

The Lexicon

LING 106

1. DETERMINING THE LEXICON

Here's the question we'll be considering for the next few classes: suppose you're given a set of strings. How do you determine (a) what the words are, (b) what syntactic category the words are, and (c) what the words mean?

In particular:

- Suppose you're an infant being given a bath, and the parent who's bathing you calls out to your other parent, "**I'm washing the baby**", or more properly, "**aimwašɪŋðəbeibi**". (Remember that when you're dealing with spoken language rather than written language, you don't get all those convenient word breaks.)
- Suppose you're a linguist in Senegal, and your bilingual friend is washing her baby. You ask her, "What do you call what you're doing in Pulaar?" and she answers, "**lo:tugol**" (the colon marks a "long vowel").
- Suppose you're a linguist in Senegal, and you don't know any English speakers, and you see someone washing a baby, who says "**kobo:boonmiwonielo:tude**".

Something in each of these strings may refer to the washing, but which? For that matter, in the first and third strings, what pieces are even candidates for having meaning? Is **wašɪŋ** a lexical unit with meaning? Is **waš**? Is **əbei**? What about **bo:boon**? **de**? **nielo**? (Good luck with that.)

2. USE SEMANTIC INFORMATION

Semantic knowledge can help with both (a) and (b) above...

2.1. *Identifying a word*

- If a phonological string recurs at the same time that a repeated meaning recurs, the two may correspond—look for a *minimal* sequence *without leaving meaningless phonological residues*. (Warning: sometimes phonology does in fact insert meaningless phonological residues in certain contexts.)

Example from Luiceño (Uto-Aztec, Southern California):

noki	'my house'	[missing]	'my son'
?oki	'your house'	?opeew	'your wife'
poki	'his house'	popeew	'his wife'
pomki	'their house'	pompeewum	'their wives'

Conclusion: we can isolate corresponding meanings and Luiseño data.

Problems: This is only as good as the actual correspondence between form and meaning. Even when you've got word breaks to look at:

- A single meaning may be expressed with different forms. Suppose you're an English speaker trying to determine the words of French.

le garçon a donné les livres à la fille

Meaning: "The boy gave the books to the girl."

"The" is a repeated unit of meaning, but there's no repeated word in the sentence. How can we tell what corresponds to it? (Answer: no one word does; *le*, *la*, and *les* all do.)

- Different meanings may be expressed with a single form. Suppose you're a French speaker trying to determine the words of English.

I know that John and Mary know Peter.

Meaning: "je sais que Jean et Marie connaissent Pierre"

"**know**" is a repeated word in the sentence, but there's no repeated meaning. How can we tell what it corresponds to? (Answer: it corresponds to both "have information that" and "be acquainted with".)

2.2. *Identifying the syntactic category of a word*

- If it's a person, place, or thing, it's a noun.
- If it's an action, it's a verb.
- If it's a quality of some sort, it's an adjective.

Thus, if someone points at a baby and says **bo:bo**, it might mean "baby" or "person" or whatever, but you can tentatively postulate (using the method above) that it may be a separate word; and if so, you can be fairly sure that what you've got is a noun.

Problems: this is only as good as the correspondence between meaning and category.

- Not all nouns are concrete objects. Not all verbs are actions—is *know* an action? Is *be*?
- When your friend tells you that her action is called **lo:tugol**, how sure are you that that's the verb "to wash" and not the noun "ablution"?
- Consider the two English sentences:

I have studied chemistry.

I am a former chemistry student.

Problem #1: these express (roughly) the same thing, but the bit of meaning about how I did all that fancy book-larnin' is a verb in the first sentence and a noun in the second.

Problem #2: we seem to have the lexical entries:

<**have**, auxiliary verb, “this was at some point in the past”>

<**former**, adjective, “this was at some point in the past”>

That is, “have” indicates that the studying was in the past (in contrast with **I am studying chemistry**), and “former” indicates that the being a student was in the past (in contrast with **I am a chemistry student**). But their syntactic categories are different.

- Conversely, we can't be sure that the language we're studying doesn't express both “noun” and “verb” meanings with a single grammatical category.

3. USE PHONOLOGICAL INFORMATION

Phonological knowledge can help with both (a) and (b) above...

3.1. *Identifying a word*

Phonology is often sensitive to word boundaries. If a certain sound or sequence of sounds, or a certain phonological process, can only occur at the start of a word, or at the end of a word, or in the middle of a word, or within a word, then locating these can help us place word boundaries. For example:

- *Final aspiration*

In Luiseño, the sounds [p], [t], [k] are usually unaspirated, but they can optionally be aspirated—pronounced with a puff of air, [p^h], [t^h], [k^h—when word-final.

- *Consonant clusters*

Many languages prohibit consonantal clusters word-internally.

- *Consonant and vowel harmony*

Many languages require “harmony” within words—all instances of a certain class of sounds must be identical along a certain dimension. In Chumash, this is true of sibilants: though either [s] or [š] (= “sh”) can appear repeatedly within a word, a word cannot contain both: all sibilants change to match the last one. For instance:

/s + api + tš ^h o + it/ =	šapitšholit	‘I have a stroke of good luck’
/s + api + tš ^h o + us/ =	sapitsholus	‘He has a stroke of good luck’
/s + api + tš ^h o + us + waš/ =	šapitšholušwaš	‘He has had a stroke of good luck’

Thus, a sequence of sounds that has both [s] and [š] must consist of multiple words.

In Turkish, this is true of vowels:

- /ip/ + [genitive] = **ipin** 'rope'
- /kiz/ + [genitive] = **kizİN** 'girl'
- /yüz/ + [genitive] = **yüzün** 'face'
- /pul/ + [genitive] = **pulun** 'stamp'

(Note: these forms are accurate, but there's also a lot of other things going on.) Thus, a sequence with both [u] and [i] must consist of multiple words.

- *Stress*

In French, lexical stress falls on the last vowel of the word. In Papago, stress falls on the first vowel of the word. (Note that this is true of a single word; putting words into phrases may affect where the stress falls.)

Problem: this kind of evidence is great if you already know the facts, but not so great otherwise. For instance, you can determine whether stress is always word-initial or word-final by isolating a lot of words and checking their stress—but you're not going to be able to use stress to determine word boundaries along the way.

3.2. *Identifying the syntactic category of a word*

- *th* in English:
 - the, they, this, that, then, though, either, whether...
 - theater, thorn, thread, thing, theory, thrust, think, thorough....
- Stress in (certain Latin-derived) English words:
 - reCORD, conVERSE, comPACT, conVICT, conVERT, obJECT
 - REcOrd, CONverse, COMpact, CONvict, CONvert, OBject

Problems: this kind of evidence is rare. It's also often imperfect (cf non-Latinate words like *shuffle*, some Latinate words like *request*).

4. USE DISTRIBUTIONAL INFORMATION

Given that semantic and phonological information is limited in usefulness, how *can* we go about deciphering an unknown language?

- Idea #1: since every utterance contains at least one morpheme, find single-morpheme utterances.

Problem, perhaps obvious: how do you know that an utterance is a single morpheme?

lo:tugol	lo:tu + gol	lo: + tu + gol
wash(v.)	wash(v.) + INFINITIVE	clean + become + cause

One of these might be right...but which one?

Goal: given a set of sentences, break down each one into a sequence of words/morphemes whose distributions can be determined.

Premise: if the syntactic rules of a language work with categories, not just lexical items, then two lexical items with the same distribution should have the same category.

- Idea #2: Find items with the same distribution: that is, find contexts in which two distinct strings can be substituted for one another. These strings are then pretty good candidates for morphemes with the same category.

Example: suppose we collect data (spoken, written, etc.) and we find the following utterances.

- (1) **The governor likes coffee.**
The governor prefers coffee.
The governor buys coffee.
The governor drinks coffee.

Conclusion: because **likes, prefers, buys, drinks** can substitute for each other in the environment

- (2) **The governor _____ coffee.**

they all share a common grammatical category. (We might in fact wonder whether **The governor _____s coffee** is an environment into which we can substitute **like, prefer...**)

We can do the same, to a more limited extent, with contexts within sentences as opposed to entire sentences. From some actual corpora (no bonus points for guessing which ones):

- Yet **across the gulf of space**, minds that are to our minds as...
- It may be that **across the immensity of space** the Martians have watched...
- ...I made out a string of black figures hurrying **across the meadows from the direction of** Weybridge.
- ...far away **across the meadows in the direction of** Kew Lodge
- ...I knew that there **I had the poorest chance of finding** my wife.
- ...There, it seemed to me, **I had the best chance of learning** what...

Conclusions: **gulf/immensity, from/in, poorest/best** share categories.

- Ms. Lewinsky testified that she met with the President privately on ten occasions **after she left her job at the White House**.
- Ms. Lewinsky testified that she met with the President in private **after she left her position at the White House** on eleven dates in 1997.
- The President stated in his civil deposition that **he could not recall whether he had** ever given any gifts to Ms. Lewinsky;
- that **he could not remember whether he had** given her a hat pin although “certainly, I could have”...
- Seventh, the President **refused to answer specific questions before the grand jury** about what activity he did engage in (as opposed to what activity he did not engage in)...
- The President refused six invitations to testify to the grand jury, thereby delaying expeditious resolution of this matter, and then **refused to answer relevant questions before the grand jury** when he testified in August 1998.

Conclusions: **job/position, recall/remember, specific/relevant** share categories.

- PETRUCHIO. **I say it is the moon that shines so bright.**
- KATHERINA. **I know it is the sun that shines so bright.**
- **Where should he find it fairer than in Blanch?**
- **Where should he find it purer than in Blanch?**¹
- **This is the strangest tale that ever I heard.**²
- **This is the silliest stuff that ever I heard.**³

Conclusions **say/know, moon/sun, fairer/purer, strangest tale/silliest stuff** etc.

¹ Both from the Citizen’s speech in *King John*, II.i

² Prince John of Lancaster, in *Henry IV Part I*, V.iv

³ Hippolyta, in *A Midsummer Night’s Dream*, V.i

4.1. *Breaking it down further*

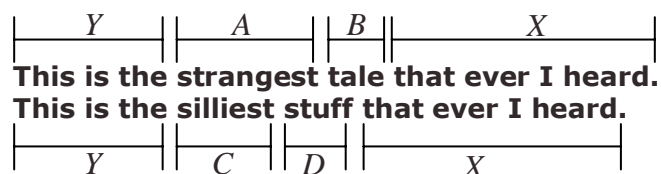
In these examples, we decided that **moon** and **sun** shared a single category; we also decided this about **fairer** and **purser**, and about **strangest tale** and **silliest stuff**. But as we know, only one of these consists of a monomorphemic pair. How do we know whether we can segment things further?

Harris's formulation of this idea: **Harris's Condition I**, a necessary condition for word or morpheme segmentation:

If, in total environment Y_X , the combination AB occurs, the combination CD occurs, and the combinations AD and CB occur (where A , B , C and D are each phonemically identifiable portions of speech), then it is possible to recognize A , B , C and D as being each of them discrete morphemic segments in the environment Y_X .

Example #1:

- Consider the last pair cited above:



The “total environment” is **this is the _____ that ever I heard**; and we have

A = **strangest**,
 B = **tale**,
 C = **silliest**,
 D = **stuff**

We know that AB and CD occur in the environment. We need to find out whether AD and CB also occur in that environment...

This is the strangest stuff that ever I heard.
This is the silliest tale that ever I heard.

We can do this by finding it in the corpus (no luck in this case), or by asking a speaker of the language. Because in fact both of these sentences are judged acceptable, we conclude that A , B , C , and D are all separate morphemes.

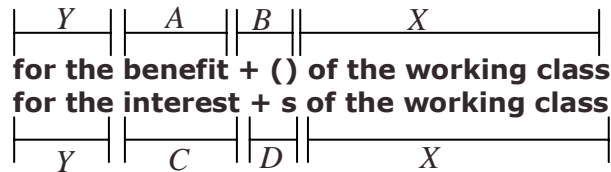
Example #2:

- Consider the following pair from yet another corpus:

Free trade: **for the benefit of the working class.**

...they are conscious of caring chiefly **for the interests of the working class...**

In this case we have $Y_X = \text{for the } ______ \text{ of the working class.}$ What are our A, B, C, D ?



That is: (a) we can look below the level of the word to find morphemes, and (b) the lack of a word or morpheme may also be informative.

In this case, we'd want to know whether AD and BC can appear in this environment:

for the benefits of the working class
for the interest of the working class

and while once again they don't appear in our corpus, we can check with a native speaker and learn that they can.

Example #3:

- Recall that we had **fairer/purer** and **moon/sun** as distinct units. But we can break both examples down a little farther:

Shared context: **Where should he find it ___er than in Blanch?**
*Morphemes:*⁴ **fair, pure**

Shared context: **It is the ___n that shines so bright.**
Morphemes: **moo, su**

...uhoh.

⁴ Ignore the spelling difference; imagine that we're working with the pronunciations.

- To take an example not from a corpus:

**Sam showed me her bug yesterday.
Sam showed me her back yesterday.**

Taking $A = \mathbf{bu}$, $B = \mathbf{g}$, $C = \mathbf{ba}$, $D = \mathbf{ck}$, we can check AD and BC :

**Sam showed me her buck yesterday.
Sam showed me her bag yesterday.**

So clearly these are all morphemes! (Ugh.)

4.2. *Solving the problem in Example #3*

Harris's Condition I was a necessary condition for identifying a morpheme, but clearly not sufficient. Thus, a refinement of the distributional method: **Harris's Condition II**.

Accord morpheme status to sequences A , B , C , if, for example, A , B and C all occur sometimes with morphemes D , E or F , but never with P or Q , where D , E and F ...constitute a distributional class against P , Q .

That is: in order for things to be morphemes, they should follow a more general distribution than simply occurring in the AB , CD , AD , BC pattern. The methodology is:

- Suppose that x is a sequence of sounds that looks like it may be a morpheme. Then:
 - Find a distributional class K such that x can appear after any element in K ;
 - Find a distributional class P such that x cannot appear after any element in P .
- Then you have additional evidence that x is a morpheme.

Why is this evidence? Because this makes it possible to explain the distribution of x . Suppose that we assign x to the syntactic category N . Then the fact that it follows K s but not P s can be explained by theorizing that somewhere in the syntax of the language is a rule like

If S is a string consisting of an A , a B , ..., a K , an N , ..., then S is a sentence.

but nowhere is there a rule like

If S is a string consisting of an A , a B , ..., an P , an N , ..., then S is a sentence.

Back to Example #2:

- The **benefit, interest, -s** distribution suggested that **-s** was a morpheme. But we want more evidence than just the fact that it follows two words; we'd like to see that it follows anything from a large set of words of a single distributional category, and not words that are not in that category.

For instance, let's suppose that **benefit** and **interest** are in a distributional class that we'll call *Set N*—we've already established that the elements in *N* have the same distribution. In addition to **benefit** and **interest**, *N* includes...

highway, vassal, hour, leader, proposal, burgher, cobweb...

Now, we can find places in the corpus where our potential morpheme **-s** follows each of the elements in *N* (*the robe of speculative cobwebs; the burghers of the Middle Ages, with their miserable highways; ...*). In contrast, consider another distributional class:

$A = \{\mathbf{miserable, speculative, critical, immediate, indistinct...}\}$

None of the elements of *A* are ever followed by **-s** in the corpus.

Therefore, we have something that can follow any element of *N* but no element of *A*. This suggests that it is in fact a morpheme.

Back to Example #3:

- It appeared that **-n** might be a morpheme, as it could follow both **moo** and **su** in the same context. And we can find a large number of other things it can follow in that context, as well as a large number of things it can't:

<i>Shared context:</i>	It is the ___n that shines so bright.
<i>Class K:</i>	moo, su, coi, crow, taver, ...
<i>Class P:</i>	po, sa, flor, ...

But are {**moo, su, coi, crow, taver, ...**} really a distributional class? Not especially, no: there's no large number of contexts in which they all appear.

- Similarly, while we can find other things that can precede both **-g** and **-ck** (e.g., **ha, jo, bri**), we're not likely to find other contexts in which they all behave the same way—they're not a distributional class. Consequently, these are not morphemes.

4.3. *The (inevitable) problems*

Harris's method is fairly solid, but there are certain challenges involved.

- You need to have a pretty large corpus to make it work. To be honest, the corpus I used for the **-s** example won't really tell you that *N* is a distributional class, because many of the words don't appear independently, and those that do don't happen to appear in the same environments.

Conversely, without a large enough corpus, you'll get a number of "false positives":

When I was *alone* with Ms. Lewinsky on certain occasions...

I remember once in particular **when I was *talking* with Ms. Lewinsky**...

Does this mean that **alone** and **talking** are the same category?

- *Bound morphemes:*

con-ceive, de-ceive, per-ceive, re-ceive

con-sist, de-sist, per-sist, re-sist

con-cur, re-cur

Should we consider the various pieces here to be morphemes? Does Harris's method identify them as morphemes?

- *Semantic issues:* while both **happy** and **tired** can appear in the context **I have been ___ all day**, many other morphemes that we'd want to consider part of the same syntactic category cannot—for instance, **American, metallic, crumbly**.

We'll end up identifying things that we know are all adjectives as being part of different distributional classes. But this brings us back to the "colorless green ideas" question: do we want to ensure that certain sentences aren't generated by the grammar, and if so, should that be in the syntax or semantics?

- *Irregular forms:* our corpora may contain places where we can swap **benefit** and **interest**, but note that **woman** and **women** won't, indeed can't, appear before **-s**. Does that make them part of a different distributional class?
- *Regularities other than suffixation/prefixation:* Many languages derive words with things other than prefixes and suffixes. For instance:
 - *Phoneme replacement.* Suppose that a language replaces **oo** with **ee** to form a plural (e.g., **foot/feet, goose/geese, tooth/teeth**—this was a fairly regular process at one point in the history of English). What will Harris's method do when it encounters these forms?

- *Noncontiguous phenomena.*

Consonantal roots, e.g. Semitic languages (in this case, Arabic):

katabat	“she wrote”
kutiba	“it was written”
yaktubu	“he writes”
kitaab	“book”

Agreement, e.g. Latin:

filius bonus	“good son”
filia bona	“good daughter”