Lexicon Acquisition

Daniel Zeman

October 17, 2024

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

About the Homework

- Goal: (partial) Morphological Analyzer for a Language
  - A morphologically interesting language (isolating languages, like Chinese, are not good candidates)
  - Preferably not Czech or English (although not strictly forbidden)
  - Ideally a low-resource language (MA not available yet), but not necessarily
  - You do not have to speak it fluently!
    - Understanding the language is obviously an advantage
    - Sufficient: you have access to description of morphology (Wikipedia?) + some data

About the Homework

- Goal: (partial) Morphological Analyzer for a Language
- Two parts:
  - Morphological lexicon (HW1)
    - List of words
    - Part of speech + inflection class ("which rules apply to this word?")
  - Morphological rules (HW2)
    - List of suffixes (and inflection classes they apply to)
    - Rules for phonological changes

Inflection Classes

pán “gentleman”

Rules

	Sing	**Plur**
Nom	*pán*	*pán + ové/i*
Gen	*pán + a*	*pán + ů*
Dat	*pán + ovi*	*pán + ům*
Acc	*pán + a*	*pán + y*
Voc	*pan + e*	*pán + ové/i*
Loc	*pán + ovi*	*pán + ech*
Ins	*pán + em*	*pán + y*

Lexicon

absolvent adresát advokát agent agresor akademik aktér alkoholik amatér anketiér antropolog architekt aristokrat asistent aspirant astrolog ...

Inflection Classes

muž “man”

Rules

	Sing	Plur
Nom	*muž*	*muž* + *i*
Gen	*muž* + *e*	*muž* + *ů*
Dat	*muž* + *i*	*muž* + *ům*
Acc	*muž* + *e*	*muž* + *e*
Voc	*muž* + *i*	*muž* + *i*
Loc	*muž* + *i*	*muž* + *ích*
Ins	*muž* + *em*	*muž* + *i*

Lexicon

akcionář autokrosař bakalář
bankéř bavič běžec běženec
bičovec brankář brusič císař
cizinec ctitel dálkař dědic
dějepisec ...

Inflection Classes

hrad “castle”

Rules

	Sing	Plur
Nom	*hrad*	*hrad* + *y*
Gen	*hrad* + *u*	*hrad* + *ů*
Dat	*hrad* + *u*	*hrad* + *ům*
Acc	*hrad*	*hrad* + *y*
Voc	*hrad* + *e*	*hrad* + *y*
Loc	*hrad* + *u*/*ě*	*hrad* + *ech*
Ins	*hrad* + *em*	*hrad* + *y*

Lexicon

adaptér aeroklub airbag akcent akt algoritmus alkohol amfiteátr antikvariát aparát apartheid appeasement areál argument arch archív ...

Inflection Classes

stroj “machine”

Rules

	Sing	Plur
Nom	*stroj*	*stroj* + *e*
Gen	*stroj* + *e*	*stroj* + *ů*
Dat	*stroj* + *i*	*stroj* + *ům*
Acc	*stroj*	*stroj* + *e*
Voc	*stroj* + *i*	*stroj* + *e*
Loc	*stroj* + *i*	*stroj* + *ích*
Ins	*stroj* + *em*	*stroj* + *i*

Lexicon

bič boj cíl děj desetiboj déšť doprodej drtič dvanáctiválec dvorec elektroodlučovač exemplář finiš hokej hrnec ...

Lexicon

Lemma	Class
abeceda	`NNF-zena`
absence	`NNF-ruze`
absolvent	`NNM-pan`
absolvování	`NNN-staveni`
adaptace	`NNF-ruze`
adaptér	`NNI-hrad`
adaptovanost	`NNF-kost`
administrativa	`NNF-zena`
adresa	`NNF-zena`
adresát	`NNM-pan`
advokát	`NNM-pan`
aeroklub	`NNI-hrad`
aféra	`NNF-zena`
...	...

Lexicon Acquisition

- Some hints only (approach must vary depending on language)
- Identify part of speech and inflection pattern
- If affixes restrict possible classes, use them!
  - 🇨🇿 Czech: the following suffixes increase likelihood of an infinitive: *-st, -át, -at, -ct, -ci, -ít, -out, -ýt, -ovat, -it, -ět, -et*
  - 🇬🇧 English: little inflection but verb forms and derivational suffixes *(-ness, -ity, -able)* can help
- Otherwise, syntax might help
  - E.g. if it's after preposition or article it's likely an adjective or a noun

Lexicon Acquisition

- Create word frequency list
- Identify closed-class words
    - Many of them will be very frequent
    - Textbook and/or bilingual dictionary may help with the rest
    - Parallel corpus + word aligner may supplement the dictionary
- What remains are mostly nouns, adjectives, verbs, and adverbs
    - Try to sort it out by iteratively looking at the word list, identifying repeating affixes etc.
    - If there are no repeating bound morphemes
        - then you may not be able to sort out the parts of speech
        - but maybe the morphology of the language is not so interesting after all

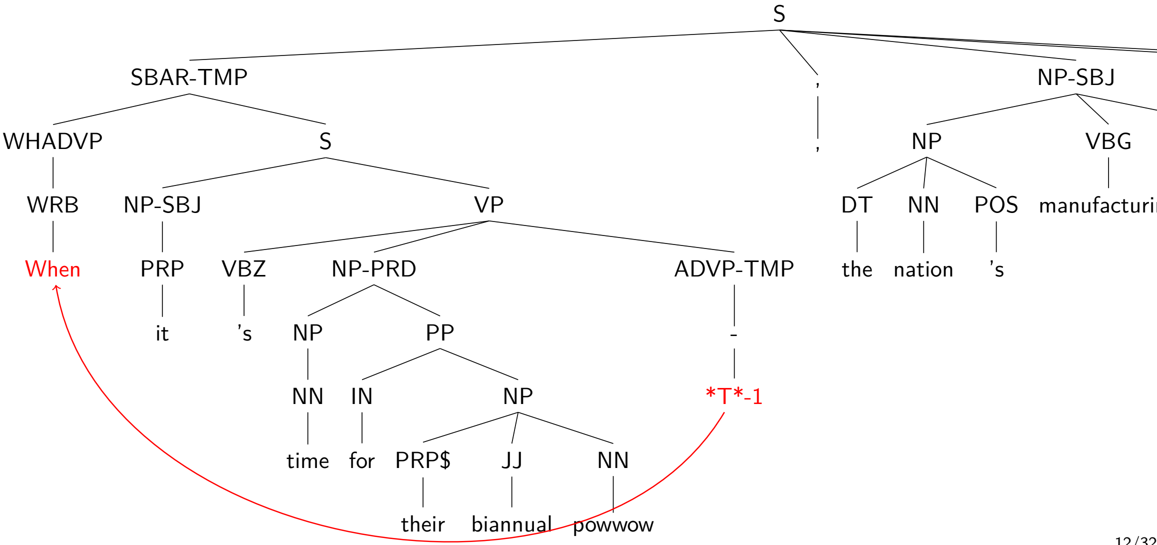
English Lexicon (LAB)

- `http://ufal.mff.cuni.cz/~zeman/vyuka/morfosynt/lab-lexicon/index.html`
- There is also a link from the NPFL094 website

English Lexicon Acquisition

- Example only! Other corpora and languages may require a different approach
- Input: plain-text (taken from Penn Treebank)
  - Tokenized (punctuation separated from words)
  - Remove traces (non-word terminal nodes in Penn Treebank): all tokens containing “*”?
  - Lowercase
    - Later we will want to identify proper nouns
    - Complicated by sentence-initial capitalization

Traces


S
SBAR-TMP
,
NP-SBJ
WHADVP
S
,
NP
VBG
WRB
NP-SBJ
VP
DT
NN
POS
manufacturi
When
PRP
VBZ
NP-PRD
ADVP-TMP
the
nation
's
it
's
NP
PP
-
NN
IN
NP
T-1
time
for
PRP$
JJ
NN
their
biannual
powwow

English Frequency Wordlist

- Penn Treebank 3 / Wall Street Journal:
- 49,208 sentences
- 1,253,013 terminal nodes (tokens and traces)
- 49,817 word types (opposed to word occurrences) including traces
- 43,764 lowercased types without traces and some other technical nodes (“error:” etc.)
- The most frequent types often have these (overlapping) properties:
  - stopwords
  - closed-class words
  - short words?

English Frequency Wordlist

,	60484
the	59459
.	48144
to	29576
of	28440
a	24781
in	21047
and	20449
’s	11556
for	10454
that	10422

$	8817
“	8735
is	8539
”	8506
it	7195
said	7141
on	6646
%	6121
at	5770
by	5705
as	5701

English Frequency Wordlist

,	60484
the	59459
.	48144
to	29576
of	28440
a	24781
in	21047
and	20449
’s	11556
for	10454
that	10422

I don’t use the counts in any heuristics. I only use them to show more frequent examples.

$	8817
“	8735
is	8539
”	8506
it	7195
said	7141
on	6646
%	6121
at	5770
by	5705
as	5701

Punctuation and Special Characters

m/\pP/

,	60484
.	48144
’s	11556
$	8817
“	8735
”	8506
%	6121
mr.	4950
n’t	4006
–	2585
u.s.	2056

third-quarter	333
buy-out	222
s&p	164
3,000	28
3.7	28
total types	10607
the rest	33157

- Caught, OK
- Not caught (but should have been caught)
- Caught (disputable)
- Caught (tokenization-related)

Numbers

`m/\pN/`

0	12447
10	668
30	607...
1988	503...
1,000	111...
1/2	105...
1.5	88...
30-year	79...
1980s	53...
ru-486	15...
mid-1980s	12...

b-2	7
19th	7
1989-90	5
80%-owned	4
xr4ti	4
total types	6123
the rest	37641
no punctuation or numbers	32218

Real Words

`!m/[\pP\pN`$]/`

the	59459	on	6646
of	28440	at	5770
to	27448	by	5705
a	24781	as	5701
in	21047	from	5438
and	20449	with	5357
for	10454	million	5335
that	10422	was	4901
is	8539	be	4586
it	7195	its	4571
said	7141	are	4528

Enumerating Closed-Class Words

- Pronouns / determiners / articles in all cases
  - Personal: *I, me, you, he, him, she, her, it, we, us, they, them*
  - Impersonal: *one* (as in *One has to be careful here.*)
  - Reflexive: *myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself*
  - Possessive: *my, mine, your, yours, his, her, hers, its, our, ours, their, theirs*
  - Demonstrative: *this, these, that, those*
  - Article: *the, a, an*
  - Interrogative / relative: *who, whom, whose, what, which, whoever, whomever, whatever*
  - Indefinite: *some, somebody, someone, something, any, anybody, anyone, anything; many, much, more, most, too, enough, few, little, fewer, less, fewest, least*
  - Total: *every, everybody, everyone, everything, each, all, both*
  - Negative: *no, nobody, nothing, none*

🇬🇧 Enumerating Closed-Class Words

- Numerals
  - Cardinal
    - *zero, one, two, three, four, five, six, seven, eight, nine, ten*
    - *eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen*
    - *twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety*
    - *hundred, thousand, million, billion*
  - Ordinal
    - *first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth* ⇒ morphology *-th*
  - In some languages written as one word, i.e., a nice morphological exercise:
    - 361,972
    - 🇬🇧 `en:` three hundred sixty-one thousand nine hundred and seventy-two
    - 🇩🇪 `de:` dreihunderteinundsechzigtausendneunhundertzweiundsiebzig

Enumerating Closed-Class Words

- Auxiliary and modal verbs
  - *be, am, are, is, was, were, been, being, ’m, ’s, ’re, aren’t, isn’t, wasn’t, weren’t, ain’t*
  - *have, has, had, having, ’ve, ’s, ’d, haven’t, hasn’t*
  - *will, would, (willing), ’ll, ’d, won’t, wouldn’t*
  - *can, cannot, could, couldn’t*
  - *shall, should, shouldn’t*
  - *may, might*
  - *must*
  - *do, does, did, done, doing, don’t, doesn’t*

Enumerating Closed-Class Words

- Pronominal adverbs
  - Demonstrative: *here, there, now, then*
  - Interrogative / relative: *where, when, how, why*
  - Indefinite: *somewhere, sometime, sometimes, somehow, anywhere, anytime, anyhow, anyway, anyways*
  - Total: *everywhere, always*
  - Negative: *nowhere, never*

Enumerating Closed-Class Words

- Prepositions ($>$60; tagged corpus?)
  - *aboard, about, above, across, after, against, ago, along, alongside, amid, among, amongst, around, as, astride, at, atop, before, behind, below, beneath, beside, besides, between, beyond, by, despite, de, down, during, en, except, for, from, in, inside, into, lest, like, minus, near, next, notwithstanding, of, off, on, onto, opposite, out, outside, over, par, past, per, plus, post, since, through, throughout, 'til, till, to, toward, towards, under, underneath, unlike, until, unto, up, upon, versus, via, vs., with, within, without, worth*

  -
    ```
    grep 'IN' wsj.mrg | perl -pe 's/^.*?\(IN(.*?)\).*$/$1/; $_=lc($_)' |
    sort -u | less
    ```

Enumerating Closed-Class Words

- Conjunctions
  - Coordinating: *and, both, but, either, et, less, minus, 'n, 'n', neither, nor, or, plus, so, times, v., versus, vs., yet*
  - Subordinating: *albeit, although, because, 'cause, if, neither, since, so, than, that, though, 'til, till, unless, until, whereas, whether, which, while*
- Particles
  - *yes, no, not, n't, to* (infinitival)

- Found in corpus:
  - 256 closed-class types (out of 307 anticipated, resp. 289 unique anticipated)
  - 413,914 occurrences (33% of total tokens)

Open-Class Words

- Now there is a nice list of some 32,000 open-class words. What remains is to read them all and sort them out manually ☺
    - (exactly: 31,962 types, covering 525,556 tokens)

    - Nouns (including proper nouns)
    - Adjectives (including those derived from proper nouns)
    - Verbs (except for auxiliaries and modals)
    - Adverbs
    - (Interjections)

- What else can help us?

Most Frequent Open-Class Words

said	7141	shares	1444
new	3257	president	1431
company	3078	years	1426
year	2753	trading	1415
market	2648	sales	1331
says	2467	only	1188
stock	2002	business	1171
also	1867	such	1164
other	1808	york	1129
share	1798	group	1102
last	1482	time	1032

Most Frequent Open-Class Words

said	7141
new	3257
company	3078
year	2753
market	2648
says	2467
stock	2002
also	1867
other	1808
share	1798
last	1482
shares	1444
president	1431
years	1426
trading	1415
sales	1331
only	1188
business	1171
such	1164
york	1129
group	1102
time	1032

Plural Nouns / 3rd Person Verbs

year	2753	years	1426	4179
company	3078	companies	1020	4098
new	3257	news	424	3680
say	878	says	2467	3345
market	2648	markets	621	3269
stock	2002	stocks	800	2802
other	1808	others	263	2071
price	929	prices	1016	1945
sale	483	sales	1331	1814
last	1482	lasts	8	1490
month	624	months	844	1468
president	1431	presidents	22	1453
business	1171	businesses	267	1438

Total 4448 pairs

Gerunds / Present Participles

market	2648	marketing	211	2859
stock	2002	stocking	2	2004
trade	525	trading	1415	1940
share	1798	sharing	9	1807
last	1482	lasting	9	1491
bank	955	banking	220	1175
time	1032	timing	33	1065
say	878	saying	172	1050
make	739	making	286	1025
price	929	pricing	59	988
sell	603	selling	353	956 ...
even	905	evening	35	940
get	572	getting	201	773 ...

Total 1927 pairs

Tagged Corpus Available?

- Having a tagged corpus does not necessarily mean we have a morphological analyzer, so it still could make sense to construct one
- Now it's trivial to distinguish nouns from verbs, adjectives etc., even if they overlap
- Still, we may need some information not encoded in the tags

- Example: declension class ("pattern") of 🇨🇿 Czech nouns:
    - `NNF*` = **feminine noun** ⇒ 4 declension classes:

žena "woman"	*-a*	*-y*	*-ě*	*-u*	*-o*	*-ě*	*-ou*	*-y*	*-0*	*-ám*	*-y*	*-y*	*-ách*	*-ami*
růže "rose"	*-e*	*-e*	*-i*	*-i*	*-e*	*-i*	*-í*	*-e*	*-í*	*-ím*	*-e*	*-e*	*-ích*	*-emi*
píseň "song"	*-0*	*-ě*	*-i*	*-0*	*-i*	*-i*	*-í*	*-ě*	*-í*	*-ím*	*-ě*	*-ě*	*-ích*	*-ěmi*
kost "bone"	*-0*	*-i*	*-i*	*-0*	*-i*	*-i*	*-í*	*-i*	*-í*	*-em*	*-i*	*-i*	*-ech*	*-mi*

And So On...

- Using similar heuristics, gradually classify more and more word forms
  - Obviously, not everything can be captured this way
    - Some sets of pairs have multiple interpretations
    - For some words no heuristics exist
    - Or the other member of the pair has not occurred in the corpus
- Semi-supervised:
  - You don't know what word form belongs where
  - However, you know how the suffixes look like
- Unsupervised:
  - You don't even know the set of affixes
  - However, you know (or assume) that the morphology is concatenative (prefix* stem+ suffix*)
  - Look at the corpus, try to find regularities

Unsupervised Morphemic Segmentation

- Morpho Challenge (shared task) since 2005
- Linguistica (John A. Goldsmith) (`http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/`)
- Morfessor (Mathias Creutz & Krista Lagus) (`http://www.cis.hut.fi/projects/morpho/`)
- ParaMor (Christian Monson) (`http://www.cslu.ogi.edu/~monsonc/ParaMor.html`)
- Affisix (Michal Hrušecký, MFF)
- Morseus (Daniel Zeman, MFF) (`http://ufal.mff.cuni.cz/~zeman/projekty/morseus/`)
- And many others...

Homework

- Pick a language, get data
- Extract lexicon
- Details and data:
  `http://ufal.mff.cuni.cz/~zeman/vyuka/morfosynt/lab-lexicon/index.html`
- Deadline:
  Wednesday November 13, 23:59 CET