COMP5046: Linguistic Fundamentals

Joel Nothman

joel.nothman@sydney.edu.au

School of Information Technologies
University of Sydney

2018-03-27

Part I

The structure of language

Linguistics: the science of language

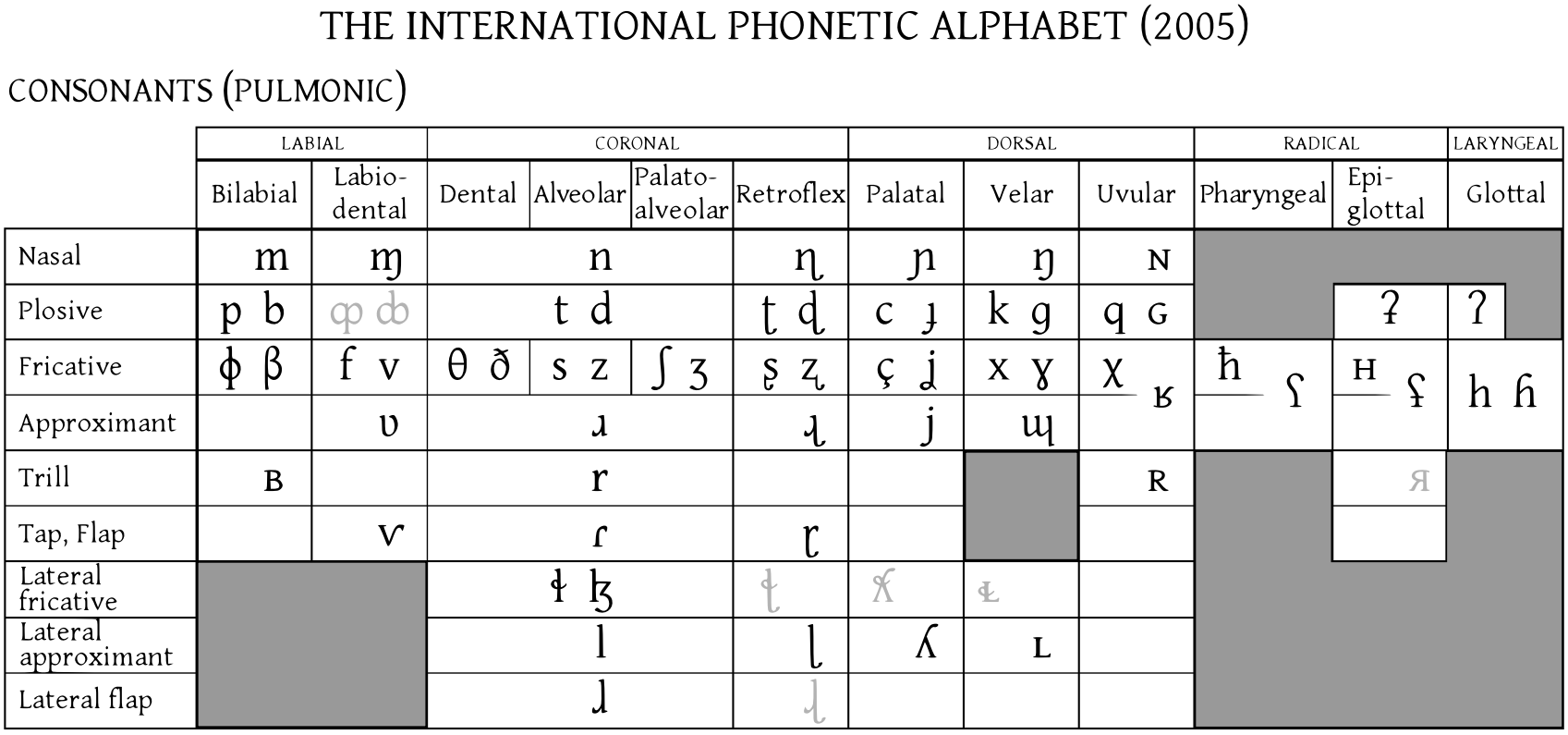
- How does language work?
- How can we analyse language?
- What qualifies as evidence?

- How can linguistics inform NLP?
- How can NLP inform linguistics?
- Why is NLP hard?

Linguistics models various components of language

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics and discourse

Phonetics: the mechanics of language sounds



THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N			
Plosive	p b	pˠ bˠ		t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ʀ̠	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		ɭ̆						

- articulation
- acoustics
- audition

Phonology: the structure of language sounds

- show that [d] and [t] are distinct in English.

Phonology: the structure of language sounds

- show that [d] and [t] are distinct in English.
- **substitution tests** and the distributional properties of sounds
- in contrast, [t^h], [t], [ɾ], [ʔ] are allophones of /t/ in English.
  butter; stow vs. tow
- valid sequences: stop vs ftop
- syllable structure, stress, intonation, . . .

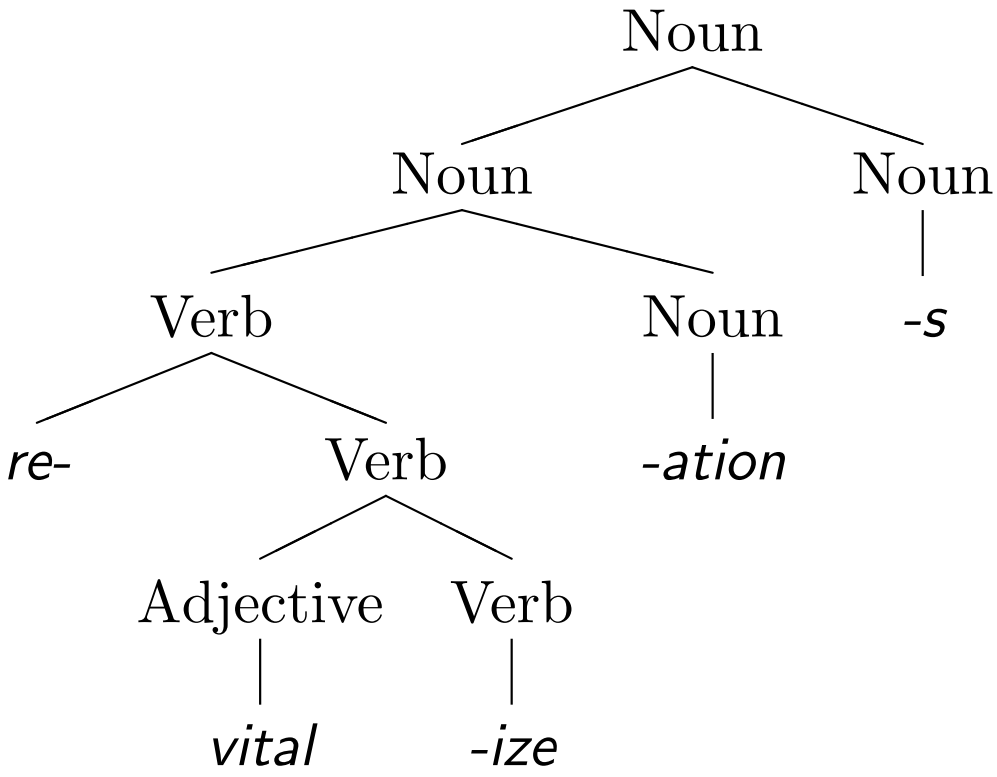
Morphophonology: Word structure can affect sound

- in-
- irrelevant
- improbable
- incapable

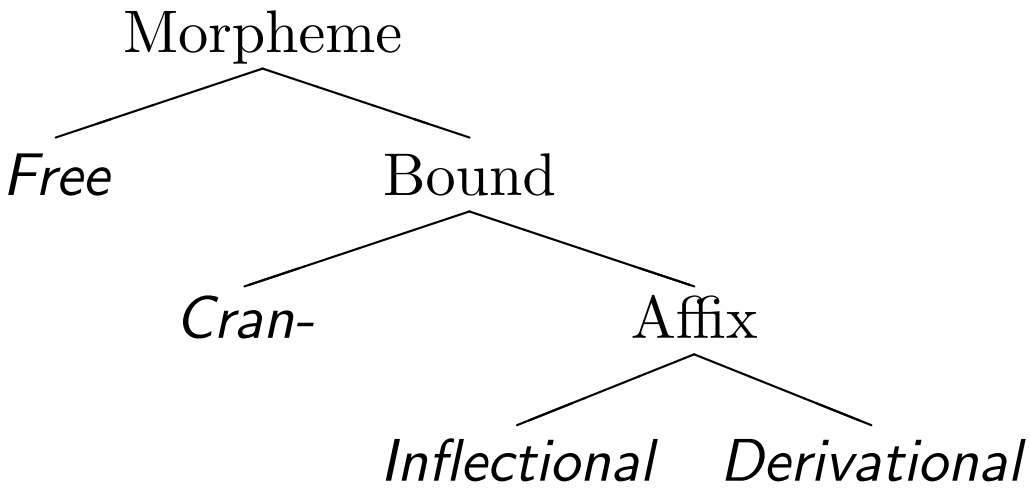
- the layers of linguistic analysis often blur

Morphology: the structure of words

a **morpheme** is the smallest discrete piece of meaning



Morphological typology



Inflection rarely affects meaning or syntactic class

Person/Number	Inflected form	Gloss
1st s.	am**o**	I love
2nd s.	ama**s**	you (sing.) love
3rd s.	amat	he/she/it loves
1st pl.	ama**mus**	we love
2nd pl.	ama**tis**	you (pl.) love
3rd pl.	ama**nt**	they love

Table: Present (tense) indicative (mood) active (voice) of amare, to love.

Full conjugation of amare
Finnish: morphology to the EXTREME!

Take-home messages about inflectional morphology

- English, like Chinese, is a low-inflection outlier
- some inflection is through prefix/suffix/infix
- most inflectional morphology is non-concatenative: the pieces fuse or the stem is altered, e.g. swim/swam
  - makes parsing/generating hard
  - even makes manual classification tricky
- a single affix encodes multiple pieces of information
  - the -o in amo specifies number, tense, mood and voice
- affixes are very frequently overloaded/ambiguous
  - context + statistics needed for morphological analysis

Derivational morphemes can change the part of speech

- noun to adjective: cheek*y*, Italian*ate*, joy*ous*, pigg*ish*
- noun to verb: chlorin*ate*, zero morpheme
- verb to noun: reflect*ion*, imprison*ment*, *'*subject
- verb to adjective: accept*able*, prohibit*ive*

Derivational morphemes might not change part of speech

- *in*sufferable
- *un*wind
- *dis*like
- *a*moral
- *co*-author
- *over*write
- cyclo*phobia*
- fancy *shmancy*

Derivational morphology is semi-productive

- not all combinations are automatically words
  sneaky vs. *discussy; action vs. *waition; agreeable vs. ?sleepable
- not always compositional: react
- some morphemes are no longer very productive: be- of become, between;
  although A sulky adolescent and his be-jeaned girl-friend. (Manchester Guardian, 1958)

Cranpheme (cranberry morpheme) problem

- a **free** morpheme can occur on its own as a word
- cran- morphemes don't have a corresponding independent word but can't be bound freely
- cran- morphemes are often **fossilised** forms
  - cran- in cranberry (from Middle English cran for crane)
  - ten- in tenable (from French tenir for hold/keep)
  - cob- in cobweb (from Middle English coppe for spider)

- berry, able and web here are clearly morphemes
- -nov- in innovate: cran- or free?

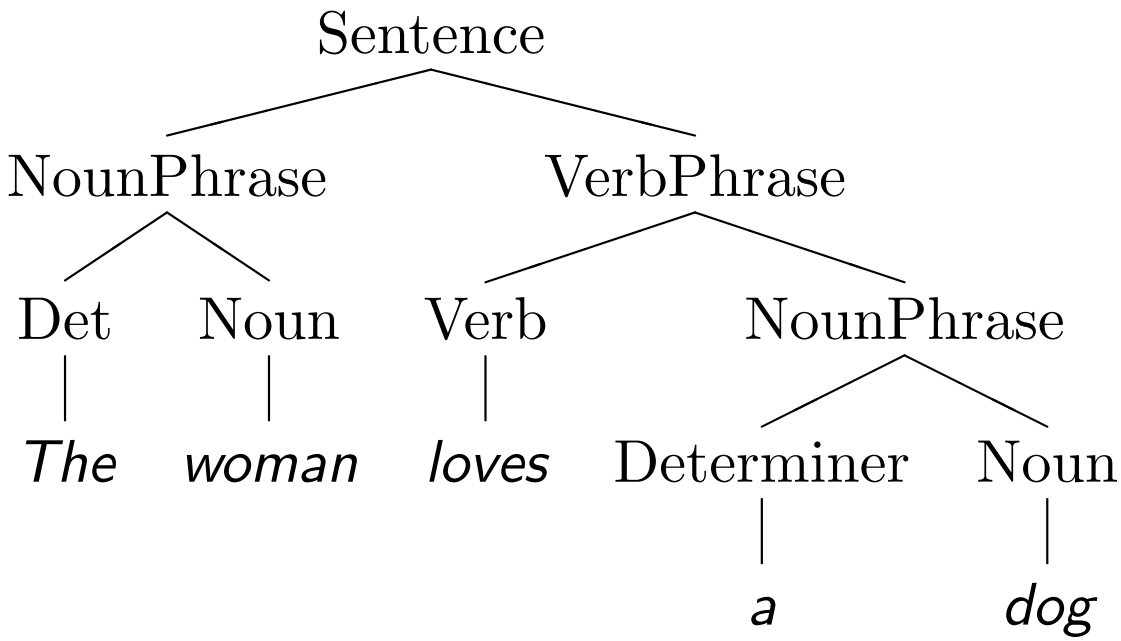
Open vs. closed-class morphemes

Open Class Lexemes

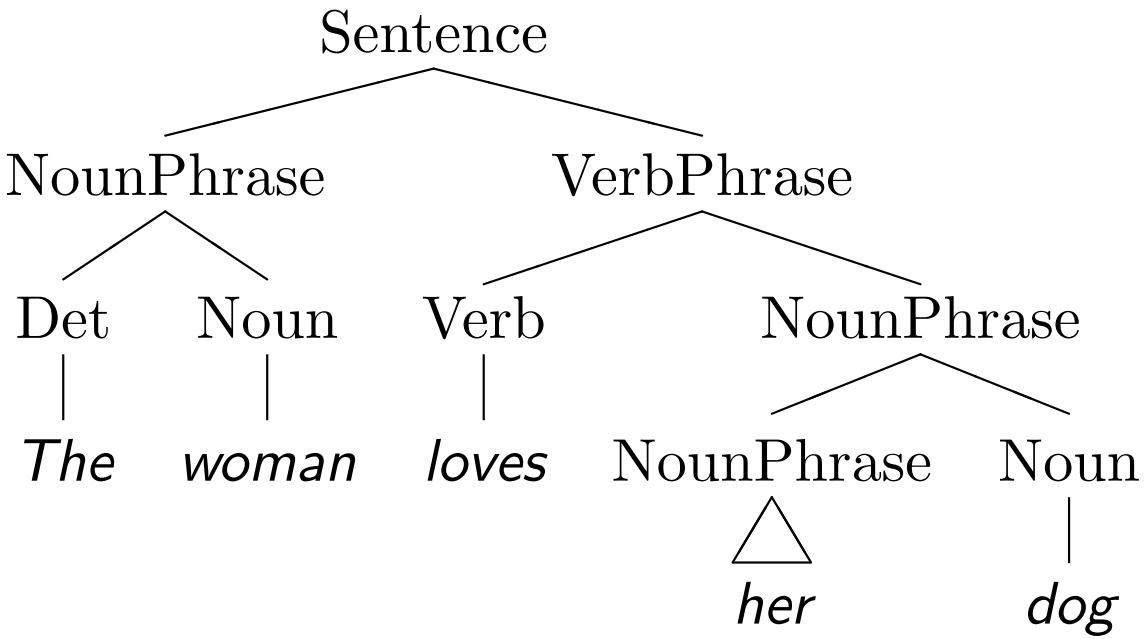
Open-class morphemes determine the content of a sentence, while closed-class morphemes and function words help determine sentence structure. Open-class morphemes are part of the lexicon, while closed-class items are part of the grammar.

- open: open, class, morpheme, determine, content, sentence, closed, class, funct(ion?), word, help, structure, part, lexicon, grammar
- closed: -s, the, of, a, are, while, and

Syntax: how words combine into phrases and sentences



Syntax: how words combine into phrases and sentences



substitutability is key to constituency

also note movement, insertion, . . .

Part of speech / syntactic class / parse nonterminal

- is a verb a *doing word*? be? think? destroy vs. destruction?
- are nouns *things*? intelligence? destruction? correctness?

Part of speech / syntactic class / parse nonterminal

- is a verb a *doing word*? be? think? destroy vs. destruction?
- are nouns *things*? intelligence? destruction? correctness?
- **test** for adjectives: that seems *X*
  - substitute and **judge grammaticality**
  - that seems correct vs. *that seems correctness
  - how about the *X* thing?
- when substitution needs context:
  - Could you set the table?
  - I have the complete set.
- syntactic subtlety means some substitutions won't work
  - cutlery is a noun, but *a cutlery
  - sleep is a verb, but *tigers sleep other tigers

Let's tag some parts-of-speech

(1) *The Pied Currawong is a medium-sized black passerine*
– – – – Adj. Adj. Adj.

bird native to eastern Australia and Lord Howe Island
Noun Adj. – Adj. – – – –

(2) *The male and female are similar in appearance*
– ?? – ?? – ?? – ??

(3) *Known for its melodious calls, the species ' name*
?? – – ?? ?? ?? – ?? –

currawong is of indigenous origin
?? – – ?? ??

Generative grammar

- **generative grammar** proposes to identify a procedure which
  - generates all valid sentences in a language
  - generates only valid sentences in a language

- avoiding both over- and under-generation is tricky. . .
- She loves her dog, She loves herself, *She loves herself's dog
- $\langle NP \longrightarrow DT\ JJ \rangle$ such as The elderly

Generative grammar

- **generative grammar** proposes to identify a procedure which
  - generates all valid sentences in a language
  - generates only valid sentences in a language

- avoiding both over- and under-generation is tricky. . .
- She loves her dog, She loves herself, *She loves herself's dog
- $\langle NP \longrightarrow DT\ JJ \rangle$ such as The elderly
  . . . but not an elderly :(

- in practice, all formal grammars are at least a little leaky
- and NP (noun phrase) remains a useful abstraction

Grammaticality and acceptability

- **autonomy of syntax**: grammatical $\neq$ meaningful
  - grammatical sense: harmless young children sleep peacefully
  - grammatical nonsense: ?colourless green ideas sleep furiously
  - ungrammatical: *peacefully children young sleep harmless

- **acceptability** is gradient. Is grammaticality?
- acceptable but ungrammatical:
  But if this ever changing world **in** which we live **in**
  Makes you give in and cry
  Say live and let die. (Paul McCartney, 1973)
- linguistic **competence** vs **performance**

Ambiguity in grammar

- natural language grammars are massively ambiguous
  - I saw the girl on the hill with the telescope
  - I saw the girl and the boy on the hill with the telescope
- PP-attachment and coordination are the major sources of ambiguity
- ambiguity tends to grow exponentially with sentence length
- not all ambiguities are meaningful
  - I saw the lions at the circus

Syntax and free word order languages

- morphologically rich languages specify much of syntactic structure in inflectional suffixes

(4) *puer* *amat* *puellam*
boy (doer) love (3s) girl (doee)
'The boy loves the girl'

(5) *puellam* *amat* *puer*
girl (doee) love (3s) boy (doer)
'The boy loves the girl' or 'It is the girl that the boy loves'

- popular phrase structure grammars fit the English-centric assumption that word order is key
- some formalisms are better for some languages, perhaps like programming languages

Abstracting semantics from syntax

- Different parses, common proposition:
  - Google bought YouTube.
  - YouTube was bought (by Google).
  - It was YouTube that Google bought.
  - Google's purchase (of YouTube) was ...
  - Did(n't) Google buy YouTube?
- require a single representation: buy(Google, YouTube)
- **semantic roles**: who did what to whom?
- semantic value of a sentence is a proposition which is true or false in some state of the world

Semantics is not just a derivative of syntax

- one of these is syntactically ambiguous
  - Kim fed her dog biscuits
  - Everyone knows one language

- lexical semantics vs compositional semantics
  - Google cancelled the purchase of X
  - Google did not purchase X
- presupposition

Why so many ways to say the same thing?

- listeners don't just need the message, they also need to understand its significance
- we signal how new content relates to the prior discourse
- natural information structure:

(6) When did Mary graduate?
She graduated in 2006.

(7) Did Mary enrol in 2006?
In 2006 she graduated.

- less natural information structure:

(8) When did Mary graduate?
In 2006 she graduated.

(9) Did Mary enrol in 2006?
She graduated in 2006.

Pragmatics: effective communication in context

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

Pragmatics: effective communication in context

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

- Grice's **cooperative principle**: make your contribution helpful given the purpose(s) of the conversation
- maxims: quality, quantity, relevance, manner
- **violation is ostensive**
- basis of modelling: humour, irony, metaphor, politeness...

Discourse structure: how sentences/clauses are related

- **Time.** I finished my thought and then went for lunch.
- **Cause.** I was hungry. I decided to get some lunch.
- **Concession.** I wasn't hungry. I only agreed to get some lunch because Robin was famished.
- **Purpose.** I went to the store so that I could get some lunch
- **Elaboration.** I was very hungry. I was simply famished.

et cetera

Reference and ambiguity

- Please shut the door.
  The house was empty. The door was broken.
- Kim thinks that he is clever.

Summary: The structure of language

- languages have rules/preferences defining their phonological, morphological, syntactic and pragmatic structure
- grammar = morphology + syntax
- meaning derives from lexicon, grammar and context
- alternatively: given context, the lexicon and grammar encode meaning as structured messages

- blurring at all boundaries
- ambiguity and change at all layers
- number of possible messages is infinite

Part II

The linguistic discipline and NLP

Evidence in linguistics

- **Acceptability judgements**: Intuitions of a native speaker.
- **Corpus linguistics**: Is the appearance or non-appearance of some phenomenon proof? What about errors and their correction?
- **Comparative linguistics**: Are there universals across languages? What distinguishes human language?
- **Language acquisition**: How do children gain linguistic knowledge? How does this compare to second language learners?
- **Psycholinguistics**: How do our minds process language? E.g. *garden path sentences*: The horse raced past the barn fell. Time flies like an arrow; Fruit flies like a banana. Reaction times.
- **Neurolinguistics**: How does this work physically inside our brains? What happens with brain damage?

Language variation

- Languages are very different from one another. Most have very little record. Field linguistics: documenting a language.
- Historical (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: Variation with social context.
  Dialect: have a shower vs. take a shower; /t/ in butter.
  Labov's *fourth floor* experiment. Code switching.
- Variation across genre

- how do humans deal with variation?
- how does variation affect NLP?

What is a language?

a language is a dialect with an army and a navy.

an ensemble of idiolects ...
rather than an entity per se.

a language as an ideal system
outside the practice of language users.

Linguistic insight in NLP

> Every time I fire a linguist, the performance of the speech recognizer goes up (Fred Jelinek)

Linguistic insight in NLP

> Every time I fire a linguist, the performance of the speech recognizer goes up (Fred Jelinek)

But remains an open question. Relates to questions of innateness.

NLP as evidence for linguistic theories

Can predictive modelling
tell us how language works?

Linguistics: the science of language

- How does language work?
- How can we analyse language?
- What qualifies as evidence?

- How can linguistics inform NLP?
- How can NLP inform linguistics?
- Why is NLP hard?