

NCSS '09: An Introduction to Language Technology

Joel Nothman jnothman@student.usyd.edu.au

January 7, 2009

The science of language

What components make up language?

Phonetics The production and perception of sounds in language. Use the International Phonetic Alphabet (IPA).

Clicks, pharyngeals and other strange sounds. Dark and light /l/: pull vs loop. Aspiration: top vs stop.

Phonology The sound structure of language. E.g. infinite vs impossible.

Morphology How we form words from small units of meaning (*morphemes*). Derivation (sing-er) and inflection (sing-s).

Rules and exceptions: walk→walked, but run→ran. Stranger morphemes: fan-bloody-tastic, but not fantas-bloody-tic.

Syntax How words form utterances: parts of speech (nouns, verbs, etc.), phrases, agreement (I walk but he walks).

Syntactic ambiguity: The boy saw the girl on the hill with the telescope

We draw *parse trees* to show how things fit together.

Semantics How language expresses meaning. *What has four wheels and flies?*

- *Lexical* semantics looks at the meaning of words:
How do you define a cup? when is it a bowl? a mug? a vase? Can it be measured?
Differentiate meanings of feel: I feel sick; I feel happy; I feel your hand; I feel like a drink; I feel like an idiot.
- *Compositional* semantics looks at how meaning is formed when words join together.
Grammatical nonsense? Colourless green ideas sleep furiously.
Two sentences, one meaning: Joel taught the class; The class was taught by Joel. Represent with symbolic logic?
Context matters: The astronomer saw the star vs the astronomer married the star.

Pragmatics How do we communicate effectively?

Q: Do you have the time? A1: Yes. A2: No, but I have a watch. A3: It's 4pm. A4: It's 4:02 and 53 seconds.

How does language operate?

Synchronic linguistics Explore the state of a single language at one time (is this possible?). Methods of research:

Field linguistics Meeting an unknown language? Find patterns, and *minimal pairs*, examples where a single change affects meaning. E.g. stress in English can be meaningful: subject, object, ?contrast, ?concrete.

Grammaticality judgements Use your own linguistic knowledge. But people disagree on what is grammatical.

Corpus linguistics Given a collection of text (newspapers, novels, conversations), what can we find out? Is the appearance or non-appearance of some phenomenon in a corpus proof?

Diversity of language Language is open-ended. Much of language is arbitrary: why is a table called table?

Language universals Are there rules which all languages abide by? What words must all languages have? What sounds?

Historical linguistics (or *diachronic linguistics*) How do languages change over time? How do languages interact?

French influence in English: sheep/mutton, cow/beef, calf/veal, swine/pork, deer/venison.

Sociolinguistics Language varies with society and social context. Dialect: have a shower vs take a shower; /t/ in butter.

Labov's fourth floor experiment. Speech accent archive: <http://accent.gmu.edu>. Code switching.

Language acquisition How do children gain linguistic knowledge? Is it better to be taught two languages or one?

Psycholinguistics How do our minds process language? E.g. *garden path sentences*: The horse raced past the barn fell.

The old man the boat. Time flies like an arrow; Fruit flies like a banana.

Neurolinguistics How does this work physically inside our brains? What happens with brain damage? Animal language?

Computing language

Aims

- to better study and model language (and its processing) using computers;
- to improve human-computer interaction;
- to perform tasks which otherwise require a lot of human work.

Tasks

Machine translation Try it: <http://translate.google.com/>, <http://babelfish.yahoo.com/>.

Speech recognition and synthesis Say aloud: recognise speech vs wreck a nice beach.

Interactive dialogue systems Meet ELIZA, a computer therapist. (Or just call Telstra; how long before you scream?)

Speaker identification and verification Is voice as good as a signature?

Intelligent word-processing Smarter red and green squiggles.

Information retrieval Find relevant documents: disambiguate words with multiple meanings.

Information extraction Extract pieces of knowledge from documents, e.g. encyclopedia articles, medical reports.

Question answering Google when was mozart born?; how tall is tom cruise?; what genus is a chimpanzee? (what broke?!)

Classification and clustering Find spam. Access your emails easier. See e.g. <http://news.google.com>.

Plagiarism detection Find students who copied their homework.

Authorship analysis Did Shakespeare write all his plays?

Summarisation Intelligently summarise documents, or multiple documents on a single topic.

Linguistic processing tools

To reach the higher goals above, we need various smaller tools:

Word sense disambiguation Which meaning of *serve*? help with food or drink; hold an office; put ball into play. Which meaning of *dish*? plate; course of a meal; communications device. If *serve* and *dish* are together, it's easier.

Pronoun resolution (a) The thieves stole the paintings. They were subsequently sold. (b) The thieves stole the paintings. They were subsequently caught. (c) The thieves stole the paintings. They were subsequently found.

Textual entailment Hypothesis: Golinkin has written eighteen books. Possible evidence: David Golinkin is the editor or author of eighteen books, and over 150 responsa, articles, sermons and books.

How might these tools help in interactive systems? machine translation? summarisation?

These in turn take advantage of lower-level tools (part-of-speech tagging, syntactic parsing, morphological analysis, etc.)

I want more!

Australian Computation and Linguistics Olympiad OzCLO was first run in 2008. If you're interested in joining in the 2009 competition for high school students, contact admin@ozclo.org.au or visit <http://www.ozclo.org.au>.

NLTK If you like the idea of playing with language using computers, check out the Natural Language Toolkit (<http://www.nltk.org>), and their book (<http://www.nltk.org/book>), which also acts as a tutorial for Python.

Sources used

- Steven Bird, Ewan Klein and Edward Loper. *Natural Language Processing*. <http://www.nltk.org/book>
- Guido Ipsen, *Linguistics for Beginners*, University of Kassel. <http://www.uni-kassel.de/fb8/misc/lfb/html/text/0frame.html>.
- North American Computational and Linguistic Olympiad. *Information about Language Technology* <http://namclo.linguistlist.org/cool.cfm>.
- Frank Richter, slides for *Introduction to Computational Linguistics*, Eberhard-Karls-Universität Tübingen. <http://www.sfs.uni-tuebingen.de/~fr/teaching/ws06-07/icl/>.

Some problems!

1 Babylonian

Source: NACLO 2009 Practice Problems <http://www.naclo.cs.cmu.edu/>

The earliest known writing system originated over 5000 years ago in what is now Iran, Iraq and other parts of Western Asia. This writing system, called *cuneiform*, was used by the ancient Persian kings to make their decrees known, and to audit the tax returns of their many subjects. The characters were inscribed on clay or stone tablets using wedge-like instruments. Although many inscriptions have survived, the writing system was not deciphered by modern scholars until 1846.

In this problem you will carry out the kind of work that these scholars had to do to decipher the cuneiform writing system. The image on the right is an actual fragment from a Babylonian educational document that was discovered in 1811. It was this document that allowed scholars to unlock the number system used by the ancient Babylonians. From this, scholars were able to extend their understanding to the entire writing system. Many of the characters are illegible because of the ravages of time. Nevertheless, it is possible to figure out what the missing characters should be. Your job is to fill in the missing characters.



2 Kannada

Source: OzCLO: By Mirjam Fried

Kannada is one of the major languages of India, spoken by more than 25 million people primarily in the South of the country, near Bangalore. It is a very old language and it uses its own writing system. For the purpose of this puzzle, the Kannada letters are transcribed using the Roman alphabet. The letters L, D, T, and N represent a special pronunciation with the tongue curled upward. Fill in the blanks! (Hint: There is no translation for “the” in Kannada.)

mane	'house'	manege	'to (the) house'	simha	'lion'	simhakke	'to (the) lion'
peeTe	'market'	peeTege	'to (the) market'	kalkatta	'Calcutta'	kalkattakke	'to Calcutta'
tande	'dad'	tandegge	'to dad'	manushya	'man'	manushyanige	'to (the) man'
roTTi	'bread'	roTTige	'to (the) bread'	amma	'mom'	ammanige	'to mom'
chaTNi	'chutney'	chaTNige	'to (the) chutney'	huDuga	'boy'	huDuganige	'to (the) boy'
hakki	'bird'	hakkige	'to (the) bird'	sneehita	'friend'	sneehitanige	'to (the) friend'
taayi	'mother'	taayige	'to mother'	hamsa	'swan'	'to (the) swan'
jooLa	'corn'	jooLakke	'to (the) corn'	akka	'older sister'	'to (the) older sister'
pustaka	'book'	pustakakke	'to (the) book'	tangi	'younger sister'	'to (the) younger sister'

3 Quechua

Source: NACLO 2009 Practice Problems <http://www.naclo.cs.cmu.edu/>

Quechua is a South American language family with about 8,000,000 speakers, most of whom inhabit the Andes mountains of Peru, Bolivia and Ecuador. Quechua was the official language of the Tawantinsuyu or Inca Empire before the Spanish invasion of 1532. For hundreds of years Cuzco, in what is now Peru, was the capital of the Inca empire. The sentences below represent the variety of Quechua currently spoken in Cuzco and in the area around Lake Titicaca.

Below are some sentences in Quechua, with their translations in random order. Indicate which translation goes with each Quechua sentence.

A	Antuqaq chakranpiqa t'ikashanmi papa.	1	Potatoes may be growing in Antuka's field.
B	Siskuq chakranpiqa wiñashanmi sara.	2	Barley may be flowering in Antuka's field.
C	Siskuq chakranpiqa rurushansi kiwña.	3	Corn is growing in Sisku's field.
D	Antuqaq chakranpiqa t'ikashanchá kiwña.	4	I've heard corn is growing in Sisku's field.
E	Siskuq chakranpiqa wiñashansi sara.	5	I've heard barley is yielding fruit in Sisku's field.
F	Antuqaq chakranpiqa wiñashanchá papa.	6	Potatoes are flowering in Antuka's field.

[I] Provide English translations for the following Quechua sentences:

- Istuchaq chakranpiqa t'ikashansi sara.
- Sawinaq chakranpiqa wiñashanchá kiwña.
- Tumasaq chakranpiqa rurushanmi papa.
- Kusiq chakranpiqa t'ikashanchá papa.
- Inashuq chakranpiqa rurushansi kiwña.

4 Better sorry than shunk

Source: 1st NAMCLO: 2007.

Here is an English sentence with a nonsense verb in it (*in italics*).

After the monster had *shunk* its prey, it dragged it back into the cave.

[I] Fill in the other forms of this verb in the following sentences:

- (a) She used to groundhogs.
- (b) Now she possums for a living.
- (c) When she was in Eugene she thirty-three possums in one day.
- (d) Then she took us possum-..... in the Cascades.

[II] Are there any other possible solutions to this problem? Please give all solutions, sorted by how likely they are correct, and explain your answer.

5 We are all molistic in a way

Source: 1st NAMCLO: 2007.

Imagine that you have heard these sentences:

- Jane is molistic and slatty.
- Jennifer is cluvious and brastic.
- Molly and Kylie are slatty but danty.
- The teacher is danty and cloovy.
- Mary is blitty but cloovy.
- Jeremiah is not only sloshful but also weasy.
- Even though frumsy, Jim is sloshful.
- Strungy and struffy, Diane was a pleasure to watch.
- Even though weasy, John is strungy.
- Carla is blitty but struffy.
- The salespeople were cluvious and not slatty.

[I] Which of the following would you be likely to hear?

- (a) Meredith is blitty and brastic.
- (b) The singer was not only molistic but also cluvious.
- (c) Mary found a dog that was danty but sloshful.

[II] What quality or qualities would you be looking for in a person?

- (a) blitty
- (b) weasy
- (c) sloshful
- (d) frumsy

[III] Explain all your answers. (Hint: The sounds of the words are not relevant to their meanings.)

6 Who are you talking about?

Source: 1st IOL: Borovetz 2003. By Maria Rubinstein.

When describing how personal and reflexive pronouns work in various languages, linguists make use of so-called subscripts—Roman letters (typically *i*, *j*, *k* . . .) which mark pronouns and some other words in sentences. The character * (asterisk) is also used. Here are some English examples.

1. John_{*i*} saw himself_{*i*} in the mirror.
2. John_{*i*} says that he_{*i/j/*k*} doesn't know Peter_{*k*}.
3. The boy_{*i*} is playing with his_{*i/j*} gun.
4. His_{*i*} teacher_{*j*}'s influence is easily seen in his_{*i/*j/k*} work.
5. The girl_{*i*} saw her_{**i/j*}.

[I] Explain the meaning of the subscripts and the asterisk.

[II] Add subscripts and asterisks (where appropriate) to the following sentences:

- (a) She doesn't like this trait in herself.
- (b) The father took his son to his room.
- (c) John knows that Peter has given his book to his son.

7 Verbal abilities

Source: 1st IOL: Borovetz 2003. By Boris Iomdin.

Consider the following pairs of verbs with closely related meanings:

accuse	rebuke
denounce	reprehend
command	instruct
advise	guide
assure	convince

It is known that all the verbs in the left-hand column have a certain ability that the verbs in the right-hand column lack.

[I] Identify the ability in question.

[II] Find the verbs that also have this ability among the following: extort, threaten, forbid, swear, shout, approve, refuse, rob, dedicate, lose, scold, give up, demand.

[III] Try to find two more verbs with the same ability.

8 A donkey in every house

Source: 1st NAMCLO: 2007

Consider these phrases in Ancient Greek (in a Roman-based transcription) and their unordered English translations:

A	ho tōn hyiōn dulos	1	the donkey of the master
B	hoi tōn dulōn cyrioi	2	the brothers of the merchant
C	hoi tu emporu adelphoi	3	the merchants of the donkeys
D	hoi tōn onōn emporoi	4	the sons of the masters
E	ho tu cyriu onos	5	the slave of the sons
F	ho tu oicu cyrios	6	the masters of the slaves
G	ho tōn adelphōn oicos	7	the house of the brothers
H	hoi tōn cyriōn hyioi	8	the master of the house

[I] Match each English translation to its corresponding Greek phrase.

[II] Translate into Ancient Greek:

- (a) the houses of the merchants
- (b) the donkeys of the slave

Note: The letter \bar{o} stands for a long o.

9 Cognates

Source: By Dragomir Radev.

Linguists group languages into families based on their historical relationship. English belongs to the Indo-European language family along with several hundred other languages and dialects. The Indo-European family of languages is further divided into branches that include Germanic languages (e.g., English, German, and Norwegian), Indo-Iranian languages (such as Hindi, Bengali, or Farsi), and Romance languages (languages descended from Latin such as French, Italian, and Romanian). All these languages share some common vocabulary. Languages in the same branch generally share a greater deal of their vocabulary. Words that are historically related in different languages are called cognates.

Many words in English are cognates of words in Romance languages, which are also related to each other. The following list includes translations of a number of English words into some number of Romance languages.

cantare, escola, stella, étoile, estrela, étudiant, scuola, escuela, cantar, studiare, école, estrella, estel, estudiar, chanter

- [I] How many different English words do the fifteen words above correspond to? Hints: Cognates are likely to have somewhat similar spelling. Try to lay out the words in a grid. Note that the same word may appear with identical spelling in more than one language.
- [II] How many languages are included in the sample?
- [III] Can you try and translate each of these words into English (knowing some words above have English cognates)?
- [IV] Escuela is one of the words in Spanish on the list above. How do you say étudiant in Spanish (which is not on the list)? Try to get as close as possible to the correct translation, even if you cannot get it quite right.

10 Weasel

Source: By John Blatz and Jason Eisner.

The following sentence, though bizarre and deliberately confusing, is actually grammatically correct:

The weasel that a boy that startles the cat thinks loves smiles eats.

- [I] Answer the following questions. In some cases, the answers may be *nobody* or *nothing in this sentence*.
 - (a) What is the subject of this sentence? (Give a single-word answer.)
 - (b) How many verbs are in the sentence?
 - (c) Who startles whom or what?
 - (d) Who thinks what?
 - (e) Who loves whom or what?
 - (f) Who smiles?
 - (g) Who eats whom or what?

11 Hieroglyphics

Source: By Tom Payne, based on research by Jean-François Champollion.

On the right are representations of two Egyptian *cartouches* from the Greco-Roman period. A cartouche is an oblong set of hieroglyphic characters that represents a name, a word or a phrase. One of these cartouches represents the name of the Queen *Cleopatra*. Your task is to figure out which one means *Cleopatra*, and what the other one probably means (Hint: the other cartouche is the name of another famous character from Ancient Egyptian history). This is exactly the kind of work that archaeological linguists do when they attempt to interpret writings in ancient languages.



12 Magic Square (difficult!)

Source: By John Blatz.

The sequence of words [dog, ore, get] has the property that taking the n th letter of each word, in order, forms the n th word. For example, the 2nd letters of dog, ore, and get are o, r, and e, which spell the second word ore.

- [I] Find a sequence of six 6-letter English words (no proper nouns, please!) with the same property. Hint: one such sequence exists containing the words spread and acetic; another includes abrupt and pierce.
- [II] Find four 4-letter words similarly forming a magic square, but whose diagonals are also words.
- [III] Computing challenge: how might you store data to make this more efficient for a computer to solve?