Computational Morphology and Syntax of Natural Languages

Daniel Zeman

http://ufal.mff.cuni.cz/course/npfl094
zeman@ufal.mff.cuni.cz
NPFL094

• Presentations and talks will be in English
  – Unless all students understand Czech
• Questions welcome in both Czech and English
• And I have many examples from Czech 😊
Caution

• **No class on**
  
  October 28
  November 4
  November 25
Getting Credits

• 2-3 smaller tasks
  – homework style
  – less flexible deadlines

• Alternatively: one larger project
  – ask me if interested
  – can be combined with your mgr. (or bc.) thesis
An “Unbalanced” Course

- 1/3 linguistics, 2/3 tools
- 1/3 lab work, 2/3 lectures
- ¾ morphology, ¼ syntax
- Mostly rule-based
  - almost no machine learning
  - no neural networks
Outline: Morphology

- Morphemic segmentation
  - un + beat + able
- Phonology ("morphonology") and orthography
  - baby + s = babies
- Inflectional vs. derivational morphology
- Morphological analysis: word form → lemma + morphosyntactic features (tag)
- Tagging (context-aware disambiguation)
- Unsupervised affix detection in corpus
- Mining of word forms from corpus
Morphological Analysis

- **Input:**
  - word form (token)

- **Output:**
  - set (possibly empty) of analyses
  - an analysis:
    - lemma (base form of the lexeme)
    - tag (morphological, POS)
      - part of speech
      - features and their values
MA Example

- Language: Czech
- Input: *malými*
- Output (only one selected analysis here):
  - lemma = *malý* (“small”)
  - tag = AAFP71A
    - part of speech = AA (adjective / přídavné jméno)
    - gender = F (feminine / ženský)
    - number = P (plural / množné)
    - case = 7 (instrumental / 7. pád)
    - degree of comparison = 1 (positive / 1. stupeň) …
MA Example

• Language: English
• Input: flies
• Output:
  – lemma 1 = fly-1 (to move in the air)
  – tag 1 = VBZ (verb, present tense 3rd person singular)
  – lemma 2 = fly-2 (an insect)
  – tag 2 = NNS (noun, plural)
• Output is not disambiguated with respect to context
MA versus Tagging

• By *tagging* we usually mean context-based disambiguation

• Most taggers employ statistical methods

• Taggers may or may not work on top of MA
  – MA may provide readings not known from training
  – If a tagged corpus is available but MA is not, a tagger can still be trained on the corpus
Morphemic Segmentation

- **Morpheme** is the smallest unit of language that conveys some meaning
- Morphemic segmentation = finding morpheme boundaries within words
- Typically part of MA:
  - input: *closed*
  - identify the morphemes: *close* + *d*
  - interpret them: *verb (close) + past tense*
  - output: *close* + *VBD*
Morphemic Segmentation

- Sometimes it is useful to know the morphemes even if we cannot interpret them
  - Data sparseness, e.g. in machine translation:
    - en: city
    - cs alignments in parallel corpus: 
      - město (nom/acc/voc sg, 42×), města (gen sg, nom/acc/voc pl, 40×), městě (loc sg, 32×), měst (gen pl, 9×), městské (adj, 7×), městem (ins sg, 7×), městských (adj, 4×), městská (adj, 4×), městský (adj, 2×), městu (dat sg, 2×), městech (loc pl, 2×)
    - missing cs: městům (dat pl), městy (ins pl), městského, městskému, městském, městským, městští, městskými, městskou (adj remaining forms)
Morphemic Segmentation

• Sometimes it is useful to know the morphemes even if we cannot interpret them
  – Data sparseness, e.g. in machine translation

• **Stemming** = stripping all morphemes but the *stem*
  – IN: *The British players were unbeatable.*
  – OUT: *the Brit play were beat.*

• **Lemmatization** = replacing all words with their lemmas (as with tagging, disambiguation may be assumed)
  – OUT: *the British player be (un)beatable.*
Inflection vs. Derivation

- **Derivational morphology:**
  - New lemma!
  - Often (but not always) new part of speech.

- **Inflectional morphology:**
  - Set of forms of one lemma (lexeme)
  - The set is called *paradigm*

- The borderline is sometimes quite fuzzy
Outline: Syntax

- Constituency vs. dependency
- Context-free grammars
- Transition network grammars
- Shallow parsing (chunking)
- Chart parsers
- Dependency parsers (transition-, graph-based)
- Clause boundaries
A record date has n't been set.
The governor couldn't make it, so the lieutenant governor welcomed the special guests.
Applications of Morphology

• First step before broader NLP applications:
  – (Input for (syntactic) parsing)
  – (Machine translation)
    • Rule-based MT: full-fledged analysis and generation
    • Statistical MT: fighting data sparseness
  – Finding word boundaries (Chinese, Japanese)
  – Dictionaries
Applications of Morphology

• Text-to-speech systems (speech synthesis)
  – Morphology affects pronunciation
    • English *th* is normally pronounced *θ* or *ð*
    • However, not in *boathouse* (*boat + house*)
    • Czech *proudit* =
      – *proud + it* (“stream” + INF = “flow”)
      – *pro + ud’ + it* (“through” + “smoke” + INF = “smoke thoroughly”)

• (Speech recognition)
  – Morphology allows for smaller dictionaries
Applications of Morphology

• Word processing
  – Spell checking dictionaries
  – Inputting Japanese text
    • Two *kana* syllabic scripts and *kanji* (Chinese characters)
    • Typically, people type in *kana* and system converts to *kanji* whenever necessary
    • Disambiguation needed!
    • Bound morphemes remain in *kana* (morpho rules)
Applications of Morphology

• Word processing: find & replace terms
  – Czech: *kniha* (*book*) ⇒ *dílo* (*work*)
    - *knihy* ⇒ *díla*, *knize* ⇒ *dílu*, *knihu* ⇒ *dílo*, *kniho* ⇒ *dílo*, *knihou* ⇒ *dílem*, *knih* ⇒ *děl*, *kniham* ⇒ *dílům*, *knihách* ⇒ *dílech*, *knihami* ⇒ *díly*

• Document retrieval
  – Keywords in query are typically base forms
  – The forms in documents are inflected
Morphology-Based Typology

• Isolating languages
  – Chinese: ǒu bú ài chī qīngcài
    = dog not like eat vegetable

• Inflectional languages
  – Romance and Slavic languages: Spanish puedo = poder +
    present indicative, 2nd person, singular

• Agglutinative languages
  – Turkish: çöplüklerimizdende kilerden migraine = çöp + lük + ler + imiz +
    de + ki + ler + den + mi + y + di = “was it from those that were in
    our garbage cans?”

• Polysynthetic languages
  – Eskimo languages
Polysynthetic Languages

- Found in Siberia and the Americas
- Intricately compose words of many lexical morphemes that are not easily told apart
  - Typically include both subject- and object-verb agreement.
- That’s why linguists decided not to separate them orthographically
- Nevertheless, *words* usually *are* separated. They are just long
- One long word may cover a whole sentence in other languages
- Chukchi example (Skorik 1962: 102):
  - *Tə-meɣ-ə-leɣ-tɼ-ə-rkən.*
  - 1.SG.SUBJ-great-head-hurt-PRES.1
  - “I have a fierce headache.”
Morphological Devices (Overview)

- Affixes (prefixes and suffixes): concatenative morphology
- Compounding
- Infixation
- Circumfixation
- Root and pattern (templatic) morphology
- Reduplication
- Subsegmental morphology
- Zero morphology
- Subtractive morphology
Affixation

• Most common way of inflection and derivation

• Three morpheme types:
  prefix + radix (stem) + suffix
  – en: dog + s = dogs
    • plural suffix –s
  – de: mach + st = machst
    • suffix –st marks present indicative 2nd person singular
  – en: un + beat + able
    • prefix un- negates the meaning
    • suffix –able converts verb to adjective, expressing applicability of the action of the verb to something
Infixation

• Languages of the Philippines, e.g. Bontoc:
  – fikas “strong” ⇒ f-um+ikas “be strong”
  – kilad “red” ⇒ k-um+ilad “be red”

• Could be analyzed as prefix to (stem minus the initial consonant)
Circumfixation

• Prefix + suffix act together as one morpheme
  – German: *legen* “lay down” ⇒ *ge+leg+t* “laid down”
  – Indonesian: *besar* “big” ⇒ *kə+besar+an* “bigness”

• Similar, but not the same as Czech superlatives
  – *nej* + *mlad* + š + í  “youngest”
  – superlative + stem + comparative + singular nominative
Templatic Morphology

- Semitic languages (Arabic, Hebrew, Amharic)
- Arabic:
  - root (usually 3 consonants): *ktb* “write”
  - vowel pattern: *aa* = active, *ui* = passive
  - template: CVCVC = first verb derivational class (*binyan*)
  - result: *katab* “write”, *kutib* “be written”
Reduplication

• Copy whole stem or part of it
  – Indonesian plural:
    • *orang* “man” ⇒ *orang+orang* “men”
  – Javanese habitual-repetitive:
    • *adus* ⇒ *odas+adus* “take a bath”
    • *bali* ⇒ *bola+bali* “return”
  – Yidin (an Australian language):
    • *gindalba* ⇒ *gindal+gindalba* “lizard”

• Reduplication cannot be modeled by finite-state automata!
Subsegmental Morphology

- Irish:
  - *cat* (/kat/) = “cat” (singular)
  - *cait* (/kæt/) = “cats” (plural)

- The plural morpheme consists just of one phonological feature (“high”), resulting in palatalization.
Zero Morphology

• Zero (empty) morpheme, marked sometimes as 0, \( \emptyset \), \( \lambda \) or \( \varepsilon \)
  
  – Czech feminine plural case endings for \( \text{žena} \) “woman”:
    • nom: \( \text{žen}+y = \text{ženy} \)
    • gen: \( \text{žen}+\lambda = \text{žen} \)
    • dat: \( \text{žen}+\text{ám} = \text{ženám} \)
    • acc: \( \text{žen}+y = \text{ženy} \)
    • voc: \( \text{žen}+y = \text{ženy} \)
    • loc: \( \text{žen}+\text{ách} = \text{ženách} \)
    • ins: \( \text{žen}+\text{ami} = \text{ženami} \)
Subtractive Morphology

- Koasati (a Muskogean language, southeast US):
  - singular verb: $pitaf + fi + n$
  - plural: $pit + li + n$
  - singular verb: $lasap + li + n$
  - plural: $las + li + n$

- Such examples are rare
- Moreover, one might argue that plural is the base form here
Compounding

• English: maximally two stems written together
• Germanic languages in general favor compounds
• de: *Hotentotenpotentatentantenatentäter*
  – *Hotentot + en + Potentat + en + Tante + n + Atentäter*
  – “Hottentot potentate aunt assassin”
  – “assassin of aunt of potentate of Hottentots”
Recommended Further Reading

- These books may be difficult to obtain from the MFF library. Reading them is not required.
  - James Allen: *Natural Language Understanding*. Benjamin/Cummings, USA, 1995
  - Daniel Zeman: *The World of Tokens, Tags and Trees*. ÚFAL, Czechia, 2018