally independent in environment \(E\) if \(X\) can be replaced by some other sequence \(Y\), without changing anything in \(E\).

**Example**
The sequence \(č\textit{ita}\) is totally independent in the environment \(yáč\textit{it}\z\textit{am}\), since it can be replaced by \(pi\textit{ye}\), without changing the environment.

**Dependency**  A sequence \(X\) exhibits a dependency in environment \(E\) if it cannot be replaced without changing something in \(E\).

**Example**
The sequence \(yá\) exhibits a dependency in the environment \(yáč\textit{ita}\). It can be replaced in this environment by \(ti\):

\[\begin{array}{c}
ti\ \ \ č\textit{ita}s\\
\text{but then the environment must also change: the } m\text{ following } č\textit{ita}\text{ must change to } s.
\end{array}\]

**Distribution**  The distribution of a sequence \(X\) is the set of all the environments in which it occurs.

**Example**
The distribution of the sequence \(č\textit{ita}\) in the dataset above is the following set of three environments:

\[\begin{array}{c}
yá\ \ \ m\\
ti\ \ \ s\\
č\textit{ita}\\
óni\ \ \ y\mu
\end{array}\]

**Distributional class**  The distributional class of a sequence \(X\) is the class of all other sequences which share the same distribution as \(X\).

**Example**
Based on the dataset above, \(č\textit{ita}\) and \(pi\textit{ye}\) belong to the same distributional class. They occur in the same set of environments.\(^1\)

• We can summarize the results of our analysis of the Serbo-Croatian data by saying:

To isolate morphemes in a text, find independent sequences which share their distribution with other independent sequences.

• We will return to this idea repeatedly in the weeks to come.

\(^1\)Actually, \(pi\textit{ye}\) they don't both occur in the environment \(óni\ \ \ y\mu\); it drops its \(e\), and the \(y\) disappears from the environment. But we might assume these are changes explained by phonological rules.
• **Exercise:** Without relying on your knowledge of English, isolate the morphemes in the following data set.

(4) onegi_run
(5) onebo_run
(6) onegir_laughs
(7) onebo_laughs
(8) twogi_run
(9) twoboysrun
(10) twogirlslaugh
(11) twoboyslaugh
(12) onegi_runfast
(13) onebo_runfast
(14) twogirlsrunfast
(15) twoboysrunfast
(16) onegir_laughsloud
(17) onebo_laughsloud
(18) twoboyslaughl oud
(19) twoboyslaughl oud

• **Exercise:** Now let’s apply what we’ve learned to a less carefully tailored set of data.

**Example**

Following is a group of sentences in Uyghur, a Turkic language of northwestern China. I have already inserted the divisions between words. Let’s find some morphemes of Uyghur.

In doing this, we should be aware that the following sounds are, often, different phonetic expressions of the same phoneme:

[a] = [ë]  [u] = [ü]  [k] = [q] = [g].

grammatika sözərniň özgərişi wä jümlä qurumisi haqqidiki pän.
sözniň yaslişi, türlinişi tätqiq qilidigan qismi morfolojiyä deyildü.
morfolojiyä haqqidä söz açqanda sintaksistin çatnigil bolmeydu.
härqandaq tilda nur gunliñ sözler bolğandin sirt, yana sözərniň mor-
folojiyilik qurumusmi wä türliniň qa’idisi, suniqdäk sözərdin jümlä tüzüüş
qa’idisi qatarlqlar bolidü.
grammatikiliq mänä söz bilän söz otturisidiki birikiş wä türkümliniş mu-
nasiwitiñ bildiridü.

çıŋki tiliniň quruşuñ materialiniña täsək qilidigan yäkkä-yäkkä sözler
grammatikiniñ başquruşığa ötkändin keyinla andin alaqlisış rolığa igä
bolğan tilga ayländigü.