Corpus Linguistics

2023-11-03

Levels of Linguistic Representation

Main topics

- Variants
- Internal vs. external conditioning
- Lexical semantics
- Synonyms and meaning clusters
- Phonetics
- Morphology
- Syntax
- Discourse

Variants

- There are different ways of saying things, or expressing certain meanings

- e.g.
  - *the boy's hat vs. the hat of the boy*
  - *The boy was hit by the car* vs. *the car hit the boy*
  - *I'll go* vs. *I will go*
  - etc.

Variants

- When we talk about variants, we can talk about the variation occurring at a specific level of linguistic structure
- The alternation between Saxon genitive *‘s* and *of* is variation at the level of syntax and partially morphology
- *sneaked~snuck*, *dived~dove, throve~thrived* etc... these variants would be morphological
- *militaristic* -> mɪlətɹɪstɪk vs. mɪlərɪɹɪstɪk

Variants

- Variants can be more or less probable contingent on different types of context
- The textbook puts these in two overarching categories
- **external context**: who is talking, their status, conversation
- **internal context:** depends on other words around it

Variants

- Saxon genitive vs. *of* possession

- What do you think the variation is based on? (external vs. internal)

Variants

- Saxon genitive vs. *of* possession
- Saxon genitive more likely to be used in speaking rather than writing
  - (external)
- Saxon genitive less likely to be used before a sibilant sound (*fish, jazz)*
  - (internal)

Variants

- **Variationist corpus linguistics:** understanding what conditions the variants we find in text.

- take away – linguistic behavior is constrained beyond grammar rules, in a probabistic fashion by internal and external context

Structural levels of variation

- Lexical semantics
- Phonetics
- Phonology
- Morphology
- Syntax
- Discourse

Lexical semantics

- Lexical semantics – the meaning of specific *lexical items*
- Lexical items are memorized forms like dictionary elements (remember the distinction between word-form and lexeme)
- A lexical item has a **behavioral profile** or **configarational profile** that reflects the different meanings it can take on in different contexts

Lexical semantics

- A lexical item has a **behavioral profile** or **configarational profile** that reflects the different meanings it can take on in different contexts

- I’ve also heard the term **potential** used to talk about all the variants

- Variation at the level of **lexical semantics** will be about *meaning variants*

The meanings of *run*

- It can be intransitive or transitive
- fast pedestrian motion
- uncontrolled movement
- escaping from something unpleasant
- toward something more pleasant
- abstract motion
- functioning
- causing to function

The meanings of *run*

- Intransitive vs. transitive

*Simons had **run** down to the villa to get help (intranstive)*

*his brother **ran** a mile (transitive)*

The meanings of *run*

- fast pedestrian motion

*Simons had **run** down to the villa to get help (intranstive)*

*his brother **ran** a mile (transitive)*

The meanings of *run*

- uncontrolled movement

Dogs ran about

He was running his mouth

The meanings of *run*

- escaping from something unpleasant

When he loses his temper with her she runs off

towards something pleasant

She ran away with him

The meanings of *run*

- Gries finds 56 senses of *run* and then uses behavior profiles to answer questions about these forms

- Can we predict the specific sense being used based on the surrounding context?

Emotions and their senses

- Krawczak (2014) analyzed the meanings of *ashamed* and words that could be considered synonyms like *humilated* ...

- How?

- By coding properties such as the source of the feeling, the time span of the feeling, etc.

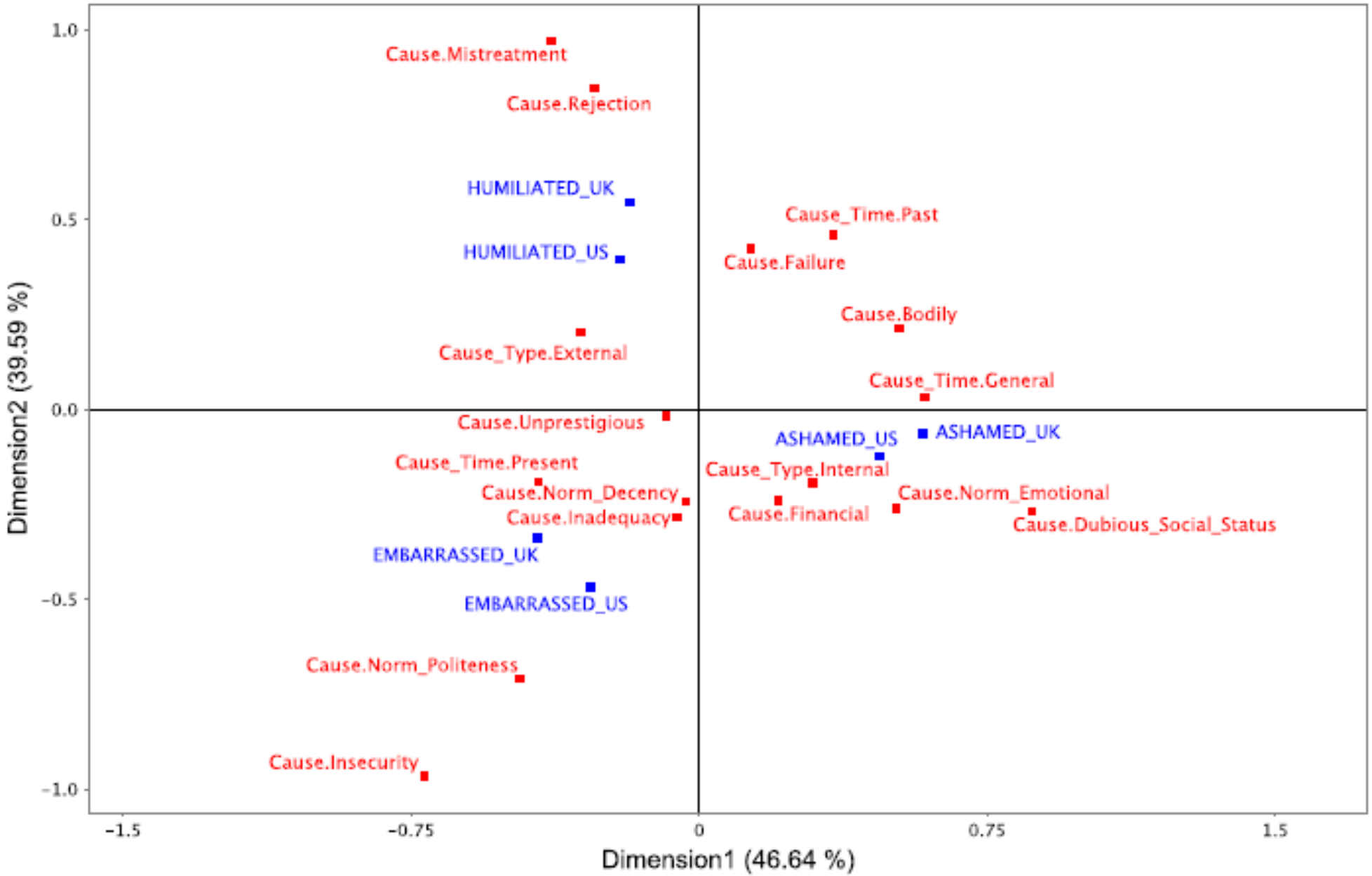


Figure 1. Binary Correspondence Analysis

Emotions and their senses

embarrased are for situations that are less serious and more fleeting

humiliated is more grevious and is used for externally cuased emotions

Phonology and phonetics

- Phonetics is the physical manifestation of speech
- Phonology is about the contrastive sounds
- Let‘s do a recap of the core concepts

Contrastive *vs.* Non-contrastive

- **Contrastive** sound: For a given language two or more sounds are contrastive if the distinction between them can be used to build distinct words that speakers do not consider homophones.
- Two sounds can be used to make a meaningful distinction in the language.

- **Non-contrastive** sound: For a given language two sounds are *non-contrastive* if their distinction cannot be used to make a meaningful distinction between words.

Contrastive *vs.* Non-contrastive

- Two sounds might be contrastive in one language and non-contrastive in another.
- Recall in our description of English we can make a distinction between aspirated and non-aspirated sounds.

  [tʰɑp] “top”

  [stɑp] “stop”

Contrastive *vs.* Non-contrastive

- But English speakers cannot exploit that the difference between these sounds in the same way that they can for the distinction between [d] and [t].
- /tɛl/ *vs.* /dɛl/
- /daɪ/ *vs.* /taɪ/
- /tu/ *vs.* /du/
- So in English [d] *vs.* [t] is **contrastive**
- [t] *vs.* [t^h] is **non-contrastive**

Contrastive *vs.* non-contrastive

- Actually we can be more general.

- English exploits a voiced:voiceless contrast but not a aspirated:non-aspirated is non-contrastive.

- Notice that saying that a feature is non-contrastive for a language is very different from saying that the language just doesn’t have that sound; English speakers systematically produce aspirated stops.

Contrastive *vs.* non-contrastive

- How we know sounds are **contrastive**;
- Try thinking of words that contrast only /k/ and /g/ in English?

Contrastive *vs.* non-contrastive

- How we know sounds are **contrastive**;
- Try thinking of words that contrast only /k/ and /g/ in English?
- Easy… “gill” *vs.* “kill”; “come” *vs.* “gum” etc…

Phonology vs phonetics

- So phonological corpus linguistics in principle should be about alternations that involve changing/deleting/adding phonemes

- Phonetic corpus linguistics should be about alternations that affect pronunciation variants.

- But the problem is that when you change one phoneme to another you are also changing the phonetics.

Phonology vs. phonetics

- For corpus studies, the difference in practice relates to measurement (for the most part)

- You are more likely doing phonological corpus linguistics if you are looking at categorical change and you are looking at alternations that can be found directly in a alphabetically written text

- You are more likely doing phonetic corpus linguistics, if you are looking at measurements.

Orthography and phonemic represetation

‘For some languages, the written words are similar enouh to their phonemic representations that text corpora can be used to examine questions relating to phonology.’ (Barth & Schnell 2022:52)

Phonology questions that could involve corpora

- What clusters of consonants are more common in natural speech?
- What phonemes are more common in natural speech?
- What phonemes rarely occur adjacent to one another inside words?

etc.

Corpus linguistics for phonetics

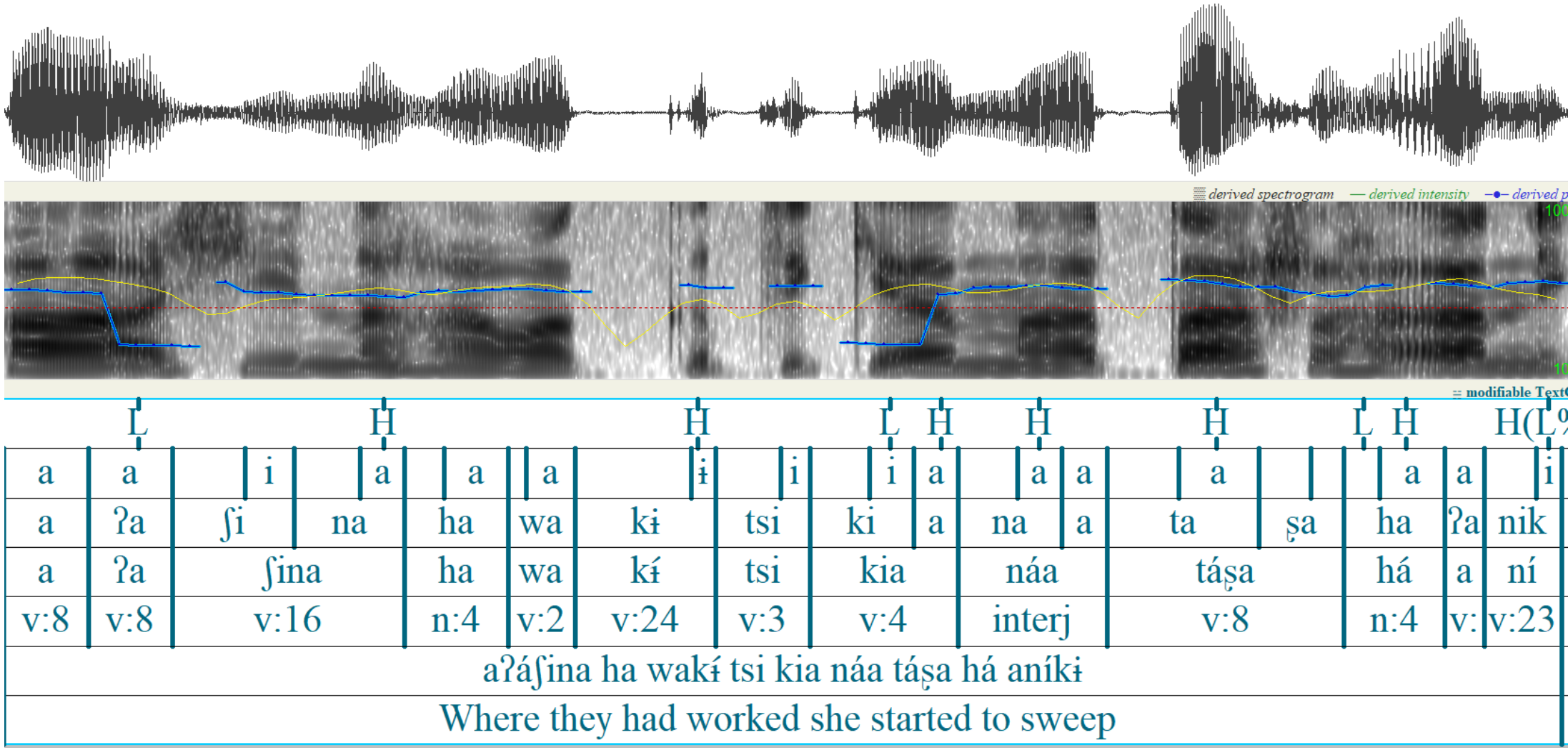
- A corpus study using phonetics would involve making phonetic measurements.

- While in principle this could be done with articulation, in practice it is always done with acoustic measurements (sound wave).

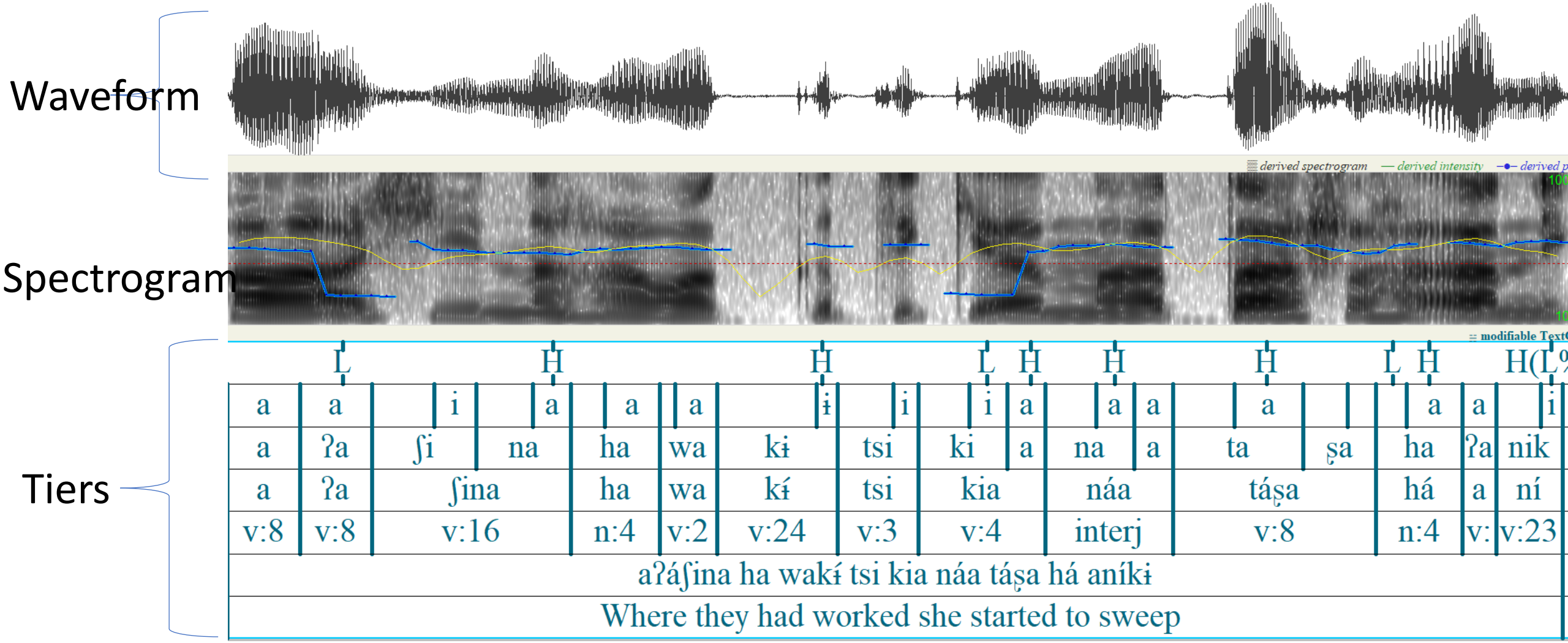
- The most common acoustic analysis program is Praat.

Praat

- website: https://www.fon.hum.uva.nl/praat/download_win.html



This is a Praat annotation of a sentence in Chácobo, a southern Pano language of the northern Bolivian Amazon.



This is a Praat annotation of a sentence in Chácobo, a southern Pano language of the northern Bolivian Amazon.

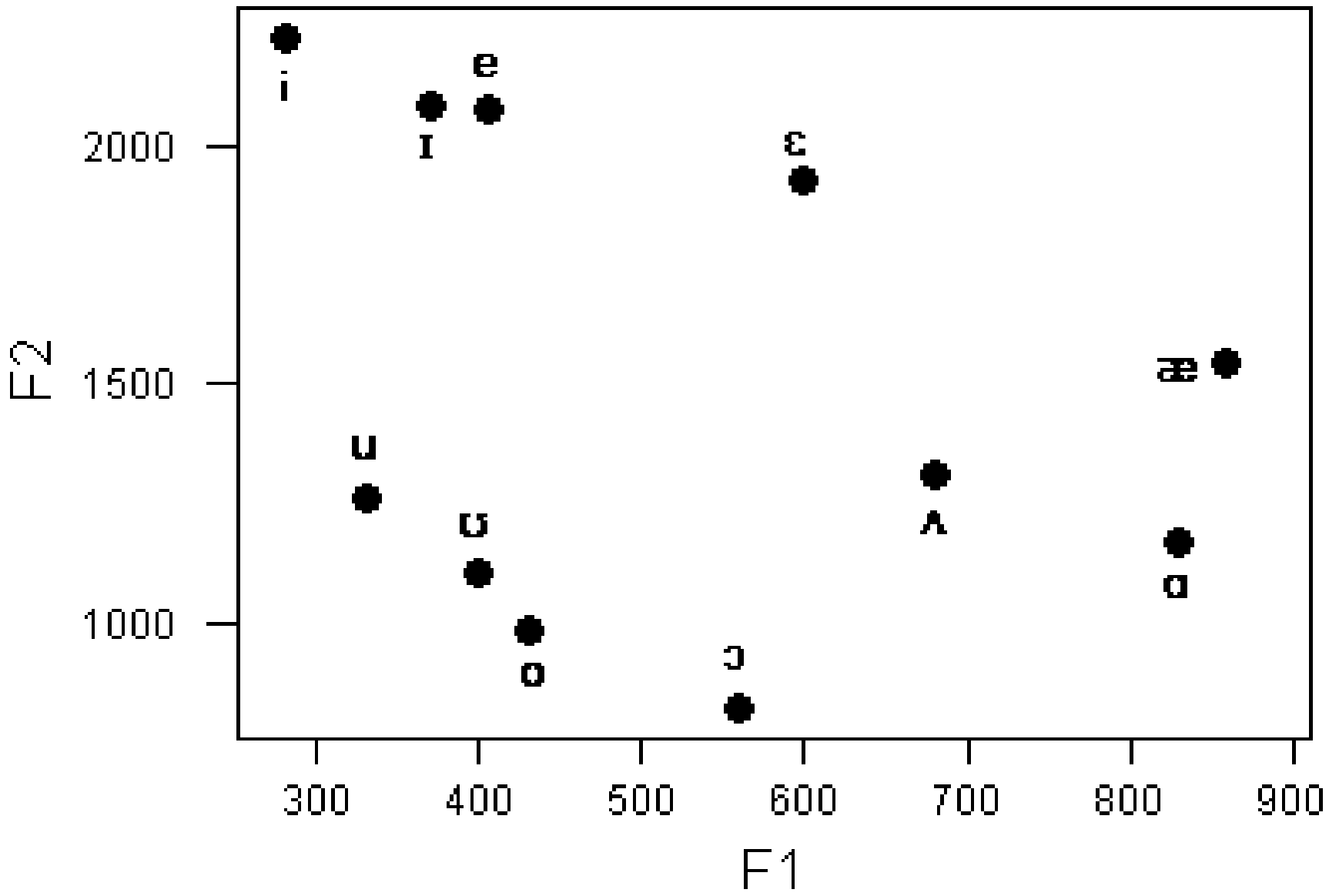
Phonetic measurements

- Phones have different phonetic measurements associated with them.

- A consonant has
  - Duration (milliseconds)
  - Voice onset and offset time (milliseconds)
  - Voicing (milliseconds or amplitude
- Vowel
  - Duration (milliseconds)
  - Pitch (F0)
  - Intensity (amplitude)

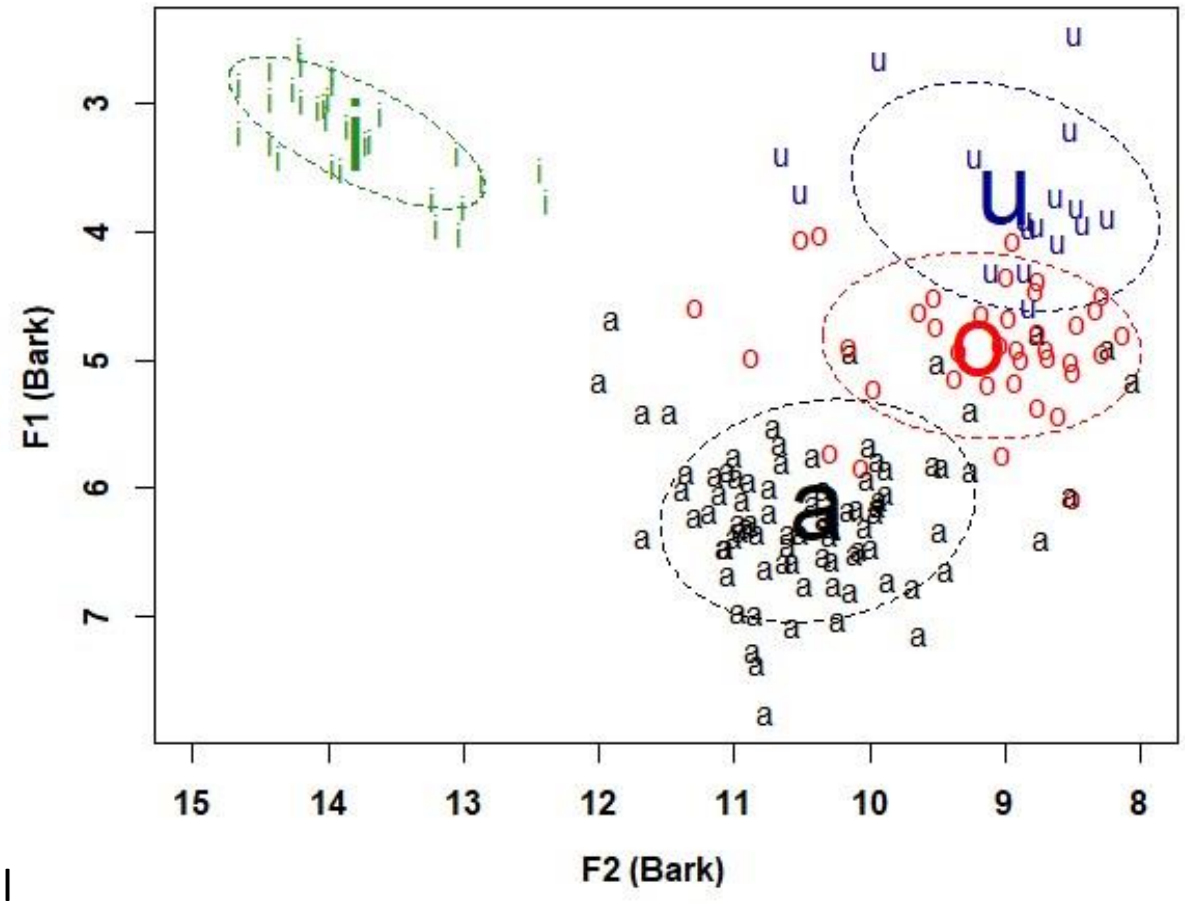
Phonetic measurements

- F1 and F2
- Vowels have an F1 and an F2
- These are the main measurements of vowel quality.
- F1: Vowel height (low F1 means high)
- F2: Vowel frontness (high F2 means front)


i
e
ɪ
ɛ
æ
u
ʊ
o
ɔ
ʌ
ɑ
2000
1500
1000
F2
300
400
500
600
700
800
900
F1

Phonetic measurements

- Actual vowel tokens are much more variable



https://fass.ubd.edu.bn/research/dusun/vowels3.html

Phonetic measurements

- How does one develop a phonetic annotation? (e.g. know the boundaries between different sounds)
- There are acoustic ‘landmarks’ that can be used to segment speech into phones.
- But
- Developing a phonetic corpus by hand is very time consuming – which is why people use forced alignment systems.

Morphology

- Morphology is concerned with word internal structure (generally)
- A central concept is the **morpheme**
- Words are made up of morphemes – morphemes are the smallest meaningful parts of words.

complete-ness has two morphemes

Morphology

- Some other core concepts:

- **Bound:** Some morphemes are ‘free’ (they can occur on their own) and others are bound (they depend on the appearance of something else to occur)

- **Roots**: The base of th Word, e.g. (*complete* in *completeness*)

- **Affixes:** The bound parts that are modifying the base (*-ness* in *completeness*

Moprhological producitivity

- Morphemes vary in terms of their productivity.
- Some morphemes like –*ness* are highly productive
- Others like –*th* in *warmth* only occur with a small number of forms.