# CS11-737:
# Multilingual Natural Language Processing

## Words

Alan W Black (originally by Yulia Tsvetkov)

Carnegie Mellon University
Language Technologies Institute

# What is a word?

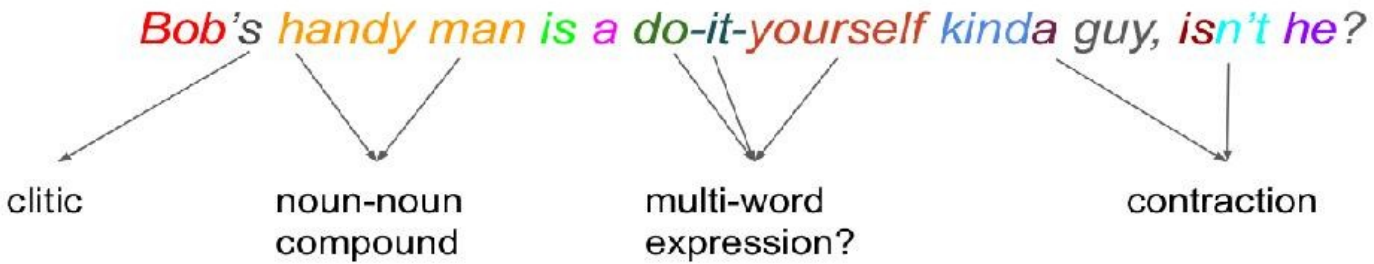
- Count the words:

*Bob's handyman is a do-it-yourself kinda guy, isn't he?*

# What is a word?

*Bob's handy man is a do-it-yourself kinda guy, isn't he?*

# What is a word?



# What is a word?

*Bob's handy man is a do-it-yourself kinda guy, isn't he?*

Much'ananayakapushasqakupuniñataqsunamá

Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

*"So they really always have been kissing each other then"*

| | |
|---|---|
| Much'a | to kiss |
| -na | expresses obligation, lost in translation |
| -naya | expresses desire |
| -ka | diminutive |
| -pu | reflexive (kiss *eachother*) |
| -sha | progressive (kiss*ing*) |
| -sqa | declaring something the speaker has not personally witnessed |
| -ku | 3rd person plural (they kiss) |
| -puni | definitive (really*) |
| -ña | always |
| -taq | statement of contrast (...then) |
| -suna | expressing uncertainty (So...) |
| -má | expressing that the speaker is surprised |

ושבתה

| | |
|---|---|
| and her saturday | ו+שבת+ה |
| and that in tea | ו+ש+ב+תה |
| and that her daughter | ו+ש+בת+ה |

# Structural subfields of linguistics

| | |
|---|---|
| Phonetics | The study of the sounds of human language |
| Phonology | The study of sound systems in human languages |
| Morphology | The study of the formation and internal structure of words |
| Syntax | The study of the formation and internal structure of sentences |
| Semantics | The study of the meaning of sentences |
| Pragmatics | The study of the way sentences with their semantic meanings are used for particular communicative goals |

# Words

- Orthographic definition
  - strings separated by white spaces
  - problems: *Bob's handy man is a do-it-yourself kinda guy, isn't he?*
  - unwritten languages, languages that don't use white spaces, etc.

# Words

- Orthographic definition
  - strings separated by white spaces
  - problems: *Bob's handy man is a do-it-yourself kinda guy, isn't he?*
  - unwritten languages, languages that don't use white spaces, etc.
- Prosodic definition
  - words have one main stress and longer words may have a secondary stress
  - problems: function words, clitics
-

# Words

- Orthographic definition
  - strings separated by white spaces
  - problems: *Bob's handy man is a do-it-yourself kinda guy, isn't he?*
  - unwritten languages, languages that don't use white spaces, etc.
- Prosodic definition
  - words have one main stress and longer words may have a secondary stress
  - problems: function words, clitics
- Semantic definition
  - words are units that describe a single idea or a semantic concept
  - problem: many semantic concepts span phrases or sentences and don't have a corresponding word

# Words

- Orthographic definition
  - strings separated by white spaces
  - problems: *Bob's handy man is a do-it-yourself kinda guy, isn't he?*
  - unwritten languages, languages that don't use white spaces, etc.
- Prosodic definition
  - words have one main stress and longer words may have a secondary stress
  - problems: function words, clitics
- Semantic definition
  - words are units that describe a single idea or a semantic concept
  - problem: many semantic concepts span phrases or sentences and don't have a corresponding word
- Syntactic definition
  - words are the syntactic building blocks of sentences

# Parts of speech

- Open classes
  - nouns
  - verbs
  - adjectives
  - adverbs
- Closed classes
  - prepositions
  - determiners
  - pronouns
  - conjunctions
  - auxiliary verbs

# Part of speech tagsets

- Penn treebank tagset (Marcus et al., 1993)

| Tag | Description | Example | Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|---|---|---|
| CC | coordinating conjunction | *and, but, or* | PDT | predeterminer | *all, both* | VBP | verb non-3sg present | *eat* |
| CD | cardinal number | *one, two* | POS | possessive ending | *'s* | VBZ | verb 3sg pres | *eats* |
| DT | determiner | *a, the* | PRP | personal pronoun | *I, you, he* | WDT | wh-determ. | *which, that* |
| EX | existential 'there' | *there* | PRP$ | possess. pronoun | *your, one's* | WP | wh-pronoun | *what, who* |
| FW | foreign word | *mea culpa* | RB | adverb | *quickly* | WP$ | wh-possess. | *whose* |
| IN | preposition/ subordin-conj | *of, in, by* | RBR | comparative adverb | *faster* | WRB | wh-adverb | *how, where* |
| JJ | adjective | *yellow* | RBS | superlatv. adverb | *fastest* | $ | dollar sign | $ |
| JJR | comparative adj | *bigger* | RP | particle | *up, off* | # | pound sign | # |
| JJS | superlative adj | *wildest* | SYM | symbol | +,%, & | “ | left quote | ‘ or “ |
| LS | list item marker | *1, 2, One* | TO | “to” | *to* | ” | right quote | ’ or ” |
| MD | modal | *can, should* | UH | interjection | *ah, oops* | ( | left paren | [, (, {, < |
| NN | sing or mass noun | *llama* | VB | verb base form | *eat* | ) | right paren | ], ), }, > |
| NNS | noun, plural | *llamas* | VBD | verb past tense | *ate* | , | comma | , |
| NNP | proper noun, sing. | *IBM* | VBG | verb gerund | *eating* | . | sent-end punc | . ! ? |
| NNPS | proper noun, plu. | *Carolinas* | VBN | verb past part. | *eaten* | : | sent-mid punc | : ; ... – - |

# The Universal Dependencies

## Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing more than 150 treebanks in 90 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Discussion on UD
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained by the Charles University in Prague
  - Kontext maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
  - INESS maintained by the University of Bergen
- Download UD treebanks

https://universaldependencies.org

| Open class words | Closed class words | Other |
|---|---|---|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

# Morpheme

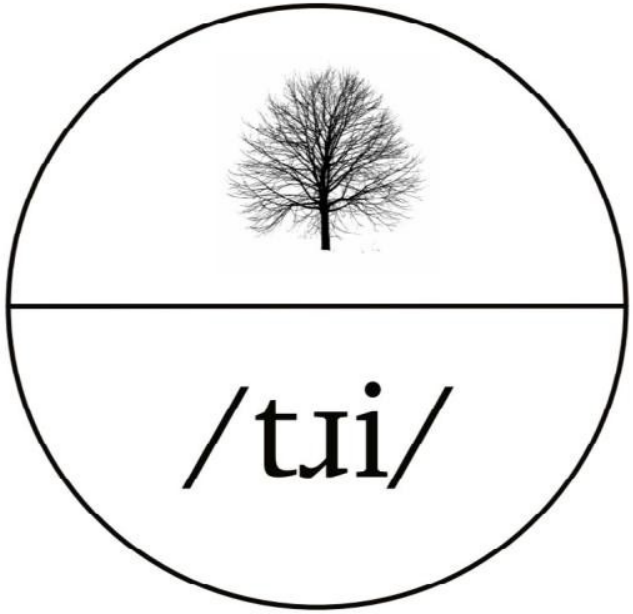
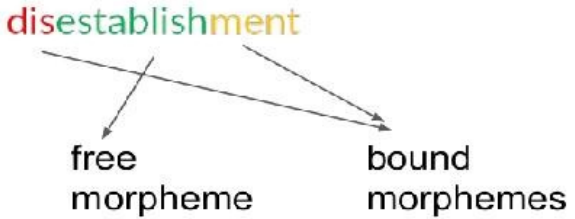



Image from Lori Levin and David R. Mortensen's draft book "Human Languages for Artificial Intelligence"

# Words are made of morphemes

*Bob's handy man is a do-it-yourself kinda guy, isn't he?*

- establish (V)
- disestablish (V)
- disestablishment (N)
- antidisestablishment (N)
- antidisestablishmentary (A)
- antidisestablishmentarian (N)
- antidisestablishmentarianism (N)




Example by Austin Matthews

# Morphological processes

- concatenation
- affixation = stem+affix
  - prefix
  - suffix
- non-concatenative affixation
  - infix
- compounding = stem+stem

- establish (V) → stem
- disestablish (V) → prefix + stem
- disestablishment (N) → prefix + stem + suffix =circumfixation

- dish (N) + washer (N) = dishwasher (N)

# Tagalog

- Tagalog
  - stem - *bundok*
  - singular - *mabundok*
  - plural - *mabubundok*
  - gloss - 'mountainous'

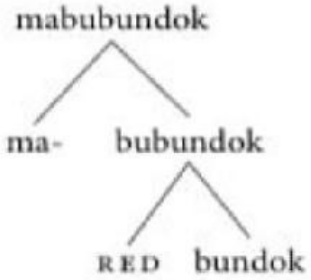



Figure 34: The morphological structure of *mabubundok*

Example from Lori Levin and David R. Mortensen's draft book "Human Languages for Artificial Intelligence"

# Arabic, Chinese

- Arabic
  - root and pattern morphology

| | |
|---|---|
| *katab-a* | 'he wrote' |
| *kaataba* | 'he corresponded' |
| *kutib-a* | 'it was written' |
| *kitaab* | 'book' |
| *kutub* | 'books' |
| *kaatib* | 'writer; writing' |
| *kuttaab* | 'writers' |
| *uktub* | 'write (to a male)!' |

Table 17: Part of the Arabic paradigm for *ktb* 'with reference to writing'.

- Chinese
  - compound words

| | | | | |
|---|---|---|---|---|
| 客厅 | 'living room' | 沙发 | 'sofa' | 'living room sofa' |
| 眼 | 'eye' | 药 | 'medicine' | 'eye medicine' |
| 马 | 'horse' | 房 | 'house' | 'manger' |
| 雨 | 'rain' | 帽 | 'hat' | 'rain hat' |

# Morphological functions

- Derivational morphemes
  - bound morphemes used to create new words
  - is these affixes are attached to a new base, the resulting combination yields a word with a new meaning
  - often derived word belongs to a different syntactic class
- Inflectional morphemes
  - bound morphemes used to mark grammatical distinctions
  - change the form but not POS tag or the key meaning of the word

- establish (V)
- disestablish (V)
- disestablishment (N)

- eat (V) + -s = eats (V)

# Morphological Levels

- **Morphosyntax:**
  - How stems and affixes combine
  - e.g verb + ed, verb + ing, un + grace + ful + ly
- **Morphophonemics:**
  - Pronunciations/Orthographic modifications at boundaries
  - e:_ < C _ + e,  “e” gets deleted when preceded by a consonant, and followed by a morpheme boundary and a morpheme starting with “e”
  - N:m < i _ + [mbp]  “n” becomes “m” at morpheme boundary followed by “m”, “b” or “p”
  - Morphophonemics can make morphology non-segmental

# Morphological typology

- Isolating or Analytic
  - Vietnamese, Chinese, English
- Synthetic
  - Fusional or Flexional
    - German, Greek, Russian
    - Templatic: Hebrew and Arabic
  - Agglutinative or Agglutinating
    - Finnish, Turkish, Malayalam, Swahili
  - Polysynthetic
    - Inuit, Yupik

# UniMorph

UniMorph
Schema and datasets for universal morphological annotation

Schema Software Publications Contact

## UniMorph

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema. The specification of the schema is described here and in Sylak-Glassman (2016).

Plus, we're now available in a Python package! `pip install unimorph`

### UniMorph Events

- SIGMORPHON 2019 Shared Task
- CoNLL–SIGMORPHON 2018 Shared Task
- CoNLL–SIGMORPHON 2017 Shared Task
- SIGMORPHON 2016 Shared Task

https://unimorph.github.io

### Annotated Languages

The following 110 languages have been annotated according to the UniMorph schema. Missing parts of speech will be filled in soon


# SIGMORPHON

Special Interest Group on Computational Morphology and Phonology



## Workshops

- **2020**: Seattle, co-located with ACL 2020
- **2019**: Florence, co-located with ACL 2019
- **2018**: Brussels, co-located with EMNLP 2018
- **2016**: Berlin, co-located with ACL 2016
- **2014** (with SIGFSM): Baltimore, co-located with ACL 2014
- **2012**: Montréal, co-located with NAACL-HLT 2012
- **2010**: Uppsala, co-located with ACL 2010
- **2008**: Columbus, co-located with ACL 2008

# The SIGMORPHON shared tasks

- Cross-lingual transfer for morphological inflection
- Morphological analysis in context
- Morphological paradigm completion

**The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection**

**Arya D. McCarthy**[♣], **Ekaterina Vylomova**[♥], **Shijie Wu**[♣], **Chaitanya Malaviya**[♠], **Lawrence Wolf-Sonkin**[♦], **Garrett Nicolai**[♣], **Christo Kirov**[♣], **Miikka Silfverberg**[♯], **Sebastian Mielke**[♣], **Jeffrey Heinz**[♭], **Ryan Cotterell**[♣], and **Mans Hulden**[♯]

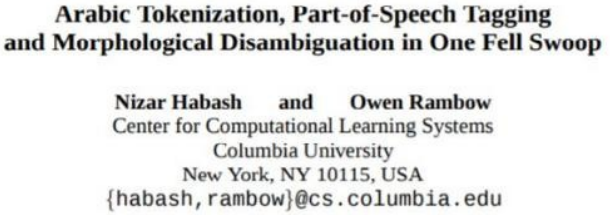
[♣]Johns Hopkins University [♥]University of Melbourne [♠]Allen Institute for AI
[♦]Google [♯]University of Helsinki [♭]Stony Brook University [♯]University of Colorado

# Morphological Analyzers

- **Finite state morphology**
  - Skilled, but not very hard (by experts)
  - Xfst, FOMA
- **Unsupervised methods**
  - Morfessor (python)
  - Assumes segemental view of morphology
- **Stemming**
  - Remove “ends” of words (doesn’t always do the right thing)
- **BPE (Byte pair encoding)**
  - Find “optimal” character level segmental split

# Related NLP problems

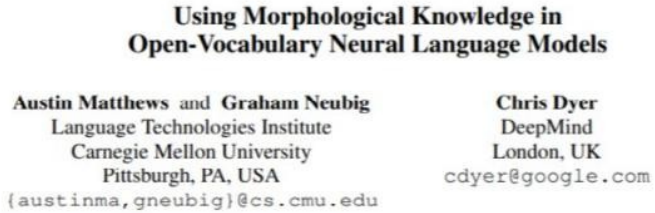
- tokenization
- lemmatization



**Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop**

**Nizar Habash** and **Owen Rambow**
Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA
{habash,rambow}@cs.columbia.edu

- processing words in multilingual NLP tasks, e.g. language modeling or machine translation
  - tokens
  - characters
  - subwords
  - +morphological knowledge



**Using Morphological Knowledge in Open-Vocabulary Neural Language Models**

**Austin Matthews** and **Graham Neubig**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
{austinma,gneubig}@cs.cmu.edu

**Chris Dyer**
DeepMind
London, UK
cdyer@google.com

- syntactic tagging (next class) and morphological analysis (later in the course)

# Further Related Problems

- **Text Normalization**
  - Replaces, numbers, symbols, abbreviations with standard words
  - Non-Standard Word expansion (Sproat et al 2001 CSL 15(3) “Normalization of Non-standard Words”)
- **Spelling correction/normalization**
  - Social media speak: lol
- **Tokenization Mismatch**
  - BERT vs what you have

# Readings and class discussion

- Read Chapter 2 in Bender E., 2013. Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax.

- Pick a language in one of the following branches of language families:
  Bantu, Dravidian, Finno-Ugric, Japonic, Papuan, Semitic, Slavic, Turkic languages.
  Tell us about some interesting aspects of morphology of that language, following examples from the assigned reading; cite your sources.

  If you would need to implement a tokenizer for that language, what language specific knowledge it would be important to incorporate into the tokenizer?