Seminar on Digital Humanities
Instructors:
Professor Nachum Dershowitz
Lena Dankin

Today:
CS in linguistics
by Yuval Ailer
Today's plan:

• What is linguistics?
• General challenges in linguistics
• Short history of the Generative grammar (Generativism)
• New horizons in the study of child language acquisition
• Anecdote on NLP
• Discussion
Movie clip

- https://www.youtube.com/watch?v=p3PfKf0ndik
What is linguistics?
Linguistics

From Wikipedia, the free encyclopedia

This article is about the field of study. For the journal, see Linguistics (journal).

"Linguist" redirects here. For other uses, see Linguist (disambiguation).

Linguistics is the scientific study of language, and involves an analysis of language form, language meaning, and language in context. The earliest activities in the documentation and description of language have been attributed to the 4th century BCE Indian grammarian Pāṇini,[4][5] who wrote a formal description of the Sanskrit language in his Aṣṭādhyāyī.[6]

Linguists traditionally analyse human language by observing an interplay between sound and meaning.[7] Phonetics is the study of speech and non-speech sounds, and delves into their acoustic and articulatory properties. The study of language meaning, on the other hand, deals with how languages encode relations between entities, properties, and other aspects of the world to convey, process, and assign meaning, as well as manage and resolve ambiguity.[8] While the study of semantics typically concerns itself with truth conditions, pragmatics deals with how situational context influences the production of meaning.[9]
why are we interested in language?

Turing test

programming languages
Were do CS and linguistics meet?
Collaborations:

**Interdisciplinary fields:**
- Semiotics
- Language documentation
- Translation
- Biolinguistics
- Clinical linguistics
- Computational linguistics
- Evolutionary linguistics
- Forensic linguistics
Collaborations:

Interdisciplinary fields:

- Semiotics
- Language documentation
- **Translation**
- Biolinguistics
- Clinical linguistics
- Computational linguistics
- Evolutionary linguistics
- Forensic linguistics
Collaborations:

Interdisciplinary fields:

- Semiotics
- Language documentation
- Translation
- **Biolinguistics (Cognitive science)**
- Clinical linguistics
- Computational linguistics
- Evolutionary linguistics
- Forensic linguistics
Collaborations:

Interdisciplinary fields:

• Semiotics
• Language documentation
• Translation
• Biolinguistics
• Clinical linguistics
• Computational linguistics
• Evolutionary linguistics
• Forensic linguistics
Collaborations:

Interdisciplinary fields:

- Semiotics
- Language documentation
- Translation
- Biolinguistics
- Clinical linguistics
- Computational linguistics
- Evolutionary linguistics
- Forensic linguistics  [https://www.youtube.com/watch?v=QCzH42efniU]
Computational linguistics:
Computational linguistics:
Computational linguistics:
Computational linguistics:

Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.

Traditionally, computational linguistics was performed by computer scientists who had specialized in the application of computers to the processing of a natural language. Today, computational linguists often work as members of interdisciplinary teams, which can include regular linguists, experts in the target language, and computer scientists. In general, computational linguistics draws upon the involvement of linguists, computer scientists, experts in artificial intelligence, mathematicians, logicians, philosophers, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists, among others.

Computational linguistics has theoretical and applied components. Theoretical computational linguistics focuses on issues in theoretical linguistics and cognitive science, and applied computational linguistics focuses on the practical outcome of modeling human language use.[1]

The Association for Computational Linguistics defines computational linguistics as:

...the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena.[2]
Computational linguistics:

• Origins:

Computational linguistics originated with efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals, into English.

It is sometimes grouped within the field of artificial intelligence.
Computational linguistics:

• Developmental approaches
• Structural approaches
• Production approaches
• Comprehension approaches
Computational linguistics:

• **Subfields:**

  • *Computational complexity of natural language*, largely modeled on automata theory, with the application of context-sensitive grammar and linearly bounded Turing machines.

  • *Computational semantics* comprises defining suitable logics for linguistic meaning representation, automatically constructing them and reasoning with them.

  • *Computer-aided corpus linguistics*, which has been used since the 1970s as a way to make detailed advances in the field of discourse analysis.

  • *Design of parsers* or chunkers for natural languages.

  • *Design of taggers* like POS-taggers (part-of-speech taggers).

  • *Machine translation* as one of the earliest and most difficult applications of computational linguistics draws on many subfields.

  • *Simulation* and study of language evolution in historical linguistics/glottochronology.
Example
How do linguists study language?

• Comparing languages
• Experiments
• Children's language acquisition
• And more..
How do linguists study language?

• Comparing languages
• Experiments
• children's language acquisition
• And more..
Foundations (up to the 1960’s)

• Structural linguistics:
  
  Ferdinand de Saussure

  • His concept of the sign/signifier/signified/referent forms the core of the field.

  • syntagmatic and paradigmatic axes of linguistic description.
Foundations (up to the 1960’s)

- Children learn by copying their parents (in Behaviorism)
- Languages can be different boundlessly
The big bang

• Noam Chomsky

Generative grammar

https://www.youtube.com/watch?v=dXE6ZafkRMI
Main ideas

• Transformational generative grammar

• deep structures and surface structures

( experiment )
Main ideas

• Transformational generative grammar
• Deep structures and surface structures
Main ideas

- **Transformational generative grammar**
- Deep structures and surface structures
Main ideas

- Transformational generative grammar
- Deep structures and surface structures

A kid broke a vase

A vase was broken by a kid
Main ideas

• Transformational generative grammar
• Deep structures and *surface structures*

A kid broke a vase

A vase was broken by a kid
Main ideas

- **Universal grammar**
  - Inherited knowledge of language structures.
  - The "Poverty of the stimulus" argument (sparsity)
Main ideas

• Chomsky hierarchy
Main ideas

• No use for statistical methods:
  • A. Colorless green ideas sleep furiously
  • B. Furiously sleep ideas green colorless

  • “It is fair to assume that neither sentence (1) nor (2) ... had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally “remote” from English. Yet (1), though nonsensical, is grammatical, while (2) is not.”

(Chomsky, Syntactic structures, 1957)
Using CS to research the problem

New horizons in the study of child language acquisition

Deb Roy 2009

New Horizons in the Study of Child Language Acquisition

Deb Roy
The Media Laboratory
Massachusetts Institute of Technology
drroy@media.mit.edu

Abstract
Naturalistic longitudinal recordings of child development promise to reveal fresh perspectives on fundamental questions of language acquisition. In a pilot effort, we have recorded 230,000 hours of audio-video recordings spanning the first three years of one child's life at home. To study a corpus of this scale and richness, current methods of developmental cognitive science are inadequate. We are developing new methods for data analysis and interpretation that combine pattern recognition algorithms with interactive user interfaces and data visualization. Preliminary speech analysis reveals surprising levels of linguistic fine-tuning by caregivers that may provide crucial support for word learning. Ongoing analyses of the corpus aim to model detailed aspects of the child's language development as a function of learning mechanisms combined with lifetime experience. Plans to collect similar corpora from more children based on a transportable recording system are underway.

Index Terms: language acquisition, rich longitudinal data, human-machine collaborative analysis, computational models

1. A New Kind of Data
Language is one of the defining features of the human species, unique in its compositional structure and referential capacity, critical for creation and transmission of cultural knowledge, devastating to an individual when impaired or lost. For all that are increasingly likely to complement audio with video recordings, the amount of video recorded tends to be exceedingly sparse due to the cost of analyzing video, and due to the disruptive observer effects of introducing video recording into home environments.

Four years ago my colleagues and I launched the Human Speechome Project with the goal of making a comprehensive and unbiased record of one child's (my son's) development at home [2]. The name of the project has two interpretations. First, our aim is to study speech in the context of the home, hence the combination of "speech + home" to yield the invented term "speechome." Second, this kind of data provides a basis for studying the environmental complement of genetic influences on language development, hence the naming parallel to the Human Genome Project.

We have completed the recording phase of the project yielding the Speechome corpus of approximately 90,000 hours of video and 140,000 hours of audio recordings spanning my son's life from birth to age three. Observational records of this magnitude are now possible due to the ease and affordability of technologies for digital data capture and storage.

The nature of recordings in our study has raised a variety of engineering, design, and privacy challenges. These have been addressed to a sufficient enough degree that I now believe ecologically-valid densely sampled observations of this kind will become pervasive in the study of child development and
New horizons in the study of child language acquisition

Deb Roy

tenured professor at MIT (media lab) and is Chief Media Scientist of Twitter
New horizons in the study of child language acquisition

Figure 1: Sample video frames from the kitchen and baby bedroom.
New horizons in the study of child language acquisition

- 4,260 hours of recording time.
- Spread over 444 of the 488 days.
- An average 9.6 hours of recordings per day.
- Approximately 70-80% of the child’s waking hours.
- In the 9-24 month period of life.
New horizons in the study of child language acquisition

Figure 2: Analysis framework: Machines that step into the shoes of a child.

10 million words
New horizons in the study of child language acquisition

Figure 3: Screen shot of TotalRecall interface showing approximately two minutes of data across three video and one audio channel. Blue and brown streaks indicate visual movement of people within and between rooms of the house.
New horizons in the study of child language acquisition

• Word Births:

Previous research:

Previous studies have shown that the frequency of a word in child directed speech predicts the age of acquisition of the word by the child.

Results:

Significant correlation of -0.29 between the log frequency of words in the child’s input and the date of the word birth (Figure 7a), and a stronger and significant correlation of -0.54 for nouns.
New horizons in the study of child language acquisition

- Word Births:

Results:

Figure 7: (a) Words plotted by their date of birth versus the log frequency of the word in caregivers’ speech over the 9-24 period. The best linear fit and r-value are shown in red. (b) The combination of how often a word is said and how it is said (based on vowel duration) predicts word births better than either alone.

Figure 8: Change in mean length of one caregiver’s utterances in relation to word births. Error bars are 95% confidence intervals.
Complexity of use by environment

The way a word is said (מבע, ביצוי)

Figure 8: Change in mean length of one caregiver’s utterances in relation to word births. Error bars are 95% confidence intervals.

https://translate.google.co.il/?um=1&ie=UTF-8&hl=iw&client=tw-ob#en/iw/Utterance
New horizons in the study of child language acquisition

• Word Births:

Evidence of fine lexical tuning of caregiver speech revealed in this analysis raises questions about how and why fine tuning occurs. Perhaps much of the caregiver speech early on is between adults and happens to contain words the child will eventually learn, and thus reflects the complexity of adult speech. As the child enters into the language, caregivers slowly adjust utterances in recognition of the child’s growing lexical abilities, bringing utterance complexity to a minimum to meet the child at the moment of birth and gently lifting him into more complex uses of each word type.
New horizons in the study of child language acquisition

Figure 9: Sixty minutes of father (green) and child (red) position traces in the living room reveal two social hot spots.
New horizons in the study of child language acquisition

Beyond N=1
Connection to the first part of the talk

• Poverty of the stimulus
• Statistical study and use
• Ability to generalize
More examples:

( past - Computational Linguistics, Volume 12, Numbers 2, April-June 1986 )

• THE MENTAL REPRESENTATION OF GRAMMATICAL RELATIONS

(MIT Press series on cognitive theory and mental representation)

Joan Bresnan, Editor

• COMMUNICATING WITH DATABASES IN NATURAL LANGUAGE

• (Ellis Horwood series in artificial intelligence)

Mark Wallace
More examples:

( present )

- Google translate

 בלשנות קיומית computational linguistics
More examples:

(present)

- Google Ngrams
More examples:
(present)

- Google Ngrams
More examples:

( present )
More examples:

( present )

- Sentiment analysis
A bit about NLP

**Goal:** Develop methods for processing, analyzing and understanding the structure and meaning of language
A bit about NLP

Applications:
• Search
• Ad matching
• Translation
• Sentiment analysis
• Speech recognition
• Chatbots
• Virtual assistants
• ...

A bit about NLP

Why is it hard?

• Ambiguity
• Variability
• Sparsity
• Grounding
A bit about NLP

Why is it hard?

- **Ambiguity** – “Look at the dog with one eye“
- Variability
- Sparsity
- Grounding
A bit about NLP

Why is it hard?

- **Ambiguity** – “kill bill”
- Variability
- Sparsity
- Grounding
A bit about NLP

Why is it hard?

• Ambiguity

• Variability

• Sparsity

• **Grounding** - Humans do not learn language by observing an endless stream of text
In popular culture
Sources:

- https://en.wikipedia.org/wiki/Linguistics
- https://www.haaretz.co.il/magazine/the-edge/.premium-MAGAZINE-1.3265954
- Advanced Methods in Natural Language Processing – Spring 2018

- New horizons in the study of child language acquisition
  https://dspace.mit.edu/handle/1721.1/65900
Questions ?