The Study of Language

Reading:

> Chapter 1 of Jurafsky and Martin

Reference:

> David Crystal, *The Cambridge Encyclopedia of Language*, Cambridge University Press, 1987.
>
> Francis Katamba, *Introduction to Phonology*, Longman, 1989.
>
> S.C. Levinson, *Pragmatics*, Cambridge University Press, 1983.

Describing A Language

- How do you describe a language ?
- Consider the spoken utterance:

  Shut the door !

  To describe it exhaustively we need to discuss such things as:
  - ◇ the sounds which make up each word
  - ◇ the order in which the words occur
  - ◇ the meaning of the individual words
  - ◇ the type of situations in which such a sentence can be uttered
  - ◇ . . .

Describing A Language (cont)

- Linguists typically make the assumption that language can be described on a number of *levels* and that these levels can largely be studied *independently*.

- Linguists do not fully agree as to what these levels are or how they are related.

- Generally agreed levels of description are:
  - ◇ phonetics
  - ◇ phonology
  - ◇ graphetics
  - ◇ graphology
  - ◇ morphology
  - ◇ syntax
  - ◇ semantics
  - ◇ discourse
  - ◇ pragmatics

Phonetics

Phonetics is the study of how to describe and classify speech sounds.

- most phonetic classification schemes use *articulatory* descriptions – e.g. descriptions which refer to how the sounds are produced in the human vocal tract;
- phonetics is independent of the language of the speaker – it is concerned with the total range of vocal sounds humans may produce and detect

E.g. *p* and *b* may be described as *bilabial* (both lips) *plosives* (air built up then released explosively).

Phonology

Phonology is the study of the principles that govern which sounds are used in actual human languages and that explain the patterns and variations that occur in individual languages.

Example (Katamba):

tpat   lsender   bintlement   zvetsin

Only one of these is a possible English word. How do we know this ?

Phonology (cont)

- One approach to phonology seeks to find minimal sound units which if changed in a word alter the word's meaning.

  E.g. if the sound [p] is replaced by [b] in *pig* we get *big*.
  These minimal sounds units are called *phonemes*

- *Comparative phonology* looks at questions such as (Crystal, 1987):
  - ◇ Which sounds occur most frequently in the world's languages? E.g.

    83% of 317 languages studied in the University of California Phonological Segment Inventory Database (UPSID) survey have some form of the dental/alveolar sibilant /s/.

  - ◇ What is the variation in number of phonemes per language?

    E.g. The language with the smallest number in the UPSID survey is Rotokas (Indo-Pacific) with 11 while the largest is !Xu (Khosian) with 141 (English has 44).

Graphetics and Graphology

- By analogy with the distinction between phonetics and phonology, a distinction can be made in *written* language between the physical characteristics of the medium and the characteristics of symbol systems used in actual languages.

- *Graphetics* is the study of the physical properties of the symbols making up writing systems.
  Graphetics involves the study of means of production of symbols – handwriting, printing, electronic – and the material used – pens, brushes, paper, tablets.

- *Graphology* is the study of the characteristics of the symbol systems used in human languages, their patterns and variations
  - ◇ on analogy with *phoneme* the notion of *grapheme* is introduced – the smallest unit in the writing system whose change affects meaning E.g. Since *cat* and *bat* have different meanings *c* and *b* are different graphemes.

Aside on Writing Systems

- Physical faculties required for spoken language
    - ◇ Broca's area, region of brain associated with language
    - ◇ vocal apparatus

    arose in homo sapiens between 2 million and 300,000 years ago
- Written language only emerged between 3500-3100BC in 3 independent centres:
    - ◇ amongst the Harappans in the Indus Valley (Indus language)
    - ◇ amongst the Sumerians in Mesopotamia (cuneiform)
    - ◇ amongst the Egyptians (hieroglyphics)

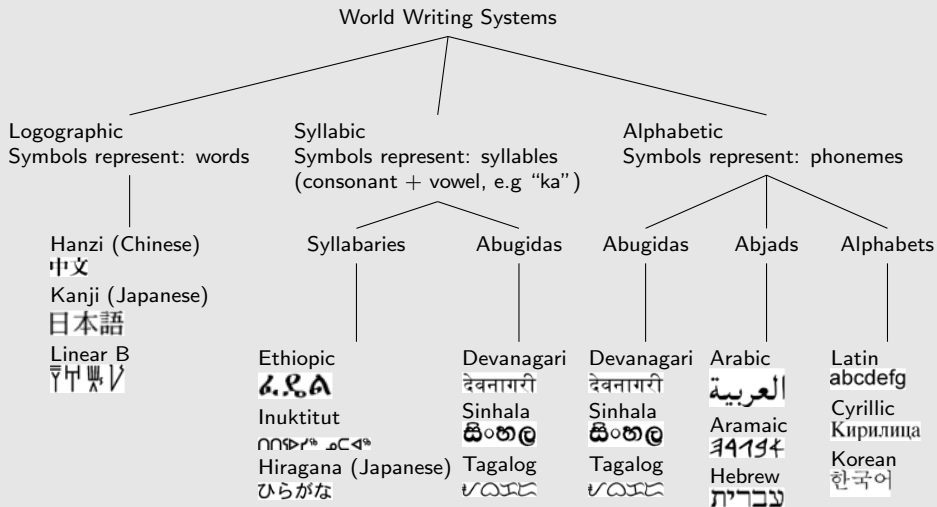   Some dispute about whether earlier evidence counts as writing

- What is a writing system?

  "a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer"

  Peter T. Daniels, *The World's Writing Systems*
- Not clear
    - ◇ How many writing systems there are – omniglot.com lists 114 + 10 undeciphered
    - ◇ How many languages lack writing systems

Taxonomy of Writing Systems

World Writing Systems

Logographic
Symbols represent: words

Syllabic
Symbols represent: syllables
(consonant + vowel, e.g "ka")

Alphabetic
Symbols represent: phonemes

Hanzi (Chinese)
中文

Kanji (Japanese)
日本語

Linear B
𐀳𐀠𐁉𐀷

Syllabaries

Abugidas

Abugidas

Abjads

Alphabets

Ethiopic
ፊደል

Inuktitut
ᐃᓄᒃᑎᑐᑦ

Hiragana (Japanese)
ひらがな

Devanagari
देवनागरी

Sinhala
සිංහල

Tagalog
ᜆᜄᜎᜓᜄ᜔

Devanagari
देवनागरी

Sinhala
සිංහල

Tagalog
ᜆᜄᜎᜓᜄ᜔

Arabic
العربية

Aramaic
𐡀𐡓𐡌𐡉𐡀

Hebrew
עברית

Latin
abcdefg

Cyrillic
Кириллица

Korean
한국어

Morphology

Morphology is the study of the *structure* of words.

- The smallest *meaningful* elements into which words can be decomposed are called *morphemes*.

dis-agree-ment-s	4 morphemes
un-happi-ness	3 morphemes
yes	1 morpheme
anti-dis-establish-ment-arian-ism	6 morphemes

- Morphology is typically divided into the two subfields of *inflectional* and *derivational* morphology

Morphology

- *Inflectional morphology* is concerned with the differing forms one word takes to signal differing grammatical roles

boy/boys	singular/plural
see/saw	present/past
go/goes	1st person/3rd person

  Inflected forms are of the same syntactic category (noun, verb)

- *Derivational morphology* is concerned with how new words may be constructed from component morphemes.

institute	institutionalisation
institution	institutionalisational
institutional	institutionalisationally
institutionalise	

oikos	+ logy	→	ecology
ecology	- logy	→	eco
eco	+ tourism	→	ecotourism

Syntax

Syntax is the study of the *structure* of sentences.

- However, there is no consensus about what constitutes a sentence! Definitions (over 200 are on offer) include:
    - ◇ 'the complete expression of a single thought'
    - ◇ 'consists of a subject and a predicate'
    - ◇ 'begins with a capital letter ends with a full stop'

  But all these have exceptions.

  Proceed by analysing linguistic constructions that actually occur and identifying those that can stand on their own and possess syntactic structure.

- Traditionally a hierarchy of sentence structure is recognised:

  sentence → *clauses* → *phrases* → *words* → *morphemes*

  Like 'sentence' the terms 'clause' and 'phrase' do not have precise definitions, but are useful in talking about sentences.

Syntax (cont)

- The structure of sentences is frequently displayed graphically by means of *phrase structure trees*.
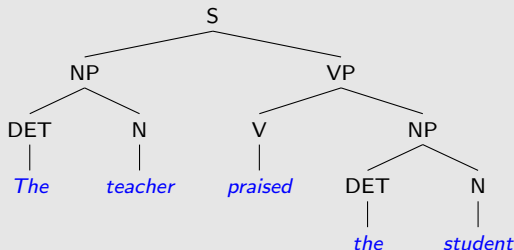
The words of the sentence label the tree's leaves, while higher nodes are labelled with *grammatical categories*, such as *NP* (noun phrase), and *VP* (verb phrase).

- A *labelled bracketting* may alternatively be given, e.g.

$[_S [_{NP} [_{DET} \text{the}] [_N \text{teacher}]] [_{VP} [_V \text{praised}] [_{NP} [_{DET} \text{the}] [_N \text{student}]]]]$

Syntax (cont)

- A catalogue of structural descriptions purporting to describe the sentence structure of a language is called a *grammar*.

- These structural descriptions are called *grammar rules*.

- Important to distinguish *descriptive* vs. *prescriptive* grammar rules (e.g. 'You should never end a sentence with a preposition')

Semantics

Semantics is the study of *meaning* in language.

- The study meaning has preoccupied generations of philosophers
- Views include:
  - ◇ Words → things (Plato)
    Words name things and their meaning is the thing they name.
    E.g. *Sheffield* names the city Sheffield. But what about words like
    tradition or *difficult* or *enquire* ?
  - ◇ Words → concepts → things (Ogden and Richards)
    Words are associated with concepts (mental entities) which may in
    turn be associated with things. But what are concepts?
  - ◇ Stimuli → words → responses (Bloomfield)
    The meaning of a word is the set of events that give rise to and follow
    upon its occurrence (behaviourist theory of meaning).
- In NLP, ignore questions about the ultimate nature of meaning
  Instead ask: Can we derive a representation of a text that can be
  usefully manipulated to achieve some goal?
  (e.g. question answering, translation, etc.).

Discourse Analysis

The study of discourse involves the analysis of the structure and meaning of discourse or text (i.e. of multiple sentence linguistic phenomena). Topics of study include:

- textual structure:
  - ◇ coreference: how to we know that two differing linguistic expressions refer to the same thing ?

    John plucked up the courage to ask the president for a raise.
    He was sitting stonily behind his mahogany desk.

    John plucked up the courage to ask the president for a raise.
   He was shaking as he entered the office.

    How do we know who *He* refers to here ?
  - ◇ ellipsis: structure may be omitted and need to be recovered from the surrounding text.

    Where did you see her ? ⊔ Next to the library.
    John chose chocolate, Sarah ⊔ butterscotch.

Discourse Analysis (cont)

◇ coherence: what makes multi-sentence texts coherent ?
Rhetorical structure theory, for example, proposes a set of relations
which hold between sentences or clauses. Tree structures are used to
define possible sequences of relations.
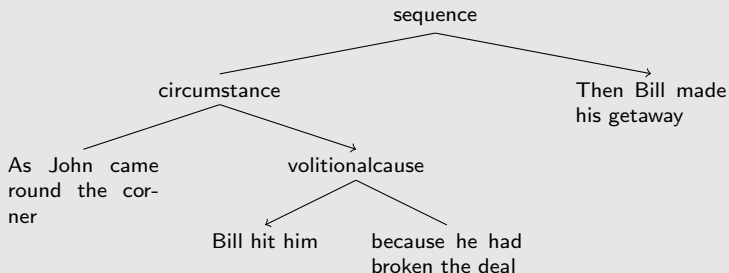
- conversational structure
  ◇ turn taking: how do people signal the end to a conversational turn or
  the desire to take one ?
  ◇ conversational maxims, such as that relevant information only is given
  in response to a query.

Pragmatics

Pragmatics is the study of how humans use language in social settings to achieve goals.

Topics of study include:

- speech acts: *promising, resigning, deploring, insisting*

- conversational implicature (Levinson):

  A: Can you tell me what time it is ?
  B: Well, the milkman has come.

  B does not literally respond to A, but says something which *implies* an answer. A is supposed to be able to *infer* the answer from B's response.

The subject area of pragmatics is not well-defined. In particular there is much overlap with discourse analysis.

Summary

- Human language is an extremely complex system.
- To study it researchers decompose language into different aspects or levels, assuming each may be studied independently.
- Broadly agreed levels of study for written language are:
  - ◇ graphetics/graphology (aka the study of writing systems)
    - study of the means/materials used to produce written language and the nature of the symbol systems used to encode it
  - ◇ morphology
    - study of the structure of words in terms of meaningful sub-elements
  - ◇ syntax
    - study of the structure of sentences
  - ◇ semantics
    - study of the meaning in language
  - ◇ discourse analysis
    - study of the structure and meaning of multi-sentence discourse
  - ◇ pragmatics
    - study of how humans use language in social settings to achieve goals
- Computational linguistics has developed computational models to address all these aspects of language ...