CMPT-413: Computational Linguistics

Anoop Sarkar
anoop@cs.sfu.ca
www.sfu.ca/~anoop/courses/CMPT-413-Spring-2003.html

Linguistics: the scientific study of language

- Language as a human instinct. It is learned but in a very peculiar way.
- Language is a system with complex (recursive) structure.
- Human language is unique to humans.
- Some misconceptions about language:
  - Language is cultural invention.
  - Children learn to talk by imitating their parents.

- Non-standard dialects are less logical than standard dialects.
- Language pervades thought, with different languages causing speakers to construe reality in different ways.
- Languages degrade over time.

Black English Vernacular (BEV)

- BEV has Negative Concord, which is attested in many Romance languages, such as French, Spanish and Italian.

(1) You **ain't** goin' to **no** heaven.

(2) Italian

a. Gianni **non** telefona a **nessuno**.
Gianni NEG telephones to nobody
'Gianni does not call anyone.'

b. **Nessuno** ha detto **niente**.
nobody has said nothing
'Nobody has said anything.'

- All languages and all dialects are grammatically complex and inherently sophisticated. To be more precise, they are equivalent from a computational point of view.

- For more, read *The Great Eskimo Vocabulary Hoax* by Geoffrey Pullum.

Language acquisition

- Children do not acquire language through language lessons from parents or by imitating their parents.

  Parents do not give explicit grammar lessons to their children. All they do in most cases is to give repetitive drills in simplified speech variety called Motherese.

  (3) Look at the doggie! See the doggie? There's a doggie!

- Children know things about language that they could not possibly have been taught.

  (4) a. A unicorn is in the garden $\Rightarrow$ Is a unicorn in the garden?

b. [A unicorn that is eating a flower] is in the garden $\Rightarrow$ Is a unicorn that is eating a flower in the garden?

c. [A unicorn that is eating a flower] is in the garden $\Rightarrow$ *Is a unicorn that eating a flower is in the garden?

– In order to form a question from a sentence containing an auxiliary verb, you need to identify the chunk that forms the subject, and invert the auxiliary verb around the subject.

  No parent would explicitly teach their three year old kid this grammatical rule about English. In fact, some adults probably aren't even aware of this rule themselves.

- Psycholinguists Stephen Crain and Mineharu Nakayama (1986) tested three-to-five year old children to see if they know this question formation rule.

  One experimenter was controlling a doll of Jabba the Hutt, and the other experimenter told the child to form a question, by saying, for example, “Ask Jabba if the boy who is unhappy is watching Mickey Mouse.”

  All the children tested gave the correct answer: “Is the boy who is unhappy watching Mickey Mouse?”

  None of the children said the ungrammatical string: “Is the boy who unhappy is watching Mickey Mouse?”

– Overgeneralization

(5) a. When she be’s in the kindergarten ....

b. He’s a boy so he gots a scary one.

c. She do’s what her mother tells her.

English has a rule of inflecting the verb in the present tense with *-s* if the subject is 3rd person singular.

The child that uttered the sentences in **??** is overgeneralizing this rule, not yet aware of the irregularities of this rule.

The fact that children make this kind of overgeneralization shows that they are aware of the rules governing the language, and that they are not just blindly imitating what their parents say.

– Dealing with unknown words: the *Wug*-test

A child successfully apply a grammatical rule to words he has never heard before, again showing that he is not just imitating his parents.

In an experiment, a child is shown a bird-like creature and is told that it is called a wug.

(6) This is a wug.

He is then shown two of them.

(7) Now there are two of them. There are two ____ .

All the children correctly say ‘wugs’.

- The speed with which children acquire language is remarkably rapid.

  By three years old, most children speak in fluent sentences, respecting detailed grammatical rules governing their community's spoken language.

- The most interesting theoretical assumption made in linguistics based on the facts of acquisition among others is that children are innately equipped with a language faculty that is programmed with a blueprint common to the grammars of all languages that tells them how to extract the patterns and rules of the speech of their parents.

  Noam Chomsky calls this common blueprint 'Universal Grammar'.

Language has structure

- Sentences are formed by putting words together. And speakers follow a certain set of rules in forming sentences, as can be seen by the fact that not all strings make a grammatical sentence.

(8) a. __ What he did was climb a tree.
b. __ What he ran was to the store.
c. __ Drink your beer and go home!
d. __ What are drinking and go home?
e. __ Linus lost his security blanket.
f. __ Lost Linus security blanket his.

- Almost all sentences a person hears or utters is new to that person. In fact, the number of possible sentences is infinite.

(9) a. This is a house.
b. This is a house Jack built.
c. This is the malt that lay in the house that Jack built.
d. This is the dog that chased the cat that killed the rat that ate the malt that lay in the house that Jack Built.
e. ...

- A brain has finite capacity. So it cannot possibly memorize infinite number of possible sentences. The number of words and the number of linguistic rules a person can know have to be finite.

- A person then must have knowledge of finite linguistic rules – grammar – that regulates how to combine these finite set of words to generate, in principle, an infinite number of sentences.

Linguistics: the various sub-fields

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics

Phonetics

- Phonetics studies the physical nature of speech sounds: their articulation, acoustic properties and perceptual characteristics.

- What physical gestures are used to articulate the speech sounds in a language? What do you do with your tongue, larynx, lips, etc., to produce different kinds of speech sounds?

- What aspects of the sounds in a language are relevant to their correct perception?

(10) Misheard Lyrics (mondegreens):
Wrong lyric: Excuse me while I squeeze this guy
Right lyric: Excuse me while I kiss the sky.

Phonology

- Phonology concentrates on the distinctive characteristics of the sounds in a language's sound system. It studies which combination of sounds are possible.

- What is the inventory of basic sounds (phonemes) in the language?

  For example, English makes a distinction between /l/ and /r/. They are different phonemes: e.g., *lock* vs. *rock*.

  But in Japanese, /l/ and /r/ are categorized as the same phoneme. So, for a Japanese speaker, it is very hard to hear the difference between *lock* vs. *rock*.

- What are the general rules for and constraints on stringing together those phonemes?

- vowel harmony  – A-na-u-pu

Morphology

- Morphology studies how words are formed: by putting morphemes (=minimal units with meaning) together, or by simply changing the category of the original word.

(11) We would be delighted if you send in your bill. However, if you don't, you will be.

(12)
Calvin: I like to verb words.
Hobbes: What?
Calvin: I take nouns and adjectives and use them as verbs. Remember when "access" was a thing? Now, it's something you do. It got verbed. Verbing weirds language.
Hobbes: Maybe we can eventually make language a complete impediment to understanding.

Syntax

- Syntax is concerned with sentence formation: we put words together to form larger linguistic units, then we put these together to build yet larger units, and so on and so forth until we build a complete sentence.

  A syntactic unit is called a ‘constituent’.

  (13) Subject-Verb-Object word order: English
  Maribel likes Olives.

(14) Subject-Object-Verb word order: Korean

Chunghye-ka olivu-lul cohaha-n-ta.
Chunghye-Nom olive-Acc like-Pres-Decl

‘Chunghye likes Olives.’

(15) I own the houses on the corner with a sign.

(16) One-eyed purple people eaters

Semantics

- Semantics investigates the interpretation of words, intermediate constituents and sentences.

- How is the meaning of a phrase related to the meaning of its parts?

- What sorts of meaning structures and meaning relations obtain in natural language?

(17) Interpretation of pronouns:
Two businessmen were discussing the sad state of sexual morality. “I didn’t sleep with my wife before we were married,” one of them declared self-righteously. “Did you?” “I’m not sure,” said the other. “What’s her name?”

(18) Interpretation of adjectives:

a. Beautiful dancer

(19) Interaction between quantifiers:
Everybody loves someone.

a. “Everybody is such that there is someone s/he loves.”

b. “There is someone that everybody loves.”

Pragmatics

- Part of the meaning of a sentence does not come from the parts of the sentence itself or the way they are combined. It comes from world knowledge or from inferences based on tacit conversational rules.

  (20) Can you pass me the salt?

  (21) This professor dresses even worse than Anoop.

Other subfields of Linguistics

- Psycholinguistics
- Sociolinguistics
- Historical linguistics
- Evolutionary linguistics
- Neurolinguistics

Terminology

- Grammar (prescriptive vs. descriptive)

  A descriptive grammar is a description of a set of rules that determines the form and meaning of words and sentences in a particular language as it is spoken in some community.

  It should not be confused with a prescriptive grammar, which is taught in school and explained in style manuals as guidelines for how one ‘ought’ to speak in a prestige or written dialect.

  When linguists talk about grammar, they are interested in descriptive grammar, and not prescriptive grammar.

- Parts of speech

  Noun: *John, he, she, cow, tomorrow, reiteration*. Among these, *he* and *she* are pronouns.

  Verb: *run, chase, teach, be, must*. Among these, *be* and *must* are auxiliary verbs.

  Determiner: *the, a, each, every , some*.

  Preposition: *in, at, between, under, beside, before*.

  Adjective: *blue, former, beautiful, good*.

  Adverb: *quickly, often, very, certainly*.

  Coordinating conjunction: *and, but, or*.

  Complementizer: *that, whether, if, than*.

- Grammatical relations

  subject

  object

  predicate

- Grammatical inflections: prefix, suffix, tense marker

  Attaching *un-* to the beginning of an adjective to mean *not*: *un-desirable, un-happy, un-sophisticated*.

  Plural *-s*: *apple-s, boy-s, car-s*.

  Past tense *-ed*: *walk-ed, return-ed, finish-ed*.

Computational Linguistics

- Linguists are happy to provide a theory which is equivalent to a Turing machine
- More constrained theories can be explored: computationally conservative theories are better
- Many ideas from programming language theory can be applied to linguistics: semantics, type theory, etc.
- The inherent learning based paradigm of natural language makes it a great challenge for machine learning.

Ambiguity: A key problem in Computational Linguistics

- pos: Reagan Wins on Budget, But More Lies Ahead
- pos: Lung Cancer in Women Mushrooms
- pos: Eye Drops off Shelf
- pos: Teacher Strikes Idle Kids

- pp: Two Sisters Reunited after 18 Years in Checkout Counter

- pp: Killer Sentenced to Die for Second Time in 10 Years

- parse: Red Tape Holds Up New Bridge
- parse: British Left Waffles on Falkland Islands
- parse: Stolen Painting Found by Tree

- wsd: Farmer Bill Dies in House
- wsd: Drunk Gets Nine Months in Violin Case
- wsd: Survivor of Siamese Twins Joins Parents
- wsd: Kids Make Nutritious Snacks
- wsd: Safety Experts Say School Bus Passengers Should Be Belted
- wsd: Iraqi Head Seeks Arms
- wsd: Prostitutes Appeal to Pope
- wsd: Local High School Dropouts Cut in Half

- discourse: Two Soviet Ships Collide, One Dies
- idiom: Chef Throws His Heart into Helping Feed Needy
- ellipsis:: Arson Suspect is Held in Massachusetts Fire

Formal Language Theory

- Σ is the alphabet, e.g. $\Sigma = \{a, b\}$

- Σ^* is the set of all strings with alphabet Σ
  The Library of Babel by Jorge Luis Borges (published in collections, e.g. *Ficciones*)

- A (formal) Language is a set of strings

- For example, a **regular language** is a set of strings constructed as follows:
  - ϕ is a RL
  - $\forall x \in \Sigma \cup \epsilon, \{x\}$ is a RL
  - If L_1 and L_2 are RLs then,
    - $L_1 \cdot L_2 = \{xy \mid x \in L_1, y \in L_2\}$
    - $L_1 \cup L_2$
    - ${L_1}^*$

    are RLs

- A (formal) Grammar is a finite description of a language using a specialized syntax
  e.g. REs are a grammar

- Each RE has an equivalent RL

- Closure properties: intersection, difference, complementation, reversal

- Equivalence of other grammars and languages: context-free languages and context-free grammars.

- Decidability or recognition for languages: given a string, decide whether it is in a language or not.

- A hierarchy of grammars and languages: The Chomsky Hierarchy regular $\subset$ deterministic CF $\subset$ context-free $\subset$ tree-adjoining $\subset$ indexed $\subset$ context-sensitive $\subset$ recursively enumerable