COM6115: Text Processing

Background: Linguistic Basics

Mark Hepple

Department of Computer Science
University of Sheffield

Linguistic Preliminaries

Peter approached the door.
He knocked and went through it.

- To describe this text we might look at:
  - ◇ the form of the words that appear
  - ◇ the order of the words within sentences
  - ◇ the meaning of individual words
  - ◇ how they combine to give the meaning of sentences
  - ◇ how sentences link together in the overall text meaning
- Linguists have assumed that language can be described at a number of levels, which can be studied independently

Levels of linguistic analysis

The levels of linguistic description include:

- **Phonetics**
  - ◇ studies how to describe and classify *speech sounds*
  - ◇ examines the range of vocal sounds humans may produce and distinguish, for use in communication
- **Phonology**
  - ◇ studies the principles that govern how speech sounds are used in human languages
  - ◇ identifies minimal units (*phonemes*) that can distinguish words

    e.g. *p/b* in *pit* vs. *bit*
  - ◇ explains how phonemes may be combined in words for each language

    e.g. *zvetsin* vs. *bintle*

Levels of linguistic analysis (contd)

For the analysis of text, rather than speech, can identify two additional levels, analogous to phonetics and phonology:

- **Graphetics**
  - ◇ studies the physical symbols making up writing systems
  - ◇ includes means of producing symbols, & materials used

    e.g. handwriting, printing, electronic

    e.g. pens, ink, brushes, paper, tablets
- **Graphology**
  - ◇ studies the *systems* of symbols used in languages, their patterns and variations
  - ◇ identifies the smallest units whose change affects meaning

    — called *graphemes* (by analogy with *phoneme*)

Levels of linguistic analysis (contd)

. . . further levels of analysis include:

- **Morphology**
  - ◇ studies the *structure* of words
  - ◇ identifies the smallest *meaningful* elements into which words can be decomposed – called *morphemes*

    e.g. disagreements ⇝ dis/agree/ment/s (4 morph's)

- **Syntax**
  - ◇ studies the *structure* of sentences, and how this differs between languages

    e.g. English shows SVO order (Subject/Verb/Object)

    other languages show other default orders: SOV, VSO, free order

Levels of linguistic analysis (contd)

. . . further levels of linguistic analysis:

- **Discourse Analysis**
  - ◇ studies interpretation of spoken & textual *discourse*

    i.e. of *multi-sentence* communications
  - ◇ various processes connect meaning *across* sentences

    e.g. pronoun *coreference* in: *Peter arrived. He knocked.*
- **Pragmatics**
  - ◇ studies how humans use language in social settings to achieve goals
  - ◇ includes how real intent of utterance may be implied by indirect statement, and so must be inferred by hearer

    e.g. *Can you reach the salt?* as a request for the salt

Syntax

- Studies the principles governing how words are combined to form sentences, and how this differs across languages

  i.e. it studies the *structure* of sentences

- A standard view:
  - ◇ *words* combine to form *phrases*

    e.g. *the* + *book* ⟶ (*the book*)
  - ◇ *words* and *phrases* combine to form *larger phrases*

    e.g. *at* + (*the book*) ⟶ (*at* (*the book*))
  - ◇ ultimately producing *sentences*

    e.g. *Bill looked* (*at* (*the book*))
  - ◇ hence, sentences have a *hierarchical* structure

Syntax: parts of speech

- Linguists group words into classes showing similar behaviour
  - ◇ called *parts of speech*
  - ◇ a.k.a. *word class*, or *syntactic / lexical category*
- A possible basic set (some disagreement):
  - ◇ Noun (N), e.g. boy, shoe, foot
  - ◇ Verb (V), e.g. eats, saw, runs
  - ◇ Adjective (Adj), e.g. red, tall, clever
  - ◇ Adverbial (Adv), e.g. quickly, smoothly, loudly
  - ◇ Preposition (P), e.g. in, of, to, from
  - ◇ Determiner (Det), e.g. the, a, an
  - ◇ Auxiliary (Aux), e.g. will, has, did
  - ◇ Complementiser (Comp), e.g. that, whether, if
  - ◇ Conjunction (Conj), e.g. and, or, but

Syntax: parts of speech (contd)

- This grouping partly based on semantic intuitions

  e.g. *prototypically*, find that:

  - ◇ nouns refer to people, animals, concepts, *things*
  - ◇ verbs used to express the *action* in a sentence
  - ◇ adjectives describe the *properties* of things

- Groupings supported by *distributional* evidence
  - ◇ words of same POS can appear in *similar contexts*,
  - ◇ tested by *substitution* (swap one for other in sentence)

    e.g. replacing *happy* with *clever* in: *He is a happy man*

    — produces a result that is a grammatical sentence

Syntax: parts of speech (contd)

- The parts of speech can be divided into two super-groupings:
  - ◇ Open-class: N, V, Adj, Adv
    - have many members
    - new ones added quite frequently
  - ◇ Closed-class: P, Det, Aux, Comp, Conj
    - a.k.a. *functional* categories
    - few in number, change little over time
    - have a clear grammatical use
- A *lexicon* (or *dictionary*) lists the words of a language
  - ◇ also specifies the POS for each word
  - ◇ words may have *more than one* POS

    e.g. *a crash* (N) vs. *to crash* (V)

Syntax: grammar

- The ways that words/phrases may be combined to form sentences may be described by a *grammar*
  - various different approaches to formulating grammars

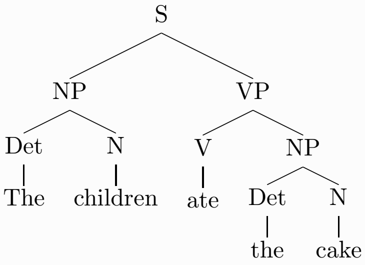
- *Phrase structure grammar* (a.k.a. *context-free* grammar)
  - gives *rewrite* rules to specify how phrases of different types constructed
  - called *phrase structure* rules, e.g.

NP	→	Det	N	(NP: Noun Phrase)
VP	→	V	NP	(VP: Verb Phrase)
S	→	NP	VP	(S: Sentence)

Syntax: grammar (contd)

- A grammar assigns a *hierarchical* structure to sentences
  - ◇ often presented in a tree-like format
  - ◇ called a *phrase structure tree*
  - ◇ is drawn *upside down*!

    i.e. with *root* at top, and *leaves* (words) at bottom



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 important for language / text processing
  - ◇ often encounter unfamiliar words
  - ◇ can use morphology to infer useful information

    e.g. of syntax (POS), and meaning

Morphology (contd)

- There are three major types of morphological processes
  - inflectional / derivational / compounding

- Inflectional morphology
  - *inflections* are sytematic modifications of a *root* by addition of *affixes* (*prefixes, suffixes*)
  - changes signal grammatical distinctions, e.g. plurality
  - inflection does not change the part of speech
  - inflection does not significantly change word meaning

e.g.	boy/boys	number	(singular/plural)
	bake/baked	tense	(present/past)
	go/goes	person	(1st/3rd)

  - inflectional variants grouped as variants of single *lexeme*

Morphology (contd)

- Derivational morphology
  - ◇ derivation creates *new* words by combining morphemes
  - ◇ commonly involves change to POS

    e.g. suffix *-en*: *dark* (Adj) → *darken* (V)

    e.g. suffix *-er*: *teach* (V) → *teacher* (N)
  - ◇ often involves significant change to meaning

    e.g. *wide* (Adj) vs. *widely* (Adv)
  - ◇ derivation is less systematic c.f. inflection
    - there are 'gaps' in what is produced
    - e.g. *quick* $\rightsquigarrow$ *quickly*, but

      fast $\not\rightsquigarrow$ *fastly*

Morphology (contd)

- Compounding
  - where two or more words merged to give a new ‘word’ or lexical unit
  - noun-noun compounds v.common in English

    e.g. *tea kettle, disk drive*
  - pronounced as a single word
  - but, often written as if separate words
  - denote a single semantic concept, deserving a separate entry in the lexicon