

Introduction to Natural Language Processing

This page is intentionally left blank

Introduction to Natural Language Processing

Aims	<ul style="list-style-type: none">• to review the structure of the discipline of linguistics, particularly as it applies to NLP• to review the structure of English grammar• to introduce the concept of a formal grammar rule
Reference	Allen, chapter 2.
Keywords	bound morpheme, parts of speech, descriptive grammar, free morpheme, lexeme, morpheme, morphology, phone, phoneme, phonetics, phonology, pragmatics, prescriptive grammar, speech act, string, syntax, wh-question, word, y/n question
Plan	<ul style="list-style-type: none">• overview of linguistics: lexicon, morphology, syntax, semantics, reference, pragmatics• parts of speech• phrases• grammar rules

Why study NLP (Natural Language Processing) ?

- NLP is part of AI
- In order to implement any application of AI, we need to know a fair amount, and maybe a lot about, the *domain* or application area.
So AI = general AI techniques + domain knowledge + domain specific techniques
We study NLP as an example of this integration process.

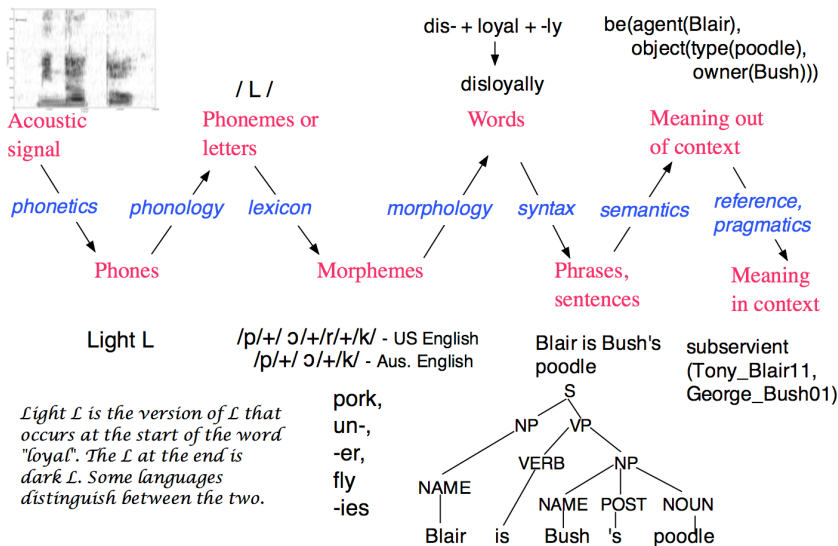
Typical Applications of NLP

- database query languages
- machine (assisted) translation: e.g. weather reports, Canadian Hansard
- information extraction: e.g. financial reports, news article summaries
- all of these operate in limited domains

Related Disciplines

- Linguistics - study of language and of languages
- Psycholinguistics - language and the mind, models of human language processing
- Neurolinguistics - neural-level models of language processing
- Logic - an unambiguous formal language useful for representing (unambiguous) meanings

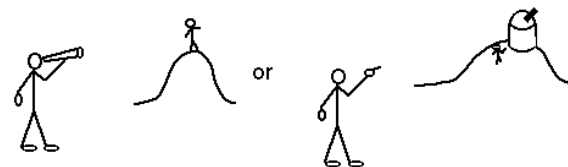
Map of the Discipline of Linguistics



Major Issue - Ambiguity

- Structural / Syntactic ambiguity
- Word-sense ambiguity
- Referential ambiguity
- Combinations of these

Structural/Syntactic Ambiguity



Word-Sense Ambiguity

dog:

1. domesticated carnivore *Canis familiaris*
2. animals related to 1.
3. the male of such an animal (as opposed to *bitch*)
4. various unrelated but vaguely similar animals (e.g. *prairie dog*)
5. a despicable fellow
6. a fellow in general (*you old dog, you*)
7. a mechanical device for gripping or holding (*dog the hatches*)
8. an andiron

and more.

Referential Ambiguity

- *Steve hated Paul. He hit him.*
- *he = Steve or he = Paul ?*

Combinations of These

- *The red eyes water*

Eliza/Doctor (1966)

Men are all alike.
In what way?

They're always bugging us about something.
Can you think of a specific example?

Well, my boyfriend made me come here.
Your boyfriend made you come here?

He says I'm depressed much of the time.
I am sorry to hear you are depressed.

It's true - I am unhappy.
Do you think coming here will help you not to be unhappy?

I need some help.

...

Earlier you said you were unhappy.

Syntax – Parts of Speech

- Words can be classified to parts of speech (POS) – practical sets for English have about 40 – here are some common ones:

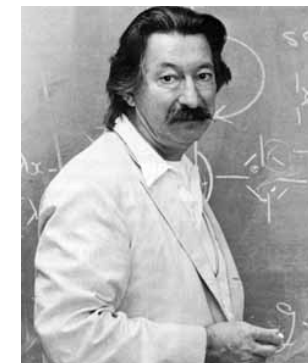
Part of Speech	Examples
NOUN	cat thing philosophy bite
VERB	assassinate graduate have bite do
ADJECTIVE	yellow happy teenage bad
ADVERB	happily well very
DETERMINER	the a an these
INTERJECTION	hallelujah! oh! arggh!
PREPOSITION	of in under at beside
CONJUNCTION	and or if but
PRONOUN	he him she it I we us they them you thou thee ye
AUXILIARY	has have had am be is was do does did can could

- Parts of speech are sometimes referred to as *lexical categories*.

Eliza 2

- Was an early rule-based / pattern-matching system
- Matched patterns (I like \$Object) in the input, discarded inessential bits, turned what was left around (I like pizza → you like pizza), and added an "attitude" (It's interesting that you like pizza).
- Has a memory, so material can be re-introduced some time later.
- You can try out a version of Eliza for yourself by running the Gnu emacs editor on one of the School's Linux computers (login and type emacs) and then typing `ESCAPE x doctor` (Type control-X control-C to exit from emacs.)
- Eliza was created by Joseph Weizenbaum of MIT.

Joseph Weizenbaum



Subclassification

- Some parts of speech can be subclassified (and the subclasses have different grammatical properties)

proper nouns	abstract nouns	mass nouns	count nouns
<i>Iraq</i>	<i>philosophy</i>	<i>sand</i>	<i>apple</i>

intransitive	transitive	ditransitive
<i>laugh</i>	<i>learn</i>	<i>give</i>
<i>You laugh _</i>	<i>You learn <u>history</u></i>	<i>You give <u>him</u> a book</i>

Summary

- linguistics forms the domain knowledge for natural language processing
- ambiguity is a major issue in NLP
- words must be classified (parts of speech, and beyond) as a basis for NLP
- phrase structures are described by grammar rules
- lexical and phrasal categories appear in grammar rules