Caution

• No class on

October 22

&

October 29
Organizational

- Your project should be in the wiki by now
- Your projects next three weeks:
  - Download corpora (or write me if I told you I have them)
  - Think about the best way of acquiring a lexicon from the corpus; start working on it
Projects until November 4

- Acquire from corpus as good a lexicon as possible
  - How many word types are there?
  - How many can we categorize (POS)
  - Are there different declension / conjugation classes? Are we able to assign them to words?
- Tagset: find existing, adapt or design a new one
  - What parts of speech (and subclasses) exist in the language / are we going to recognize?
  - What categories (gender, number, case…) for each POS?
- What types of productive derivational morphology can be covered?
- Keep this information for your final presentation.
Parts of Speech

Daniel Zeman

http://ufal.mff.cuni.cz/daniel-zeman/
zman@ufal.mff.cuni.cz
Part of Speech

• Vague definitions, criteria of mixed nature

• **Looong tradition**… (difficult to change)
  
  – Traditional linguistics:
    • Classification differs cross-linguistically!
    • (Even among established classes, not just endemic minor parts of speech.)

  – Computational linguistics (tagsets):
    • Dozens of classes and subclasses
    • Significant differences even within one language
History

- 4th century BC: Sanskrit
- European tradition (prevailing in modern linguistics): Ancient Greek
  - Plato (4th century BC): sentence consists of nouns and verbs
  - Aristotle added “conjunctions” (included conjunctions, pronouns and articles)
  - End of 2nd century BC: classification stabilized at 8 categories (Διονύσιος ο Θράξ: Τέχνη Γραμματική / Dionysios o Thrax: Art of Grammar)
Ancient Greek Word Classes

- **Noun** (Ουσιαστικό *ousiastiko*)
  - inflected for case, signifying a concrete or abstract entity
- **Verb** (Ρήµα *rîma*)
  - without case inflection, but inflected for tense, person and number, signifying an activity or process performed or undergone
- **Participle** (Μετοχή *metohî*)
  - sharing the features of the verb and the noun
- **Interjection** (Επιφώνηµα *epifônîma*)
  - expressing emotion alone
- **Pronoun** (Αντωνυµία *antônymia*)
  - substitutable for a noun and marked for person
- **Preposition** (Πρόθεση *prothesî*)
  - placed before other words in composition and in syntax
- **Adverb** (Επίρρηµα *epirrîma*)
  - without inflection, in modification of or in addition to a verb
- **Conjunction** (Σύνδεσµος *syndesmos*)
  - binding together the discourse and filling gaps in its interpretation
Where Are Adjectives?

- The best matching Ancient Greek definition is that of nouns, and perhaps participles.
- Adjectives are a relatively new (1767) invention from France:
Traditional English Parts of Speech

1. Noun
2. Verb
3. Adjective
4. Adverb
5. Pronoun
6. Preposition
7. Conjunction
8. Interjection

“Traditional” means: taught in elementary schools, marked in dictionaries.

Linguists (and especially computational linguists) may see other categories, e.g. determiners.
Traditional Czech Parts of Speech

1. Noun *(podstatné jméno, substantivum)*
2. Adjective *(přídavné jméno, adjektivum)*
3. Pronoun *(zájmeno)*
4. Numeral *(číslovka)*
5. Verb *(sloveso)*
6. Adverb *(příslovce, adverbiunm)*
7. Preposition *(předložka)*
8. Conjunction *(spojka)*
9. Particle *(částice)*
10. Interjection *(citoslovce)*
A Mixture of Criteria

• Parts of speech are defined on the basis of morphological, syntactic and semantic criteria
• In many cases they are just rough approximation
• Because of long tradition in some languages, it is difficult to redesign the system
• Sets of POS tags strive to
  – keep reasonable consistency with tradition
  – partition the word space systematically
Morphological Criteria

- By definition language-dependent. In Czech (simplified):
  - Nouns: (gender), number, case. Include some pronouns (někdo) and numerals (pět, tisíc, sedmero, polovina)
  - Adjectives: gender, number, case, sometimes degree; agr. with N. Include some pronouns (který, žádný) and numerals (první, druhý, čtverý)
  - Personal pronouns: person, gender, number, case
  - Possessive pronouns: possessor’s person, gender & number; possessed gender & number
  - Verbs:
    - infinitive
    - finite: mood (indicative/imperative), tense (present/future), person, number
    - participle: voice (active/passive), gender, number
    - transgressive: tense (present/past), gender, number
  - Non-inflectional words
Syntactic Criteria

- Slightly less language-dependent
  - Nouns: arguments of verbs (subject, object), nominal predicate \((he\ is\ a\ teacher)\) etc. Also attribute of other nouns. Include personal pronouns \((I,\ you)\), some numerals in some languages.
  - Adjectives: modify noun phrases.
  - Verbs: predicates of clauses.
  - Adverbs: modify verbs, usually as adjuncts (non-obligatory).
  - Prepositions: govern noun phrases, dictate their case, semantically modify their relation to verbs or other nouns.
  - Coordinating conjunctions \((and,\ or,\ but)\).
  - Subordinating conjunctions \((that)\): join dependent to main clause.
  - Relative \((not\ interrogative)\) pronouns \((which)\): merger of nouns/adjectives and subordinating conjunctions.
Syntactic Nouns

- Arguments of verbs (subject, object), nominal predicate (*he is a teacher*) etc.
- Attributes of other nouns (**CS**: *auto prezidenta* = *president’s car*)
  - **en**: *Christmas present*: is *Christmas* a syntactic adjective or noun?
  - Even if definitions are purely syntactic, consensus across languages is not guaranteed because every language has its own set of syntactic constructions
- Including
  - pronouns: personal (*I, you, he, we*), indefinite (*somebody*), negative (*nothing*), totality (*everyone*), some demonstratives (*this in this is ridiculous*)
  - **CS**: some numerals in some cases (*pět, deset, tisíc, miliarda, třetina, sedminásobek, desatero*)
Syntactic Adjectives

• Modify a noun phrase, typically agree with it in gender, number and case. Include:
  – Possessive pronouns (determiners?) (*my, your, his, our*)
  – Demonstrative pronouns in some contexts (*this apple is sweet*)
  – Some indefinite and other pronouns in some languages (*CS: nějaký (some), každý (every), žádný (no]*) (in other languages these may not be traditionally considered pronouns)
  – Cardinal numerals (but see next slide) (*one, two, three*)
  – Adjectival ordinal numerals (*first, second, third*)
  – Adjectivally used participles (*traveling salesman, mixed feelings*)
  – Possibly even adjectivally used nouns (*Christmas present, car repair, New York Times advisory board member*)
Syntactic Behavior of Czech Cardinal Numerals

- *jeden* (one), *dva* (two), *tři* (three), *čtyři* (four) are syntactic adjectives. They agree in case (and also gender and number) with the counted noun.
- *pět* (five) and higher may behave as syntactic nouns:
  - whole phrase in nominative / accusative / vocative: the numeral governs the counted noun, forces it to genitive: *pět /nom židlí* (five chairs) /gen, not *pět *židle /nom ⇒ pět is syntactic noun
  - whole phrase in other cases: the numeral agrees in case with the counted noun ⇒ it modifies the noun: *k pěti/dat židlím/dat* (to five chairs) ⇒ pěti is a syntactic adjective
- *tisíc* (thousand), *milion* (million), *miliarda* (billion) in both Czech and English can be used as:
  - nouns (morphologically and syntactically): *z banky zmizely milióny = millions vanished from a bank*
  - traditional numerals, syntactic nouns: *dluží mi milión dolarů = he owes me one million dollars*
Syntactic Verbs

• Predicate of a main clause
• Predicate of a dependent clause
• Auxiliary verb, modal verb or another part of a complex verb form:
  - en: would have been willing (to) keep smiling 😊
  - cs: bych byl býval mohl chtít udělat
    (= (I) could have wanted to do)
• Copula in nominal predicates:
  - en: he is a teacher
Syntactic Adverbs

• Modify verbs, optionally specify circumstances such as location, time, manner, extent, cause…
• Can also modify adjectives (*very large*) or other adverbs (*very well*)
• Including:
  – some ordinal numerals: cs: *poprvé* (*for the first time*)
  – multiplicative numerals: cs: *dvakrát* (*twice*), *pětasedmdesátkrát* (*seventy-five times*)
  – transgressives: cs: *čekajíc na autobus všimla si ho* (*she noticed him while waiting for a bus*); hi: दरवाजा खोलकर वह कमरे में आई *darvāzā kholkar vah kamre mē āī* (*having opened the door she came in*)
Conjunctions

• Coordinating conjunctions join phrases of same or similar type or even whole clauses (independent)
  – single coordinators:
    • *Peter and Paul; today or tomorrow; he wanted to go but she didn’t like the idea*
  – paired coordinators:
    • *neither here nor there; the sooner the better; as soon as possible*

• Subordinating conjunctions join dependent clauses or phrases to the governing node, specifying their function
  – single subordinators:
    • *that; so; if; whether; because*
  – paired subordinators:
    • *hi: जब मैं कहूँगा तब आना jab maĩ kahūgā tab ānā (lit: when I tell then come)*
Relative Pronouns, Determiners, Numerals and Adverbs

- Merge properties of syntactic nouns / adjectives / adverbs and of subordinating conjunctions
  - relative syntactic noun: *those who know*; a car *that never breaks*; *the man whom I met*; who *knows what you find*
  - relative syntactic adjective: *the man whose son is this boy*; you *decide from what time on you work*; …*which color you like*
  - **cs**: relative numerals: *pověz mi, kolik máš peněz* (tell me *how much money you have*); …*kolikátý jsi byl* (where did you rank; lit. *how-many-th you were*)
  - relative syntactic adverb: *I don’t know when she came*; …*where it is*; …*how to say*; …*why he’s here*
- Interrogative pronouns (adverbs etc.) may have same form (in some languages) but not the same joining function.
Adpositions

- Govern syntactic noun (dictate its case marking), specify its role as argument of
  - a verb (*believe in* something)
  - another noun (*lack of* something)
  - or adjective (*acceptable for* me)

- Appear before, after or around the noun phrase:
  - Preposition: *in the house; under the table; beyond this point*
  - Postposition: *हिसकमरेमें kamre mē* (lit. *room in*)
  - Circumposition: *de: von diesem Zeitpunkt an (from this moment on)*
Semantic Criteria

- **Semantic noun**: a concrete or abstract entity
  - **cs**: otcův (*father’s*) is traditionally a possessive adjective but could be regarded as a form of the semantic noun *otec* (*father*); not to confuse with genitive case *otec/otců*

- **Semantic adjective**: a quality, property
  - **en**: *cleverly* could be regarded as a form of the semantic adjective *clever*
  - How far should we go? Is *cleverness* an adjective, too? What purpose would such classification serve?

- **Semantic adverb**: a circumstance (location, time, manner)
  - **cs**: traditional adjective *zířejší* could be regarded as a form of the semantic adverb *zítra* (*tomorrow*)

- **Semantic verb**: a state or an action
  - **cs**: deverbative nouns (*dělání = the doing*) and adjectives (*dělající = doing; udělavší = the one that did; udělaný = done*) could be regarded as forms of the semantic verb

- **Pronoun**: any referential word (trad. pronoun, determiner, numeral, adverb / personal, possessive, indefinite, absolute, negative, interrogative, relative, demonstrative)

- **Numeral**: a number, amount (*one, two, three; first, second, third; once, twice, thrice; twofold; pair, triple, quadruple*)

- **Adpositions + conjunctions + particles + auxiliaries (glue material)**
Openness vs. Closeness

- **Open classes (take new words)**
  - verbs (non-auxiliary), nouns, adjectives, adjectival adverbs, interjections
  - word formation (derivation) across classes
- **Closed classes (words can be enumerated)**
  - pronouns / determiners, adpositions, conjunctions, particles
  - pronominal adverbs
  - auxiliary and modal verbs
  - numerals (mathematically infinite, linguistically closed)
  - typically they are not base for derivation
- **Even closed classes evolve but over longer period of time**
  - *es*: *Vuestra Merced* (*Your Mercy, Your Grace*) $\Rightarrow$ *usted* (new singular 2nd person pronoun in formal/honorific register)
Universal POS Tags

Lexicon Acquisition

Daniel Zeman

zeman@ufal.mff.cuni.cz
Lexicon Acquisition

- Some hints only (approach must vary greatly depending on language)
- Identify part of speech and inflection pattern
- If affixes restrict possible classes, use it!
  - E.g. in Czech, the following suffixes increase likelihood of an infinitive: -st, -át, -at, -ct, -ci, -ít, -out, -ýt, -ovat, -ít, -ět, -et
  - English does not inflect but verb forms and derivational suffixes (-ness, -ity, -able) can help
- Otherwise, syntax might help
  - E.g. if it’s after preposition or an 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
  - A textbook and/or a bilingual dictionary may help with the rest
  - Parallel corpus + word aligner may supplement the dictionary (Addicter)
- 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 Acquisition

• Example only! Other languages and corpora may require a different approach.

• Input: a plain-text corpus (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
When it's time for their biannual powwow, the nation's manufacturing titans typically jet off to the sunny confines of resor
English Frequency Wordlist

- Penn Treebank 3 / Wall Street Journal:
- 49,208 sentences
- 1,273,255 terminal nodes (tokens and traces)
- 52,494 word types (opposed to word occurrences) including traces
- 46,074 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

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>60484</td>
</tr>
<tr>
<td>the</td>
<td>59459</td>
</tr>
<tr>
<td>.</td>
<td>48144</td>
</tr>
<tr>
<td>to</td>
<td>29576</td>
</tr>
<tr>
<td>of</td>
<td>28441</td>
</tr>
<tr>
<td>a</td>
<td>24781</td>
</tr>
<tr>
<td>in</td>
<td>21257</td>
</tr>
<tr>
<td>and</td>
<td>20449</td>
</tr>
<tr>
<td>'s</td>
<td>11556</td>
</tr>
<tr>
<td>for</td>
<td>10454</td>
</tr>
<tr>
<td>that</td>
<td>10422</td>
</tr>
<tr>
<td>$</td>
<td>8817</td>
</tr>
<tr>
<td>`` (&quot;&quot;</td>
<td>8735</td>
</tr>
<tr>
<td>is</td>
<td>8539</td>
</tr>
<tr>
<td>&quot; (&quot;</td>
<td>8506</td>
</tr>
<tr>
<td>it</td>
<td>7195</td>
</tr>
<tr>
<td>said</td>
<td>7141</td>
</tr>
<tr>
<td>on</td>
<td>6646</td>
</tr>
<tr>
<td>%</td>
<td>6121</td>
</tr>
<tr>
<td>at</td>
<td>5770</td>
</tr>
<tr>
<td>by</td>
<td>5705</td>
</tr>
<tr>
<td>as</td>
<td>5701</td>
</tr>
</tbody>
</table>
# Punctuation and Special Characters

$m/\backslash pP/$

<table>
<thead>
<tr>
<th>Character</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>60484</td>
</tr>
<tr>
<td>.</td>
<td>48144</td>
</tr>
<tr>
<td>'s</td>
<td>11556</td>
</tr>
<tr>
<td>$</td>
<td>8817</td>
</tr>
<tr>
<td>`` (“)</td>
<td>8735</td>
</tr>
<tr>
<td>&quot; (”</td>
<td>8506</td>
</tr>
<tr>
<td>%</td>
<td>6121</td>
</tr>
<tr>
<td>mr. (tokenization?)</td>
<td>4950</td>
</tr>
<tr>
<td>n’t</td>
<td>4006</td>
</tr>
<tr>
<td>--</td>
<td>2505</td>
</tr>
<tr>
<td>u.s.</td>
<td>2056</td>
</tr>
<tr>
<td>third-quarter</td>
<td>333</td>
</tr>
<tr>
<td>buy-out</td>
<td>222</td>
</tr>
<tr>
<td>s&amp;p</td>
<td>166</td>
</tr>
<tr>
<td>3,000</td>
<td>28</td>
</tr>
<tr>
<td>3.7</td>
<td>28</td>
</tr>
<tr>
<td>total types</td>
<td>12888</td>
</tr>
<tr>
<td>the rest</td>
<td>33186</td>
</tr>
<tr>
<td>Caught, OK</td>
<td></td>
</tr>
<tr>
<td>Not caught (but should have been caught)</td>
<td></td>
</tr>
<tr>
<td>Caught (disputable)</td>
<td></td>
</tr>
<tr>
<td>Caught (we want better tokenization)</td>
<td></td>
</tr>
</tbody>
</table>

*Caught, OK*

*Not caught (but should have been caught)*

*Caught (disputable)*

*Caught (we want better tokenization)*
<table>
<thead>
<tr>
<th>Numbers</th>
<th>1203</th>
<th>105...</th>
<th>8416</th>
<th>37658</th>
<th>32237</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1203</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>673</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>610...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>503...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>111...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>105...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>88...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-year</td>
<td>79...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980s</td>
<td>53...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ru-486</td>
<td>15...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mid-1980s</td>
<td>12...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b-2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19th</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1989-90</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80%-owned</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xr4ti</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total types</td>
<td>8416</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the rest</td>
<td>37658</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no punctuation or numbers</td>
<td>32237</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Real Words

\( \text{\textbf{!m/ [\backslash pP\backslash pN` $] /}} \)

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>59459</td>
<td>on</td>
<td>6646</td>
</tr>
<tr>
<td>to</td>
<td>29576</td>
<td>at</td>
<td>5770</td>
</tr>
<tr>
<td>of</td>
<td>28441</td>
<td>by</td>
<td>5705</td>
</tr>
<tr>
<td>a</td>
<td>24781</td>
<td>as</td>
<td>5701</td>
</tr>
<tr>
<td>in</td>
<td>21257</td>
<td>from</td>
<td>5438</td>
</tr>
<tr>
<td>and</td>
<td>20449</td>
<td>with</td>
<td>5357</td>
</tr>
<tr>
<td>for</td>
<td>10454</td>
<td>\textbf{million}</td>
<td>5335</td>
</tr>
<tr>
<td>that</td>
<td>10422</td>
<td>was</td>
<td>4901</td>
</tr>
<tr>
<td>is</td>
<td>8539</td>
<td>be</td>
<td>4586</td>
</tr>
<tr>
<td>it</td>
<td>7195</td>
<td>its</td>
<td>4571</td>
</tr>
<tr>
<td>said</td>
<td>7141</td>
<td>are</td>
<td>4528</td>
</tr>
</tbody>
</table>
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
  - Indefinite: some, somebody, someone, something, any, anybody, anyone, anything, every, everybody, everyone, everything, each, all, both; many, much, more, most, too, enough, few, little, fewer, less, least
  - 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, ..., nineteen
    • twenty, thirty, forty, sixty, seventy, eighty, ninety
    • hundred, thousand, million, billion
  – Ordinal
    • first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth \( \Rightarrow \) morphology “-th”
  – In some languages written as one word, i.e. nice morph. 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*
  - *have, has, had, having, ’ve, ’s, ’d*
  - *will, would, (willing), ’ll, ’d*
  - *can, cannot, could*
  - *shall, should*
  - *may, might*
  - *must*
  - *do, does, did, done, doing*
Enumerating Closed-Class Words

- Pronominal adverbs
  - Demonstrative: here, there, now, then
  - Interrogative / relative: where, when, how, why
  - Indefinite: somewhere, sometime, somehow, anywhere, anytime, anyhow, anyway, everywhere, always
  - Negative: nowhere, never
- 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 | more
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:**
  - 263 closed-class types (out of 289 anticipated)
  - 419,915 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 😊
  – 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 OC Words

- said 7141
- new 3257
- company 3078
- year 2753
- market 2648
- says 2467
- pos 2423
- tags 2317
- stock 2002
- also 1867
- other 1808
- share 1798
- last 1482
- shares 1444
- president 1431
- years 1426
- trading 1415
- sales 1331
- fixing 1195
- only 1188
- business 1171
- such 1164
Most Frequent OC Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>said</td>
<td>7141</td>
</tr>
<tr>
<td>new</td>
<td>3257</td>
</tr>
<tr>
<td>company</td>
<td>3078</td>
</tr>
<tr>
<td>year</td>
<td>2753</td>
</tr>
<tr>
<td>market</td>
<td>2648</td>
</tr>
<tr>
<td>says</td>
<td>2467</td>
</tr>
<tr>
<td>pos</td>
<td>2423</td>
</tr>
<tr>
<td>tags</td>
<td>2317</td>
</tr>
<tr>
<td>stock</td>
<td>2002</td>
</tr>
<tr>
<td>also</td>
<td>1867</td>
</tr>
<tr>
<td>other</td>
<td>1808</td>
</tr>
<tr>
<td>share</td>
<td>1798</td>
</tr>
<tr>
<td>last</td>
<td>1482</td>
</tr>
<tr>
<td>shares</td>
<td>1444</td>
</tr>
<tr>
<td>president</td>
<td>1431</td>
</tr>
<tr>
<td>years</td>
<td>1426</td>
</tr>
<tr>
<td>trading</td>
<td>1415</td>
</tr>
<tr>
<td>sales</td>
<td>1331</td>
</tr>
<tr>
<td>fixing</td>
<td>1195</td>
</tr>
<tr>
<td>only</td>
<td>1188</td>
</tr>
<tr>
<td>business</td>
<td>1171</td>
</tr>
<tr>
<td>such</td>
<td>1164</td>
</tr>
</tbody>
</table>
## Plurals / 3\textsuperscript{rd} Person Verbs

<table>
<thead>
<tr>
<th>Word</th>
<th>Singular</th>
<th>Plural</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>2753</td>
<td>years</td>
<td>4179</td>
</tr>
<tr>
<td>new</td>
<td>3257</td>
<td>news</td>
<td>3680</td>
</tr>
<tr>
<td>say</td>
<td>878</td>
<td>says</td>
<td>3345</td>
</tr>
<tr>
<td>market</td>
<td>2648</td>
<td>markets</td>
<td>3269</td>
</tr>
<tr>
<td>stock</td>
<td>2002</td>
<td>stocks</td>
<td>2802</td>
</tr>
<tr>
<td>pos</td>
<td>2423</td>
<td>poses</td>
<td>2428</td>
</tr>
<tr>
<td>po</td>
<td>1</td>
<td>pos</td>
<td>2424</td>
</tr>
<tr>
<td>tag</td>
<td>7</td>
<td>tags</td>
<td>2324</td>
</tr>
<tr>
<td>other</td>
<td>1808</td>
<td>others</td>
<td>2071</td>
</tr>
<tr>
<td>last</td>
<td>1482</td>
<td>lasts</td>
<td>1490</td>
</tr>
<tr>
<td>month</td>
<td>624</td>
<td>months</td>
<td>1468</td>
</tr>
<tr>
<td>president</td>
<td>1431</td>
<td>presidents</td>
<td>1453</td>
</tr>
<tr>
<td>business</td>
<td>1171</td>
<td>businesses</td>
<td>1438</td>
</tr>
</tbody>
</table>

**Total 3246 pairs**
### Gerunds / Present Participles

<table>
<thead>
<tr>
<th>Gerund</th>
<th>Present Participle</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>market</td>
<td>marketing</td>
<td>2859</td>
</tr>
<tr>
<td>pos</td>
<td>posing</td>
<td>2429</td>
</tr>
<tr>
<td>stock</td>
<td>stocking</td>
<td>2004</td>
</tr>
<tr>
<td>trade</td>
<td>trading</td>
<td>1940</td>
</tr>
<tr>
<td>share</td>
<td>sharing</td>
<td>1807</td>
</tr>
<tr>
<td>last</td>
<td>lasting</td>
<td>1491</td>
</tr>
<tr>
<td>fix</td>
<td>fixing</td>
<td>1203</td>
</tr>
<tr>
<td>bank</td>
<td>banking</td>
<td>1175</td>
</tr>
<tr>
<td>say</td>
<td>saying</td>
<td>1050</td>
</tr>
<tr>
<td>make</td>
<td>making</td>
<td>1025</td>
</tr>
<tr>
<td>price</td>
<td>pricing</td>
<td>988</td>
</tr>
<tr>
<td>even</td>
<td>evening</td>
<td>940</td>
</tr>
<tr>
<td>get</td>
<td>getting</td>
<td>773</td>
</tr>
</tbody>
</table>

**Total 1842 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“ -a, -y, -ě, -u, -o, -ě, -ou, -y, -0, -ám, -y, -y, -ách, -ami
    • „růže“ -e, -e, -i, -e, -i, -e, -i, -ím, -e, -e, -ích, -emi
    • „píseň“ -0, -ě, -i, -0, -i, -i, -ě, -i, -ím, -ě, -ě, -ích, -ěmi
    • „kost“ -0, -i, -i, -0, -i, -i, -i, -i, -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/](http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/))
- ParaMor (Christian Monson) ([http://www.cslu.ogi.edu/~monsonc/Paramor.html](http://www.cslu.ogi.edu/~monsonc/Paramor.html))
- Affisix (Michal Hrušekčy, MFF)
- And many others…