Lexical Analysis
Ling 106, Maribel Romero
Sept. 23, 2002

1 Introduction: language, grammar, lexicon.

- A language is a set of strings-finite sequences of minimal units (words, morphemes, phonemes...) with meaning. The ”machine” that generates those strings and their corresponding meanings is its grammar. A grammar must specify:

  1. a lexicon which contains every minimal unit with meaning (= every word, for the time being) and its grammatical category;
  2. a syntax, that is, a set of rules that tells you how the minimal units combine to form longer units, how this longer units combine to form yet longer units, and so forth until we form full (complex) sentences.
  3. a semantics, which determines what semantic operation or function corresponds to each syntactic rule.

- Example: (A fragment of) English.

1. Lexicon:

<table>
<thead>
<tr>
<th>Grammatical Category</th>
<th>Lexical Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proper Nouns</td>
<td>Joan, Bill, Philadelphia</td>
</tr>
<tr>
<td>Pronouns</td>
<td>he, she, it, they</td>
</tr>
<tr>
<td>Nouns</td>
<td>dog, cat, candy</td>
</tr>
<tr>
<td>Intransitive Verbs</td>
<td>walk, sleep, snore</td>
</tr>
<tr>
<td>Transitive Verbs</td>
<td>see, find, hug</td>
</tr>
<tr>
<td>Ditransitive Verbs</td>
<td>give, put, send</td>
</tr>
<tr>
<td>Propositional Verbs</td>
<td>know, claim, believe</td>
</tr>
<tr>
<td>Auxiliary Verbs</td>
<td>will, would, could, must, might</td>
</tr>
<tr>
<td>Determiners</td>
<td>the, a, some, every</td>
</tr>
<tr>
<td>Prepositions</td>
<td>with, in, on, to, before</td>
</tr>
<tr>
<td>Adjectives</td>
<td>tall, short, green</td>
</tr>
<tr>
<td>Adverbs</td>
<td>quickly, carefully, very</td>
</tr>
<tr>
<td>Complementizers</td>
<td>that, if, whether</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>and, or, but</td>
</tr>
</tbody>
</table>

2. Syntax:
If $\phi$ is a proper noun and $\psi$ is an intransitive verb, then the sequence $\phi\psi$ (disregarding inflection) is a sentence. Or, more simply: $S \rightarrow N_{pr} V_{intr}$

If $\omega$ is a proper noun, $\phi$ is an auxiliary verb and $\psi$ is an intransitive verb, then
the sequence $\omega \psi$ (disregarding inflection) is a sentence. Or, more simply: 

$$S \to N_{pr} \ AUX \ V_{intr}$$

... 

Nothing else is a sentence.

(1) a. Bill snores.
   b. Joan might snore.
   c. *Joan snore might.

Note that we are using the notion of English grammar (in particular, English syntax) in a descriptive and not in a prescriptive way. The job of a linguist is to construct a grammar that generates all and only the utterances that a given group of speakers consider well formed in their dialect. This grammar may coincide or not with prescriptive grammaticality. Also, grammaticality has to be distinguished from mere semantic anomaly (the form of the sentence is fine, though the meaning is strange) and processing difficulty.

(2) a. The hired the guy is nice.
   b. Who do you wonder whether they will come.
   c. Who do you think that saw me?
   d. Who do you think that I saw?
   e. Do you wanna see what happens?
   f. I didn’t see nobody.
   g. The professor met with Sam and I yesterday.
   h. Colorless green ideas sleep furiously.
   i. The cat the dog chased escaped.
   j. The building the guy John hired built is nice.

3. **Semantics**

If $\phi$ is a proper noun and $\psi$ is an intransitive verb, then $[\phi \psi] = \text{TRUE}$ iff ... If $\omega$ is a proper noun, $\phi$ is an auxiliary verb and $\psi$ is an intransitive verb, then $[\omega \phi \psi]$ is true iff ...

...

- This part of the course concerns the Lexicon. Consider the following cases:

(3) An infant, knowing no other language, must determine the words in its caretakers’ speech and assign them to grammatical categories.

(4) A linguist is dropped into the field and must organize a grammar of the language spoken locally. Consider two cases:
a. The linguist shares at least one language with some cooperative local.

b. The linguist has no language in common with any of the locals and must induce words, their meanings and their grammatical behavior from his observations in the field.

The following questions arise:
(i) How do we determine, out of a sequence of sounds, what sub-strings are minimal units? What criteria would help us determine that?
(ii) Once we have isolated a minimal unit, what properties or criteria determine whether this unit falls into a grammatical category or another?

2 Semantic Criteria

- Recurrence can help isolate minimal linguistic units. A word can be partially defined as a phonological form that recurs with constant meaning. We can pursue the following strategy:

  (5) Semantic recurrence:
      Find recurrent phonological sequences that always correlate with a constant component of meaning. The units posited as words should be the minimal sequences meeting this condition without leaving meaningless phonological residues.

  (6) Spanish, p. 38 in Langacker.

  (7) Luiseño, p. 39 in Langacker.

- Problem: there is no word-to-word correspondence between all languages. The meaning component that is expressed in a separate word in English may not correspond to an independent word in another language (cf. the, as, is), and vice-versa.

- Semantic cues may help us determine the category of a word. A common way to teach grammatical categories in schools relies on semantics cues like the following:

  (8) Relation between grammatical category and type of denotation:
      a. Nouns denote persons, places or things.
b. Verbs denote activities.
c. Adjectives denote qualities or properties.
d. Etc.

The idea is that the linguist would correlate a word with a denotation and try to deduce from that denotation the grammatical category of the word.

• Problems:
There are many exceptions to this rule of thumb. E.g., nouns don’t always denote persons, places or things. Imagine the following scenario:

(9) The linguist and his informant come upon a couple necking in the woods. The following exchange occurs (in the local language):
Linguist: What do you call what they’re doing?
Native Speaker: Osculation!

Here the linguist notes something that is presumably an activity and the native speaker responds with a noun (perhaps the verb is taboo!). By the semantic criteria, the linguist wrongly concludes that osculation is a verb.

Furthermore, maybe some languages lack some of our familiar grammatical categories. How do we know that a language does not distinguish the verb category from the noun category and expresses all the corresponding meanings under one single category?

3 Phonological Criteria

• Some phonological evidence can help us determine what parts out of a phonological string are words. Some examples:

(10) Final aspiration:
In Luiseño, stops (the sounds [p], [t], [k]) are usually unaspirated, but they can optionally be aspirated (as [pʰ], [tʰ], [kʰ]) when they are word-final.

(11) Consonant clusters:
Some consonantal clusters are prohibited word-internally in many languages. If we encounter one such cluster, there must be a word break inbetween.
(12) Consonant harmony:
In Chumash, sibilants must harmonize within a word: we can have several occurrences of [s] or we can have several occurrences of [ś], but we cannot have both kinds within the same word. Hence, if we encounter a sequence with both types of sibilants, there must be a word break inbetween.

(13) Vowel harmony:
In Turkish, the vowels in a word must be all front or all back.
   a. ćayličtik  b. odadabirčojukgördüm

<table>
<thead>
<tr>
<th></th>
<th>FRONT</th>
<th>BACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH</td>
<td>i</td>
<td>ü</td>
</tr>
<tr>
<td>NON-HIGH</td>
<td>e</td>
<td>ö</td>
</tr>
</tbody>
</table>

(14) Stress position:
In French, lexical stress falls on the last vowel of a word.
   a. ārišenšešegans eteližá

(15) Stress position:
In Papago, stress falls on the first vowel of a word.

QUESTION: Locate the word boundaries in the following Papago sentences. Assume that all words end in a vowel or in a single consonant. Once you have demarcated the words, associate each phonological unit to its meaning.

(16) máagina?omíkl. "The car is running"
(17) wákial?očíkpan. "The cowboy is working"
(18) wísilo?ohúhu?idgmístol "The calf is chasing the cat"
(19) mlíd?ogmáagina. "The car is running"
(20) číposid?ógwákialgwísilo. "The cowboy is branding the calf"
(21) místol?ohúhu?idgwákial. "The cat is chasing the cowboy"
(22) číposid?og?áalignístol. "The child is branding the cat"
(23) ?áali?očíkpan. "The child is working"
• Problem: Any phonological evidence useful for word demarcation presupposes a good deal of knowledge about the phonology of the language at issue. The problem is that part of this phonological knowledge can only be achieved once numerous word boundaries have been located.

• Phonological cues can help determine the grammatical category or a word.

(24) English pronunciation of *th*:
   a. The, they, this, that, then, than, although, either, whether, etc.
   b. Theater, thorn, thread, thing, method, theory, thrust, think, thorough, etc.
   c. Observation:

• Problem: this type of evidence is very limited.
4 Distributional Criteria

- Given the limitations of the semantic and phonological evidence, structural linguists were very concerned with finding a method that would break utterances into fundamental units based on distributional evidence.

The goal, then, is to use the corpus in such a way that any sequence of sounds corresponding to an utterance can be broken down into a sequence of words or morphemes, and distributional properties of these words or morphemes can be discovered and stated.

- First idea: Since every utterance contains at least one morphemic segment, find short utterances that consist of only one segmental morpheme.

Obvious problem: How do we know that an utterance is only one morpheme long, even if it is short?

- A better idea: Words (and morphemes) have dependencies with other elements in the environment in which they occur. The type of dependencies that a given word has is shared by the rest of the units in the same grammatical category or distributional class.

This means that distinct morphemes are separable from their environment by substituting them with another unit of the same grammatical category (or with nothing).

For example, we might scan our corpus of texts and discover the utterances in below. We would conclude that likes, prefers, sells and drinks are distinct words or morphemes that belong to the same grammatical category and hence can substitute for each other in the environment (26):

(25) The governor likes coke.
    The governor prefers coke.
    The governor sells coke.
    The governor drinks coke.

(26) The governor ______ coke.

Similarly in (27)-(28): yesterday, late and silently are distinct words that belong to the same class:

(27) The priest arrived yesterday with her sister.
    The priest arrived late with her sister.
    The priest arrived silently with her sister.
    The priest arrived Ø with her sister.

(28) The priest arrived ______ with her sister.
• Harris’ formulation of this idea: **HARRIS’ CONDITION I.**

Harris (p. 158) proposes that the following is a necessary condition for word or morpheme segmentation:

\[(29) \text{ If, in total environment } \_\_X, \text{ the combination } AB \text{ occurs, the combination } CD \text{ occurs and at least one of the combinations } AD \text{ or } CB \text{ occurs (where } A, B, C \text{ and } D \text{ are each phonemically identifiable portions of speech), then it is possible to recognize } A, B, C \text{ and } D \text{ as being each of them discrete morphemic segments in the environment } \_\_X.\]

By “total environment \_\_X” Harris means a sequence missing a part somewhere. For example: The governor \_\_ coke.

• Example 1.

Let us apply this first condition for word or morpheme segmentation to an example:

- Consider the following expressions:

\[(30) \quad \underbrace{\text{the junkie}}_{X} \quad \underbrace{\text{stole}}_{A} \quad \underbrace{\text{the car stereo}}_{X} \quad \underbrace{\text{\_\_}}_{B} \]

\[(31) \quad \underbrace{\text{the junkie}}_{X} \quad \underbrace{\text{trashed}}_{C} \quad \underbrace{\text{my car stereo}}_{X} \quad \underbrace{\text{\_\_}}_{D} \]

- The “total environment” in this case is:

\[(32) \quad \text{The junkie \_\_ car stereo.}\]

- The potential units we are investigating are:

\[(33) \quad A = \text{stole} \quad B = \text{the} \quad C = \text{trashed} \quad D = \text{my} \]

- We can now apply the test. If any of the following is an acceptable utterance, \textit{stole, the, trashed} and \textit{my} fulfill the first condition to be considered distinct morphemes.

\[(34) \quad \text{The junkie stole my car stereo.} \quad \text{The junkie trashed the car stereo.}\]
• Example 2.
We can apply Harris’ first condition to demarcate not just words but morphemes. Morphemes are the minimal linguistic units with meaning. Often times, words contain just one morpheme, but there are certainly many words than contain two or more morphemes:

(35) a. apple + s
    b. thought + ful
    c. un + wind
    d. un + forget + able

The reasoning we saw above can apply to the following utterances:

(36) the smurf hid my \( A \) apple \( s \) \( B \) yesterday.

(37) the smurf hid my \( A \) wallet \( \emptyset \) yesterday.

(38) The smurf hid my _____ yesterday.

(39) \( A = \) apple
    \( B = s \)
    \( C = \) wallet
    \( D = \emptyset \)

Given that the following are valid utterances, we conclude that \( \text{apple} \), \(-s\) and \( \text{wallet} \) can be considered distinct morphemes:

(40) The smurf hid my wallets yesterday.
The smurf hid my apple yesterday.

• Example 3 and problem:
Couldn’t we follow the same reasoning for \( \text{bug} \) and split it in two morphemes: \( \text{bu} + g? \)

(41) the smurf washed my \( A \) \( \text{bu} \) \( \emptyset \) \( B \) \( g \) yesterday.
(42) the smurf washed my ba \( A \) \( \text{X} \) \( B \) \( \text{X} \) \( (c)k \) yesterday.

(43) The smurf washed my ______ yesterday.

(44) \( A = bu \)
    \( B = g \)
    \( C = ba \)
    \( D = (c)k \)

(45) The smurf washed my buck yesterday.
The smurf washed my bag yesterday.

Intuitively, we don’t want to say that \(-g\) in \( bug\) is a distinct morpheme (with its own meaning!). Thus, we need to refine this distributional approach.

- **Harris’ refinement of the distributional method: HARRIS’ CONDITION II.**

  Idea: The potential units isolated by the first condition must be classifiable into a grammatical category/class in order for them to be considered morphemes. This means that the potential units would have to participate in the standard distributional environments typical of a grammatical category/class.

  (46) Accord morpheme status to sequences \( A, B, C \), if, for example, \( A, B \) and \( C \) all occur sometimes after morphemes \( D, E \) or \( F \), but never after \( G \) or \( H \), where \( D, E \) and \( F \) … constitute a distributional class against \( G, H \).

- **Applying new rule to example 2: apple + s.**

  We isolated \(-s\) as a potential unit by Harris’ condition I. Now we have to make a set \( \alpha \) with the acceptable preceding units and another set \( \beta \) with unacceptable preceding units: (48)-(50).

  If we can find lots of environments in which all the member of \( \alpha \) behave the same (as opposed to the members of \( \beta \)), then \( \alpha \) constitutes a distributional class against \( \beta \).

  If the preceding environment set \( \alpha \) for \(-s\) constitutes a distributional class, then \(-s\) is a morpheme.

(47) The smurf hid my apples yesterday.
The smurf hid my wallets yesterday.
The smurf hid my crayons yesterday.
The smurf hid my pencils yesterday.
The smurf hid my nuts yesterday.
The smurf hid my squirrels yesterday.
Etc.
(48) Set $\alpha = \{ \text{apple, wallet, crayon, pencil, nut, squirrel, ... } \}$

(49) *The smurf hid my very+s yesterday.  
*The smurf hid my late+s yesterday.  
* The smurf hid my for+s yesterday.  
* The smurf hid my think+s yesterday.  
Etc.

(50) Set $\beta = \{ \text{very, late, for, think, ... } \}$

(51) Behavior of elements of $\alpha$ in the following environments:
   a. The good ______ fell.
   b. The ______ hit the ground.
   c. I'll take the ______ now.
   d. The ______ is good.
   e. * The ______+s is good.
   f. * The ______ are good.
   g. The ______+s are good.

• Applying new rule to example 3: bu + g.  
(52) I love my bug.  
    I love my rug.  
    I love my mug.  
    I love my slug.  
    I love my jug.  
    I love my bag.  
    I love my rag.  
    I love my hag.  
    I love my nag.  
    I love my leg.  
    Etc.

(53) Set $\alpha = \{ \text{bu, ru, mu, slu, ju, ba, ra, ha, na, le, ... } \}$

We cannot find a good number of environments in which all the elements in $\alpha$ would behave in the same way.  
This means the set $\alpha$ does not constitute a grammatical category/class.  By Harris’ condition II, this means that -g is not a distinct morpheme.

• QUESTION: What about the sound [s] in box? Can we consider that it constitutes a distinct morpheme? Justify your answer with the above procedure.

• Notice, however, the hedge in Harris’s procedure. We have to consider a lot of distributions before we can make secure judgements about morpheme demarcation!!!
4.1 A Difficult Case

Consider words like:

(54)  con-ceive, de-ceive, per-ceive, re-ceive  
con-sist, de-sist, per-sist, re-sist  
con-cur, re-cur

These words involve *bound morphemes*. These are units that are involved in forming words, but which never occur independently as words.

(55)  con-, de-, per-, re-  
-receive, -sist, -cur

Does Harris’s method identify the above as morphemes?

4.2 Special cases

The units that the method identifies don’t have to be contiguous; that is, they don’t have to form a continuous string of elements.

1. Consonantal roots in Semitic:

(56)  *kataba* ‘He wrote.’  
*katabtu* ‘I wrote.’

(57)  *kataba* ‘He corresponded.’ (‘:’ indicates vowel length.)  
*katabtu* ‘I corresponded.’

The above give the following units:

(58)  k__t__b ”write”  
_____a___a Perfective (completed in the past)  
____a ”he”  
_____tu ”I”  
vowel length: Reciprocal

2. Repeated elements are identified as units:

(59)  *filius bonus* — ‘good son’  
*filia bona* — ‘good daughter’

3. Other noncontiguous agreement phenomena:

(60)  *le bon-Ø fils*  
*la bonne fille*
4. Phoneme replacement:

(61) take — took
give — gave
sing — sang
hang — hung

5. Intonation contours and other suprasegmentals:

(62) You ate the dog.
You ate the dog?
You ate the dog!?