Linguistic terminology

Peter Petré

Contact details

- Peter Petré
- Office: D.216
- peter.petre@uantwerpen.be
- Office hours
  - right after class
  - make an appointment by email (also for online consult)
- I'm responsible for
  - the *Corpus Studies* course module
  - internships (together with Walter Daelemans) – there'll be a Q&A 8/10

Overview

- Basic terminology: not so harmless
- Linguistic theory and digital text analysis are intertwined
- Modularity of language?
- Syntax and symbolic NLP
- Statistical approaches: where’s linguistic theory?
- Neural approaches: why would we still need linguistic theory?

Some basic terminology, part I

- Grammar =
  - Phonology/phonetics
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics
- ???Spelling
  - Only marginal part of grammar
  - More of a nuisance than anything else
  - *Can* be meaningful though

Phonology versus phonetics

- Phonetics studies how we make and perceive sounds
- Phonology is one level of abstraction higher

keep	/kip/	The place of articulation is fronter in the mouth	[k_{+}^{h}]
cart	/kt/	The place of articulation is not so front in the mouth	[k^{h}]
coot	/kut/	The place of articulation is backer, and the lips are rounded	[k^{hw}]
see**k**	/sik/	There is less aspiration than in initial position	[k`]
s**c**oop	/skup/	There is no aspiration after /s/	[k]

- Both make use of the International Phonetic Alphabet https://www.internationalphoneticassociation.org/

Information lost in written text

- Phonetic
  - Emphasis, prosody
- Phonological
  - Stress patterns differ between verbs & nouns
    - pro*test* vs. *pro*test etc.
  - Stress patterns differ between compounds and Adj + N combinations
    - *ˈblackbird, ˈblackboard, White House*
    - *ˈblack ˈbird, ˈblack ˈboard, white house*

Morphology

- Study of word formation
- What is a word?

Word segmentation

- How many words are there in the following sentence?

Our taxi driver took us to the sea side in a

bullet-proof car

Word segmentation

- How many words are there in the following sentence?

Our taxi driver took us to the sea side in a

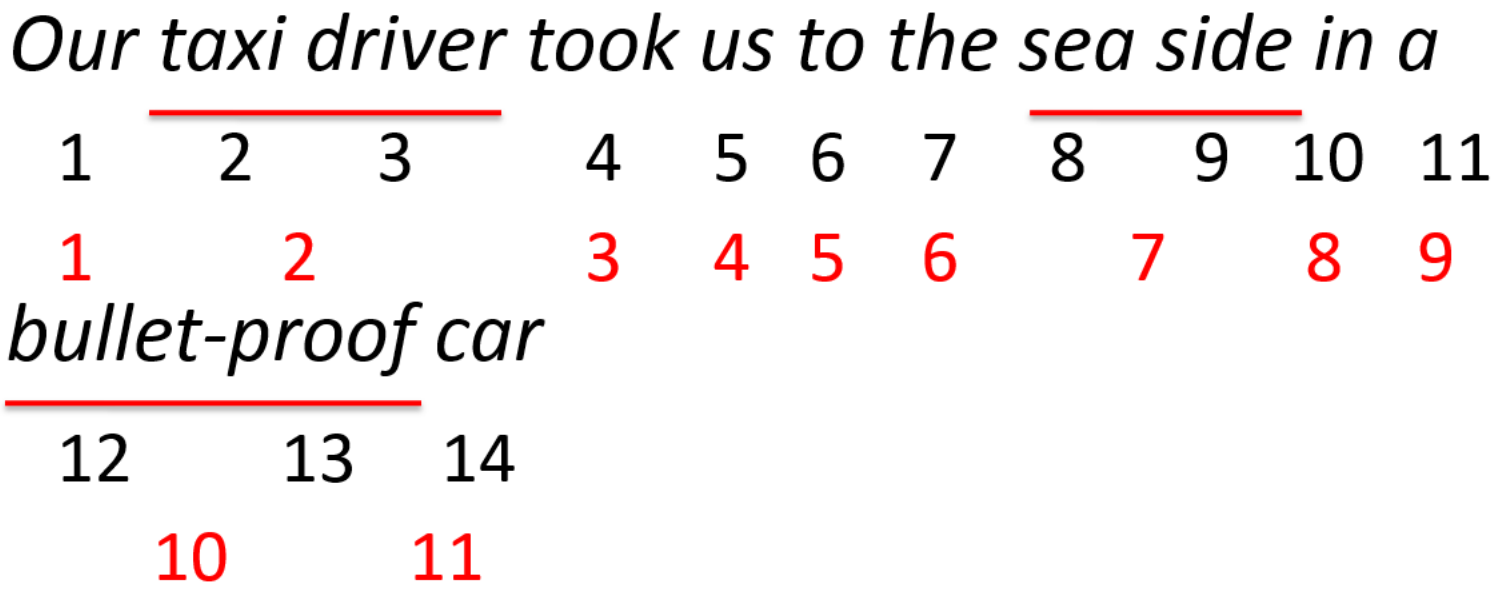
1 2 3 4 5 6 7 8 9 10 11

bullet-proof car

12 13 14

Word segmentation

- How many words are there in the following sentence?



Word segmentation

- How do you define a word?

  Deceivingly simple question; answer not so straightforward

- 4 definitions
  1. orthographical definition
  2. integrity definition
  3. semantic definition
  4. syntactic definition

1. Word – Orthographic definition

- **Any string of letters separated by a space**
- Problems:
  - Compounds and the like:
    - *taxi driver*; *sea side*; *bullet-proof*; *top rack dishwasher safe*: can be regarded as 1 word or as 2 / 5
    - *do-it-yourself*
    - *we'll*, *isn't; kinda* (= *kind of*), *wanna* (=*want to*)
  - Some languages (e.g. Chinese) don’t use spaces
  - Unwritten languages have no words then? Writing is **not intrinsic** to language

2. Word – Integrity definition

- **Any string of phonemes whose integrity cannot be broken**

  happy => *un+happy*, *happy+ness*

  **hap+un+py*, **hap+ness+py*

  (note: * means ungrammatical)

- Problems
  - Some languages allow insertions (infixes)
  - Some languages have two-part words (*Ich komme gerade an*)
  - Occasional (expletive) infixation even in English

    abso-fucking-lutely, *fan-fucking-tastic*,

    kanga-bloody-roo, etc.

=> definition (cross-linguistically) not fully reliable

3. Word – Semantic definition

- Expresses one single idea or concept

  love, *happy*, *banana*, *university, taxi driver, sea side*

  Problems:
  - *love*, *university*: complex concepts
  - some 'single' ideas or concepts do not have words to express them, e.g., "area between nose and lip", "smell of freshly ground coffee"
  - vagueness & cross-linguistic differences, e.g., kinship terms:
    - *Cousin(e) – Neveu/Nièce*
    - English *cousin*: male/female
    - Dutch *neef* = both *cousin* & *neef*

=> semantic definition not so reliable

4. Word – Syntactic definition

- The smallest syntactic building blocks of sentences (= external boundary)

  Examples

  - *Look at that [X]* => X is a word that belongs to the syntactic category of **nouns**, e.g. *car*, *toy*, etc.
  - *Look at that [X] thing* => X is a word that belongs to the syntactic category of **adjectives**, e.g. *lovely*, *shiny*, *big*, *unfashionable*

$\Rightarrow$ fairly robust definition (despite some problems)

Summary

- Words are units that
  - form the syntactic building blocks of sentences
  - mostly express a single concept or idea
  - usually are indivisible

A word in Digital Text Analysis?

- Does it matter?
- In digital text analysis the distinction is made between *token* and *type*
- A *token* is an instance of a word
  - Running text consists of tokens
  - Your definition or conceptualization of what a ‘word’ is will have an impact on how you ‘tokenize’ your data (split up the string of characters into tokens)
- A *type* is the abstract representation of a word
  - *go, goes, going, gone, went, gonna* = 1 type GO / 6 tokens in a text

Kinds of words

- Common to distinguish different kinds of words:
  - content vs. function words
  - simplex vs. complex words

Content words and function words

- **Content words**: nouns, verbs, adjectives, (most) adverbs
  - denote objects, actions, attributes, ideas
    e.g. ***child***, ***anarchism***, ***sour***, ***purple***, ***run***, ***liberty***
  - **open** class : new words regularly added
    e.g. ***download, byte, email, podcast, obamania***

Content words and function words

- **Function words**: conjunctions (***and***, ***or***, …), prepositions (***in***, ***of***, …), articles (***a***, ***the***), pronouns (***it***, ***he***, …), …
  - No clear lexical meaning (but still meaningful)
  - No obvious concepts associated with them
  - **Closed** class (e.g., ***per*** as unsuccessful attempt at creating a gender neutral pronoun to replace *he*/*she*)
  - Typically have a grammatical role

Also called ***lexical*** vs ***grammatical*** words

Simplex vs. complex words

- simplex words: consist of a single part, e.g.,
  father, *child*, *kill*, etc.
- complex words: consist of different parts, e.g.,
  father-s, *child-hood (neighbour-hood, brother-hood)*, *kill-ing*, etc.

=> strong regularities in how words are built up

- recurrent patterns
- rule-governed

= studied in **morphology**

Recurrent patterns

- *bullet(-)proof, explosion(-)proof*
- *water(-)proof, fire(-)proof*
- *scratch(-)proof, splash(-)proof*
- *baby(-)proof, mother-in-law-proof, rabbit(-)proof, student-proof*

=> **general pattern: *X-proof***:

- 2 words "fused" into 1 (= compouding)
- meaning: ± "not destroyed/affected by X"

Morphology and digital text analysis

- Many vanilla implementations are focused on *tokens*
- Because of this focus they miss out on structure

• desirable	**un**desirable
• likely	**un**likely
• inspired	**un**inspired
• happy	**un**happy
• developed	**un**developed
• sophisticated	**un**sophisticated
ADJECTIVE	Form: *un*-ADJECTIVE
	Meaning: "NOT-ADJECTIVE"
	! ***un*- is NOT a word/token**

Syntax

- Rules of sentence-formation
- Chomsky's 1957 *Syntactic structures*
  - promoted the study of 'generative syntax'
  - syntax is seen as a phenomenon that could and should be studied in isolation from other 'linguistic modules' (phonology, semantics, ...)
- Caused great advances to be made in linguistic analysis
- At the same time problematic on *many* levels

Generative syntax

- **Generate all and only the grammatical sentences of the language**
- Assign an appropriate syntactic structure to the sentences concerned which accounts for the native speaker's intuitions about the structural relations between the words in a sentence

Generative syntax

(2) (a) I gave back the car to him
(b) I gave the car back to him
(c) I gave him back the car
(d) I gave him the car back
(2) (e) *I gave the car to him back
(f) *I gave back him the car

Generative syntax

- Generate all and only the grammatical sentences of the language
- **Assign an appropriate syntactic structure to the sentences concerned which accounts for the native speaker's intuitions about the structural relations between the words in a sentence**

Representing underlying structure

- Only linear structure would be

  (3) Thisboywillspeakveryslowlytothatgirl

- Spelling identifies word boundaries
- Other levels of structure remain unrepresented
- Native speaker recognizes that sound-sequences are organized in successively larger groups which we call *constituents*
  - Constituents = theoretical constructs
  - This assumption is not quite borne out in psycholinguistic research
- Syntactic analysis may provide a way of representing this structure

Constituency tests: coordination

- Coordinating conjunctions (*and, or, yet, but*)
- Join constituents that are at the same structural level

  I met *John* and *Mary*

  *John rang *up his mother* and *up his sister*

  He lived *in New York* and *in Belgium*

  *I knocked her *up* and *on the door*

Constituency tests: intrusion

Clearly, this guy has been drinking again.

**This clearly guy has been drinking again.*

- With contrastive focus

  ?The cat will eat, almost certainly, his dinner
  (= *and not yours/and not his blanket*)

$\Rightarrow$Doesn't have scope over the entire sentence anymore
(Cf. different paraphrase of *What the cat ... is ...*)

Constituency tests: anaphora by proforms

- pronouns = pro-Noun-Phrases

  (89)*what do you think of the man next door*

  I like him a lot. He's very amiable

  (90)*I like the woman in the blue hat*

  I like her

  **I like the her in the blue hat*

- Proforms presuppose constituents

Constituents at different levels

- Smallest (syntactically relevant) level = word/token/lexical item
- Can be grouped into ‘lexical categories’
  - Noun (common noun, proper noun) N
  - Verb (transitive verbs, intransitive verbs) V
  - Preposition P
  - Adjective A
  - Determiner Det
  - ?Adverb Adv
- Based on their *distributional* properties

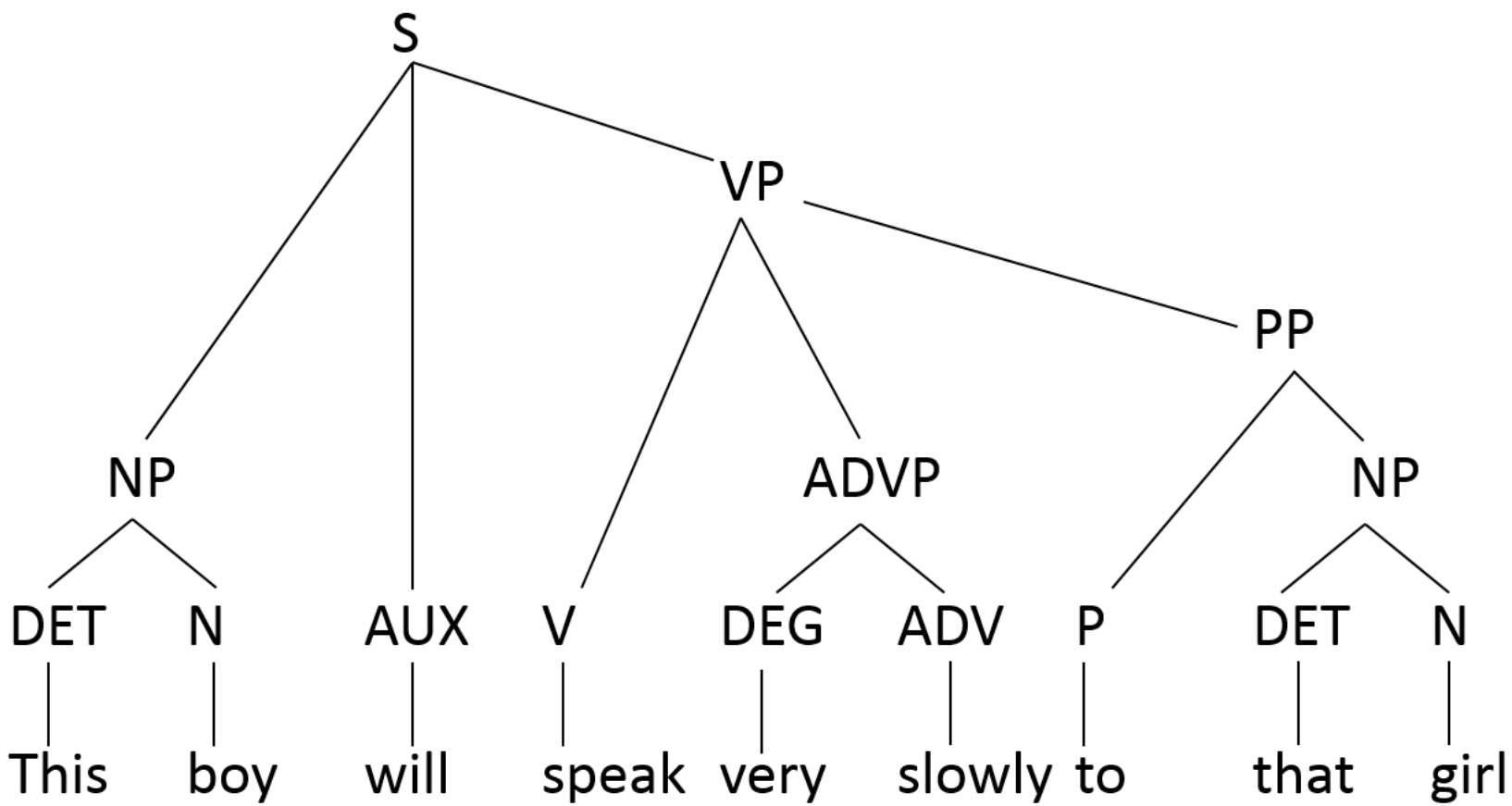
  (1) a. Jack devoured the doughnut.
  b. *Jack slept the doughnut.

  (2) a. *Jack devoured.
  b. Jack slept.
- Theory of constituency is fraught with difficulty
  - Discontinuity: cf. German example *Ich komme gerade an* ‘I’ll be right with you’

From lexical categories to phrases

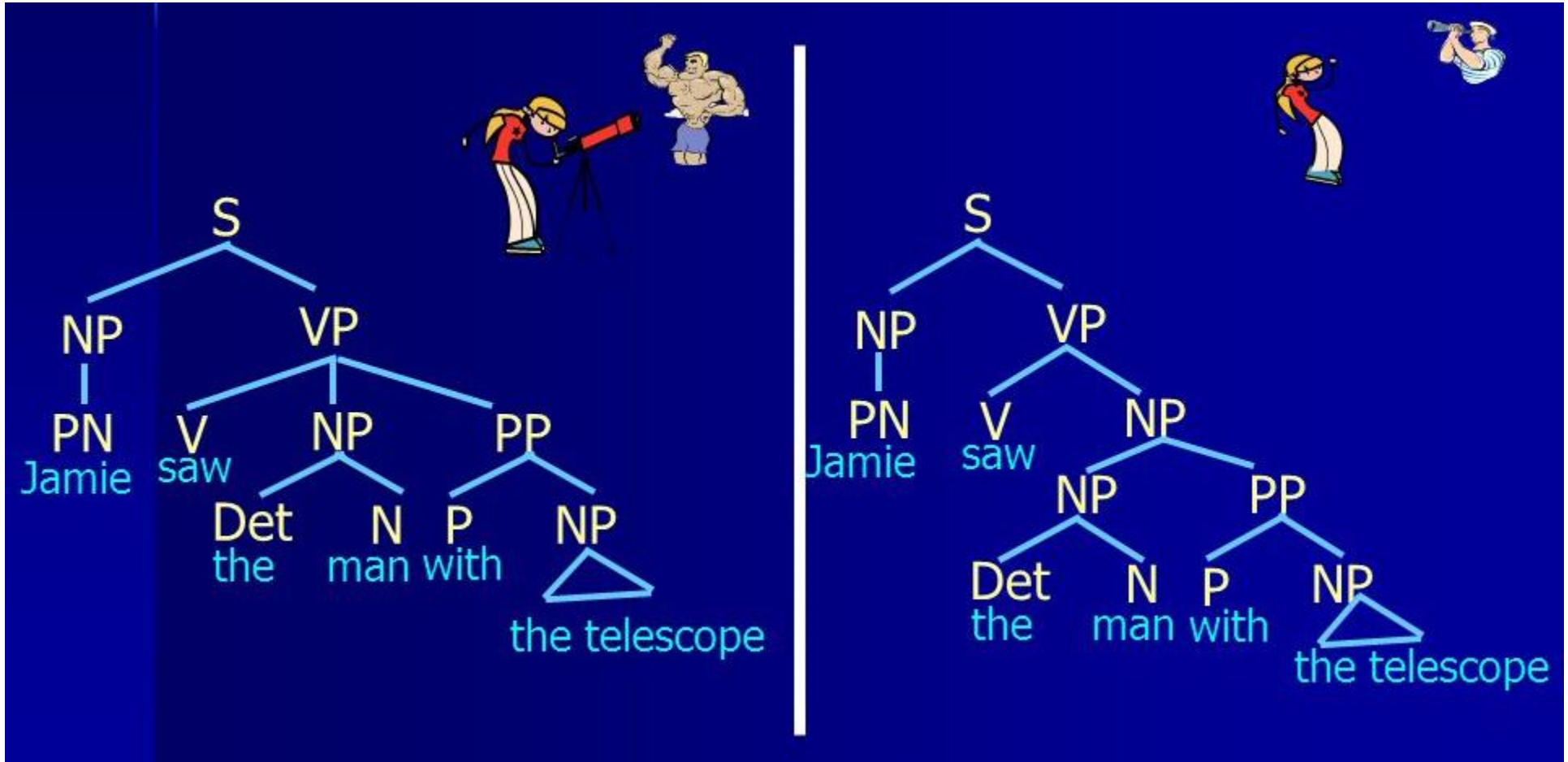
- (37) S → ***DET – A – N*** – AUX – V – ***DET – A – N*** – P – ***DET – A – N***
- *DET – A – N* is a recurrent set, so it makes sense to make the following generalization:
  - they are structurally connected into an NP
  - (38) a. S → NP – AUX – V – NP – P – NP
    b. NP → DET – A – N
    c. PP → P – NP
    d. VP → V – NP – PP

Phrase-marker / tree-diagram



Resolving ambiguities

- Jamie saw the man with the telescope

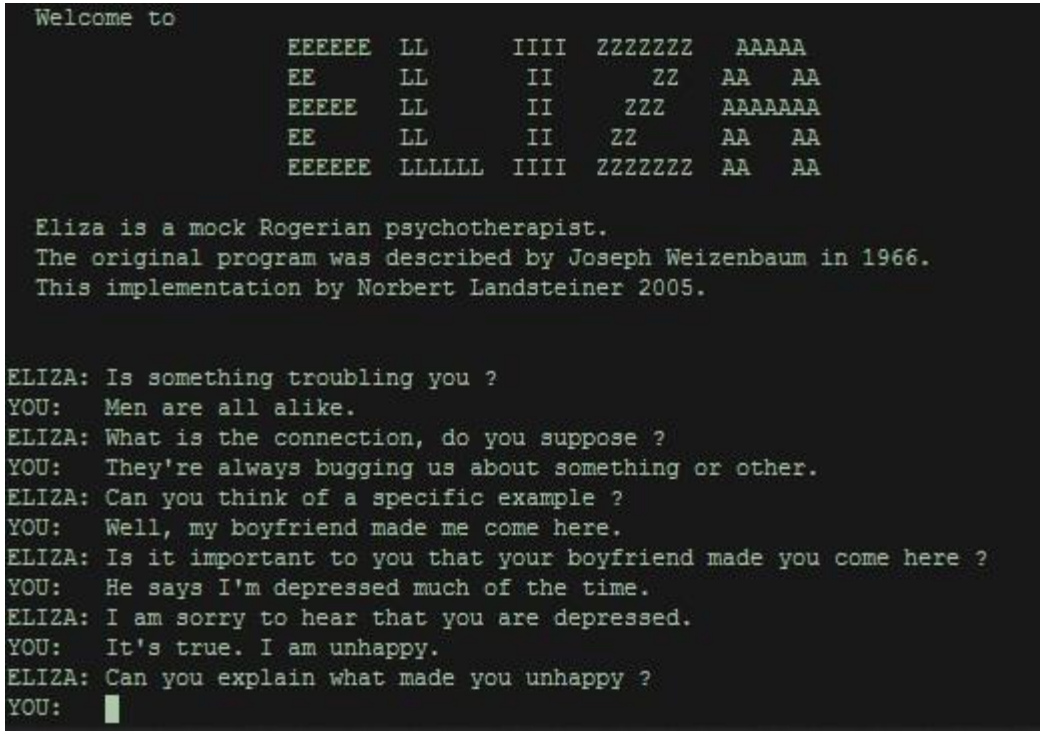


Basic linguistic terminology: pro's

- Heuristic aid for organizing one's thoughts about language
- E.g. if you want to do a case study on possessive *have*
  - *I had an ox*
  - *I had an ox plow my field*
- Impact on programming
  - In coding we also speak of syntax and semantics
- Fair enough

First wave of NLP

- Based on symbolic rules such as those formulated in Generative Grammar
- Some early implementations
  - Automatic translation
  - ELIZA
    one of the first fit for the Turing test

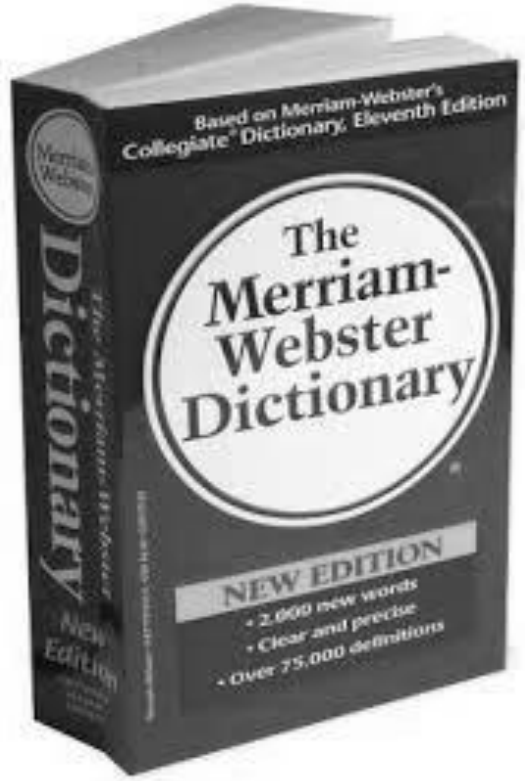


Basic linguistic terminology: Not so harmless

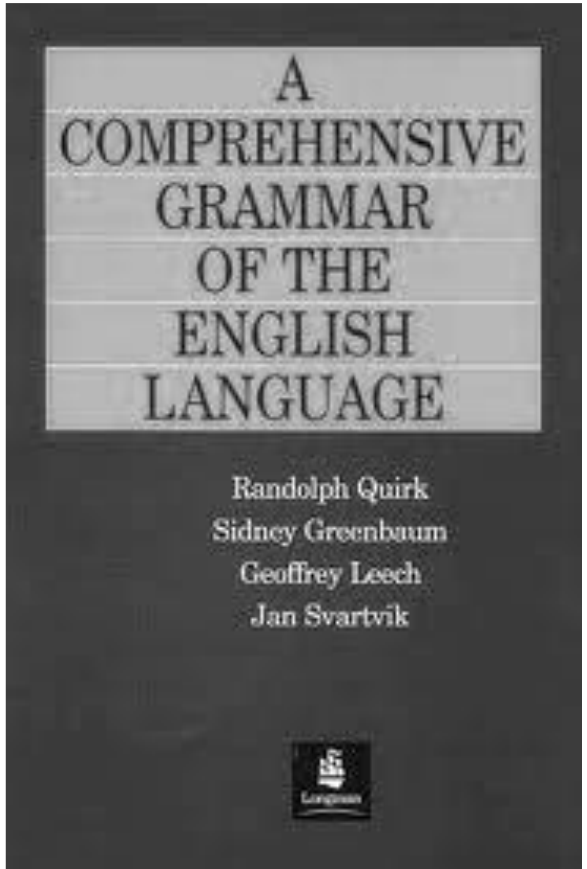
- Imbued with outdated (?) assumptions
  - Symbolic NLP implementations quickly reached a ceiling
    - Led to funding for machine translation being heavily reduced in the late 1960s
    - Why?
  - Phonology, morphology, syntax, semantics = different **modules** of language
  - Is hierarchical syntax real?

Basic linguistic terminology: Not so harmless

- Imbued with outdated (?) assumptions
  - Phonology, morphology, lexicon, syntax, semantics = different **modules** of language



≠



Basic linguistic terminology: Not so harmless

- They are rather like interconnected areas with overlap
  - phoneme > morpheme (typically a case of ‘exaptation’)
    - 17th century *my **n**ose* vs. *mine **ea**r* > *my book* vs. *this book is mine*
  - morpheme > word
    - *-ism (racism, sexism) > ism* (‘an ideology’)
  - word > morpheme
    - *will* (I will) > *I’ll*
  - morpheme > syntax
    - *I’m going to > I’m gonna > I Ø gonna*
  - lexical categories
    - nouns may be used as adjectives or verbs: *4 Ways to Mother-In-Law Proof Your Kids*

Not so harmless (cont'd)

- Syntax is also *not* meaningless
- This was an argument by Chomsky in the 50s
  - Colorless green ideas sleep furiously
- But: Nonsensical sentences are a by-product of competence rather than an illustration of its core workings
- Language is about communication
- Communication is inherently meaningful

Usage-based turn in linguistics

- The poor performance of symbolic NLP is an indication that something's wrong with the idea of 'competence = symbolic rules'
- Cf. The Ptolemaic view (earth first) > all kinds of problems > Copernican revolution (sun first)
- 'Syntactic competence' first > Performance (semantics/communication) first
- Corpus studies study actual language in use
- Also not without problems
  - Intentions of speakers remain a black box

Collocations & second wave of NLP

- Statistical NLP
- Unlike purely symbolic approaches, statistical approaches mine (parallel) corpora for recurrent patterns
- Frequency of co-occurrence is integrated into machine translation
- Captures idiomatic information better
  - *Maria no **daba una bofetada** a la bruja verde*
  - *Mary did not **slap** the green witch* (better than ***give a slap to***)
- *dar una bofetada* = a **collocation**
  - a series of words or terms that co-occur more often than would be expected by chance

Cognitive linguistics (e.g. Ronald Langacker)

- Generally claims that language is *not* modular
- Specific branch: construction grammar (e.g. Goldberg 1995, Croft 2001)
- Idiomatic expressions are everywhere
  - *by and all*
  - *the bigger they come, the harder they fall*
  - *possess much?*
- Appendix to grammar?
- No, core of grammar: constructions
  - May contain fixed words in them (unlike traditional syntactic rules)
  - They are often productive (unlike traditional lexical items)
    the X-er (Y Z), the X'-er (Y Z)

Semantics vs. pragmatics

- Semantics: entrenched meaning of a word
- Pragmatics: additional meanings a word gets from the context
- Blurry line
- *Could you pass me the salt please?*
  - Is this really just a pragmatic interpretation of a sum of words?
  - More likely stored as a chunk in our memory
- Things such as sarcasm (*Yeah, great!*) or understatement (*This is not ideal*)
  - May cause a mess in sentiment analysis, for instance

Issues with data-driven NLP

Original	Google translate	deepl.com
• M: And..umm…what he would do would be if myself ..or..or..and one of my colleagues were having a conversation that was in any way, uhh, related to something other than the immediate job that we were working on…	M: En..umm ... wat hij zou doen zou zijn als ik ... of..of..en een van mijn collega's een gesprek hadden dat op enigerlei wijze, uhh, verband hield met iets anders dan de directe baan die we werkten aan ...	M: En...umm...wat hij zou doen zou zijn als ik...of...of...en een van mijn collega's een gesprek hadden dat op enigerlei wijze, uhh, gerelateerd was aan iets anders dan de directe baan waar we mee bezig waren...
• L: Uh huh.	L: Uh huh.	L: Uh huh.
• M: …he would, uhh, jump in there and and and, uhh, tell us off! For doing it…	M:… hij zou, uhh, erin springen en en, uhh, vertel het ons! Om het te doen ...	M: ...hij zou, uhh, daar in springen en, uhh, ons vertellen! Om het te doen...
• L: Really?	L: Echt?	L: Echt waar?

What all these notions are struggling with

What all these notions are struggling with

- Meaning is not like a label attached to a word
- Meaning is holistic
  - distributed across a sentence / text
  - co-determined by non-verbal context
- Human cognition extracts **patterns** out of the constant flow of consciousness

What all these notions are struggling with

- Salient patterns may function as prototypes / pivots for linguists to use for analytical purposes
  - Noun : typically a thing
  - Verb : typically an action
  - etc.
- It is very dangerous (wrong) to assume that these inferred categories are the basis on which humans act when speaking
- Cf. Dąbrowska (e.g. 2008) work on passives
  - *The man was bitten by the dog*
  - *The dog was bitten by the man*

The promise of neural approaches?

- State-of-the-art AI uses artificial neural networks (ANNs) to uncover **patterns**
- Are they similar to the patterns that human cognition extracts?
- Is there cognitively plausible emergence of hierarchy?
- Hard to tell
- Linguistic terminology may help to avoid certain pitfalls

Linguistic theory as an ill-news bearer

- Word-embeddings (e.g. Word2Vec) transfer a word into a real vector based on its distributional properties
- However, the input for word-embeddings typically is
  - a written text
  - naively tokenized (on the basis of whitespace)
- Issues
  - Homonyms: words that sound the same but aren't

    The modal auxiliary *may* and the month *May*
  - Word-embeddings will still assign only one vector to 'may'
  - This may cause serious issues.

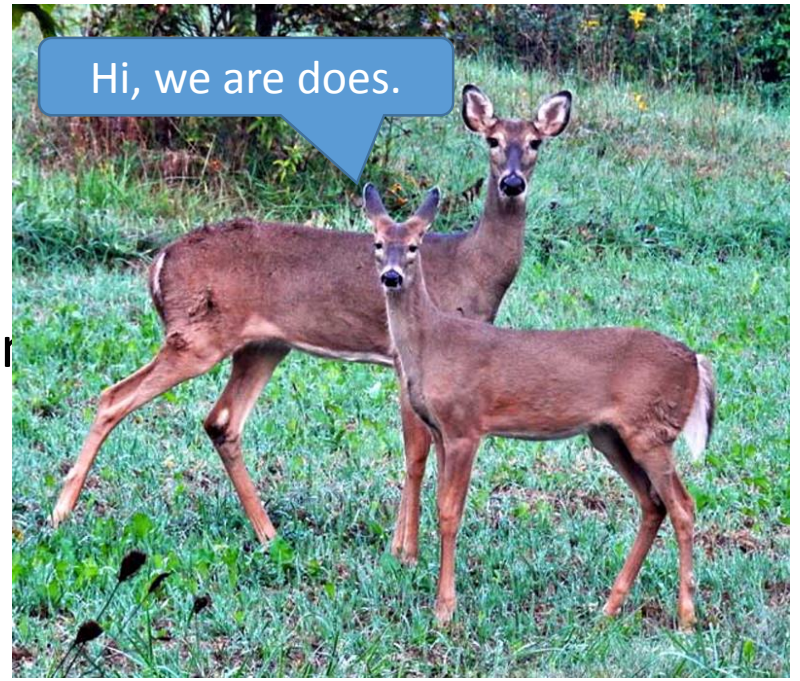
    Example: *does* vs. *doth* in the 17th century

Linguistic theory as an ill-news bearer

- Word-embeddings (e.g. Word2Vec) transfer a word into a real vector based on its distributional properties
- However, the input for word-embeddings typically is
  - a written text
  - naively tokenized (on the basis of whitespace)
- Issues
  - Homonyms: words that sound the same but aren't

    The modal auxiliary *may* and the month *May*
  - Word-embeddings will still assign only one vector to '
  - This may cause serious issues.

    Example: *does* vs. *doth* in the 17th century



Linguistic theory as a requirement

- Proper design of a classifier heavily depends on
- Neural networks may works unsupervised (Word2Vec e.g.)
- May start from training data as well for more specific tasks
- E.g.
  - *It is God's will that we work and do of his good pleasure.*
  - *It is a good divine that follows his own instruction.*
  - *It was hog's flesh that she had got of the Soldiers.*

Linguistic theory as a requirement

- Proper design of a classifier heavily depends on
- Neural networks may works unsupervised (Word2Vec e.g.)
- May start from training data as well for more specific tasks
- E.g.
  - *It is God's will that we work and do of his good pleasure.* = Extraposition
  - *It is a good divine that follows his own instruction.* = Cleft
  - *Her Child being long missed, her acquaintance asked her where it was, and how she came by that Flesh, she replyed, It was Hogs flesh that she had got of the Souldiers.* = referential it

Paradigmatic vs. syntagmatic

- Paradigmatic axis
  - Complete set of related word forms associated with a given lexeme
  - Verb paradigm (*I do, you do, he does etc.*)
    - In corpus studies also referred to as a *type*
  - Broader definition: Words/morphemes that can fill the same slot
- Syntagmatic
  - Words that co-occur
    - In corpus studies near-synonymous to *collocation/co-occurrence*

From syntagmaticity to paradigmaticity

[He] [went] [home] [to] [feed] [the] ([cat] [dog] [kangaroo?] [~~train~~])

- Traditional collocational analysis = syntagmatic:
  - “What are the collocates of these constructions?”
- ANNs = from syntagmatic to paradigmatic:
  - “Based on their collocates, how similar are these constructions?”

Linguistic theory as an opportunity

- A good grasp of linguistic theory also offers many opportunities
- Complement vs. adjunct
  - complement = required part to have a full sentence
  - adjunct = optional part
  - Interpretation may depend on what the function is
    - *He got liberated out of prison* vs. *He got out of prison, liberated*
- Variation in complementation (Givón 1980
  - *She told him to eat*
  - *She told that he should eat*
  - *She ordered him to do it*
  - *She ordered that he do it*
  - **She insisted him to do it*

More terminology if you can’t get enough

- Onomasiology vs. semasiology
- Ambiguity vs. vagueness
- Collocation vs. collostruction
- Construction vs. phrase vs. constituent
- Descriptive vs. prescriptive grammar
- Rhyme, metrum, prose, verse
- Spoken vs. written
- Register, genre, style
- Idiom, chunk, priming
- cognitive schema, network
- Individual vs. aggregate levels