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 22
&
October 29
Projects

• Each student will work on a different language
• Depending on complexity and resources available, you will create a lexicon, morphosyntactic analyzer, syntactic grammar etc.
  – Desired coverage of the final product will be determined later
• You will give short presentations to keep others informed about your project.
Pick up a Language

• Until next week, think about “your” language
• Typically, each student would work on a different language ⇒ wiki
• More interesting are those languages for which few or no resources exist
  – Model situation: doing MT with unknown language, need morphology, nothing is freely available
• Do not choose English or Czech. Possibilities are:
  – your native language if it’s not English or Czech
  – another language you have some knowledge of
  – or even a language you don’t know but can get and read some information about its grammar
SVN Accounts

[NPFL094://]
...
your_login = rw
...

svn.ms.mff.cuni.cz
  > htpasswd your_login
Subversion (svn)

- ÚFAL svn server: svn.ms.mff.cuni.cz
  - You need an svn login name and password!
- Repository “NPFL094”
- Check out your working copy (this will create new subfolder npfl094 in your current folder):
  
  ```
  svn --username $USER checkout https://svn.ms.mff.cuni.cz/svn/NPFL094 npfl094
  ```
- Subsequent operations on this working copy do not require password
  
  ```
  svn update
  ```
  - get new changes made by others and merge them with your working copy
  - your working copy may still contain other changes invisible to others!
  
  ```
  svn commit -m 'changed encoding of lexicon'
  ```
  - save your changes to the repository (and describe in the –m string)
  
  ```
  svn add file.txt
  ```
  - add new file under version control (otherwise it will be ignored by svn commit)
Language Resources

- Raw corpus (can be acquired from the web)
  - lexicon, affix repertory
- Language knowledge (conventional textbook / web)
- Tagset(s) with description
- Tagged corpus
  - i.e. someone has morphology but it’s not available
- Textbook or website on grammar of the language
- Morphemic segmenter, morphological analyzer, tagger
- Parser (syntactic analyzer)
Language Resources

• How do I know what resources are available?
• Start at the ACL wiki:

• Google
• NLP departments in countries where the language is spoken

• If morphology is publicly available, you may want to work on a (syntactic) parser (but first get the morph. running!)
Various Sorts of Work

- Designing grammar rules … formal linguistics
- Processing word lists, acquiring lexicons … programming, web crawling, improvisation
- Using existing tools
- Own programming
  - should be as platform-independent as possible
    - e.g. Perl is better than shell scripts
    - non-standard libraries, OS-dependent features should be avoided
  - if it must be platform dependent, then prefer Linux!
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.*
Outline: Syntax

- Constituency vs. dependency
- Context-free grammars
- Transition network grammars
- Shallow parsing (chunking)
- Chart parsers
- Dependency parsers (Malt, MST)
- Clause boundaries
A record date has n't been set.
The governor could n'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 *proudít* =
      – *proud* + *it* (“stream” + verb = “flow”)
      – *pro* + *ud’i* + *it* (“through” + “smoke” + verb = “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*, *kníze* ⇒ *dílu*, *knihu* ⇒ *dílo*, *kniho* ⇒ *dílo*, *knihou* ⇒ *díleme*, *knih* ⇒ *děl*, *knihám* ⇒ *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: góu bú ài chī qīngcài
    = dog not like eat vegetable

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

• Agglutinative languages
  – Turkish: çöplüklerimizdeki+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):
  - $T\-\text{metry-}a\-\text{levt-p\text{-}ryt-a-rk\text{\-}n}$.
  - 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* (/katʃ/) = “cats” (plural)

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

- Zero (empty) morpheme, marked sometimes as 0, Ø, λ or ε
  - Czech feminine plural case endings for žena “woman”:
    - nom: žen+y = ženy
    - gen: žen+λ = žen
    - dat: žen+ám = ženám
    - acc: žen+y = ženy
    - voc: žen+y = ženy
    - loc: žen+ách = ženách
    - ins: žen+ami = 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: *Hotentotenpotentatatantenatenentäter*
  - *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
Homework

• Pick a language, check available resources
• Write me to book it for you in the wiki
• Summarize your choice in 1 – 2 slides (PowerPoint / OpenOffice / PDF):
  – Interesting facts about the language and its grammar
  – Available resources; what you think needs to be done
• Send me the slides by Wednesday 6pm
• Present them to others on Thursday