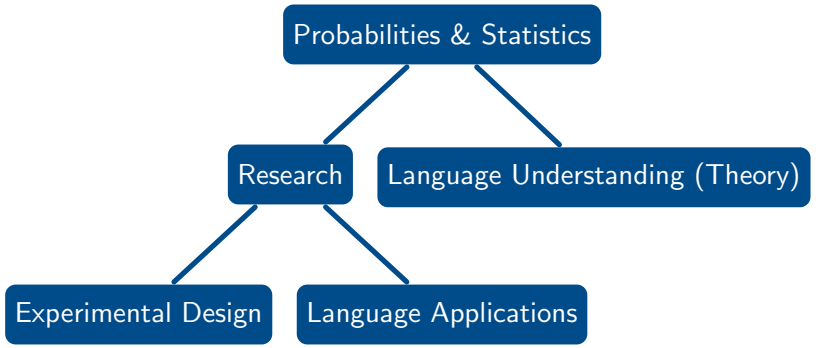
Statistics and Language Technology

Introduction

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability



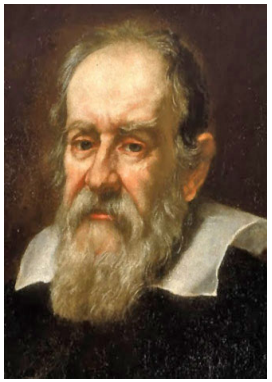
Introduction

- As I speak, you are taking part in an unconscious game. You make **hypotheses** and try to **predict** what I am going to say, you might have guessed the following words, you also continuously restructure the utterances in your minds as I utter new ones. This can be easy in cases such, **went to the post-X**. (office).
- We do predictions all the time and this is not based on some kind of ocular knowledge but based on something real and really remarkable and astonishing: our innate ability to do linguistic predictions using the knowledge we have about the world.
- We will see that this basic idea is fundamental for understanding language structure, modeling language, finding information in texts and translating them. These methods and applications are some of the topics of this class.

Section 1

Probabilities as a window to language understanding

Probabilities as a window to language understanding



"La filosofia scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si pu intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali scritto. Egli scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezzi impossibile a intenderne umanamente parola; senza questi un aggirarsi vanamente per un oscuro laberinto." (Galileo Galilei (1564–1642), Il Saggiatore, Cap. VI)

Language Acquisition

1. Infants show sensitivity in their environments.
2. Do infants use statistics to acquire their native language?
3. Recent findings show that are sensitive to statistical patterns in their environment:
   - **finding basic regularities:** identifying frequencies: sound sequences, phonotactic patterns, stress patterns, prosodic patterns and how they related to external reality, objects, people, etc.
   - **identifying more complex relationships:** words, categories, such as grammatical parts of speech, basic syntactic structures, and meanings.
   - **finding conditional probabilities:** in word segmentation, in part of speech identification, etc., the conditional probability of Y given X in a string XY.
   - **identify linguistic structure:** phonetics, phonology, syntax,
   - employ the knowledge to different environments.

Language Acquisition and Statistics

What do we mean when we say that children employ statistics when they learn a new language?

- Finding patterns in data provided by their immediate environment.
- The learning can be supervised (in children and algorithms!):
  - reinforcement: Bravo! You said "mama", have a cookie!!!
  - punishment: Punishment (it comes in many forms!): don't say *mouses* again say mice!
- Unsupervised: the child identifies regularities in the environment: dog dogs, horse horses, so mouse $\rightarrow$ *mouses is an error that can occur from pattern observation.
- More complex learning: Bayesian models

Language Acquisition and Statistics

Probably, most linguist agree that some sort of statistical learning takes place in language acquisition. How do children apply this knowledge?

- Children use statistics to identify syllables, such as CVCVCV sequences of a consonant (C) and a vowel (V) in the speech of their parents.

Language Acquisition and Statistics

- The sensitivity to cues carried by duration and intonation emerge within the early months of life. For instance, Langus, (2010) reports that
  - **1–2 months**: infants use prosodic cues to segment speech: Infants can discriminate pitch change.
  - **4.5 months**: infants prefer passages with artificial pauses inserted at clause boundaries rather than other places in the sentence.
  - **6 months**: infants are able to use prosodic information consistent with clausal units, and also demonstrate some sensitivity to prosodic information consistent with phrasal units.
  - **9 months**: infants show a preference for passages with pauses coincident with phrase boundaries over passages where the pauses are inserted elsewhere in the sentence.
  - **13 months**: infants can use Phonological Phrase boundaries to constrain lexical access. In sum, the sensitivity to cues carried by prosody appears to emerge within the first year of life.
- All these cues are consistent with the idea that patterns identification play a role in language learning.

Language Acquisition and Statistics

- Identifying stress and prosodic patterns can help children to identify word boundaries and gradually develop a sense of linguistic structure: word groupings in phrases and sentences.
- From simple frequency calculation of patterns, children can discriminate and distinguish categories in speech: sound segmentation, word segmentation, phrase segmentation.

Probabilities and Language Change

- Since children learn the language from they parents and environments, why do language change takes place?
- So, how does language change begin? - Actuation problem - Labov and Weinreich, Labov, and Herzogs (1968).
- How does an innovation spread?

Grad[illegible] Synchronic sociophonetic variation and diachronic sociophonetic variation

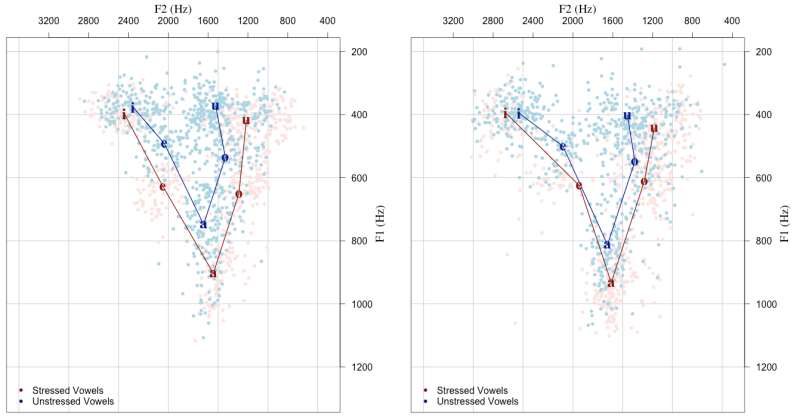


Figure: Stresssed vs. Unstressed Athenian Greek vs. Cypriot Greek vowels.

Gradience: Synchronic sociophonetic variation and diachronic sociophonetic variation

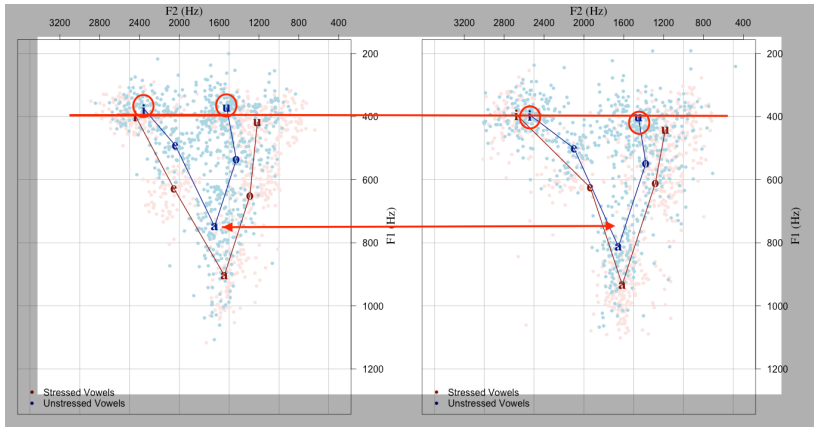


Figure: Stresssed vs. Unstressed Athenian Greek vs. Cypriot Greek vowels.

Section 2

Probabilities and Experimental Linguistics

Experimental Design: Setting and Designing Experiments

1. Psycholinguistics
2. Phonetics
3. Research on Language Acquisition
4. ...

Null and alternative hypothesis

A. The **null hypothesis** states that there is no difference between the two results:

$$GroupA - GroupB = 0 \quad (1)$$

B. The **alternative hypothesis** states that there is a significant difference between the two results.

$$GroupA - GroupB \neq 0 \quad (2)$$

The research hypotheses guide the design of any experiment.

Confidence Intervals

Ronald Aylmer Fisher (1890–1962): Trust the **95%**!

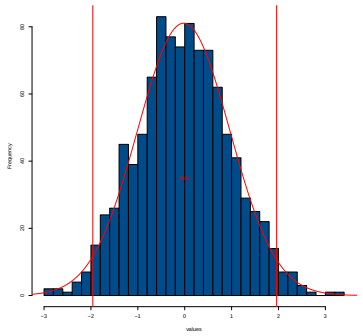


Figure: Confidence Intervals 95%

Systematic and non-systematic variation

Equally important is the error or deviation from the model to understand how good a statistical model. There are two types of variation:

1. the **systematic variation**: it occurs from the experimental modification, e.g., Drug vs. Placebo.
2. the **non-systematic variation**: physiological differences between patients or subjects. A doctor conducting an experiment might want the non-systematic variation to be as small as possible.

$$\frac{systematic\ variation}{non-systematic\ variation} \quad (3)$$

Models: Bad and non so Bad

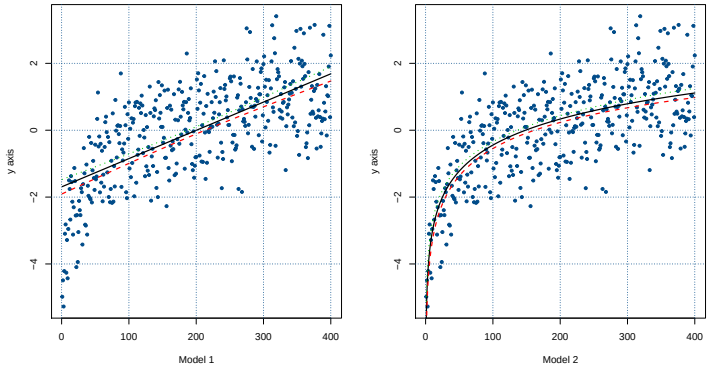


Figure: Model 1 and Model 2

Section 3

Probabilities and Linguistic Applications

Applications in the 1970s and 1980s (Klatt, 1989)

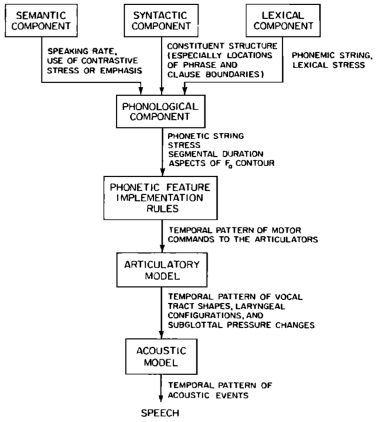


FIG. 2. Simplified block diagram of how a linguist might view the sentence generation process. An abstract linguistic representation for a sentence that is provided by the semantic component, syntactic component and lexical component undergoes various intermediate transformations before becoming an acoustic waveform.

Rule-based systems vs. Probabilities: Early Ambivalent Views

Jelenek(1976) argues for the use of decision strategies that are based on the collection of an appropriate set of probabilities determined experimentally. **Several of the speech understanding systems used estimates of the probability of a phonetic or lexical decision given the acoustic data in scoring the goodness of a theory, and each seems to have gotten into trouble by so doing. The problem is to analyze enough data to be sure of the probability of infrequent confusions. This is nearly impossible if one wants to take into consideration factors such as phonetic environment**.

- Early models of Speech to text and other applications were based on Rules.
- Linguists had to define the rules. But rules cannot capture all possible cases.

See Klatt (1977). Review of the ARPA Speech Understanding Project. *JASA* 62(6), 1345-1366

Current Approaches

- **Supervised Machine Learning**
- **Unsupervised Machine Learning**

Common Machine Learning Algorithms

1. Support Vector Machines,
2. linear regression,
3. logistic regression,
4. Naive Bayes,
5. linear discriminant analysis,
6. decision trees,
7. k-nearest neighbor algorithm,
8. Hidden Markov Models
9. Neural Networks (Multilayer perceptron).

Authorship Attribution

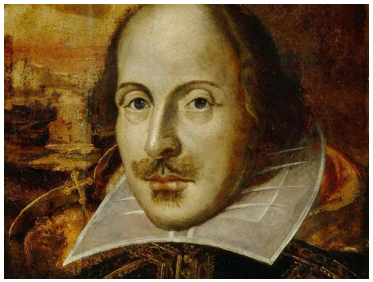


Figure: Did Shakespeare Write Double Falsehood?

Who is the author of **"Double Falsehood"**?
Boyd and Pennebaker (2015), in their article **"Did Shakespeare Write Double Falsehood? Identifying Individuals by Creating Psychological Signatures With Text Analysis"** investigate different linguistic aspects of the text and employed machine learning algorithms (L) to classify and find the author of the text (linear discriminant analyses (LDAs), J48 decision trees (DTs), support vector machines (SMO SVMs)).

Google Translate

Corpora: Texts and Applications

- **General corpora**: attempt to represent a language as a whole, e.g., the Brown Corpus, one million words from 500 American English texts of approximately 2000 words each, distributed across fifteen genres.
- **Specific corpora**: (e.g., Commerce, Finance, Food, or Law).
- Raw or **non-annotated** corpora.
- **Annotated corpora**.
- **Parallel corpora**. A parallel corpus is a corpus that consists of the same text in two or more languages. Parallel corpora can inform linguistic theory in general, enable the comparison of languages, and can have applications, such as in translation.

Conversational Agents: From Eliza to Siri

```
I am the psychotherapist.  Please, describe your problem
Each time you are
finished talking, type RET twice.

Hello
How do you do?  What brings you to see me?
I have problems
Why do you say you have problems?
Because I have
Can you elaborate on that?
Yes
I see...  Well, what makes you believe this is so?
A dream I saw last night
Is it because of your plans that you say a dream you saw
```

An implementation of Eliza on Emacs.

Conversational Agents: From Eliza to Siri

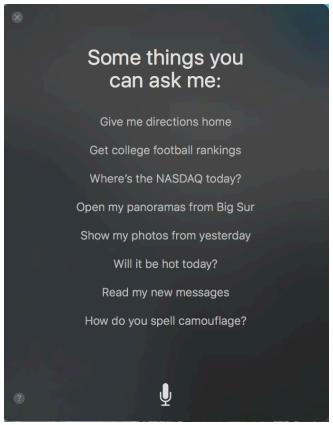


Figure: Siri on MacOS.

Other Applications

- Information Retrieval
- Topic Segmentation
- Speech to text and text to speech
- Domain Specific language applications: health, commerce, financial services.

Problems

- Acquiring that data
- Data are chaotic, noise

Ethical and philosophical considerations

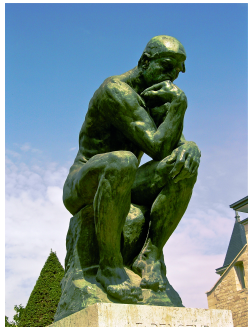


Figure: The Thinker (French: Le Penseur) by Auguste Rodin (1840 – 1917)

References

- Langus, Alan (2010). Struggling for Structure: cognitive origins of grammatical diversity and their implications for the Human Faculty of Language. *PhD Thesis*. International School for Advanced Studies.
- Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. Wiley Interdisciplinary Reviews. *Cognitive Science*, 1(6), 906914. http://doi.org/10.1002/wcs.78
- Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of the Accoustical Society of America*, 82(3), 737-793.
- Klatt, D. (1977). Review of the ARPA Speech Understanding Project. *Journal of the Accoustical Society of America*, 62, 1345-1366.