Tagsets and Parts of Speech
(continued)

Daniel Zeman

http://ufal.mff.cuni.cz/course/npfl094/
zeman@ufal.mff.cuni.cz
Review: The Big Four

• Nouns
  – Proper nouns

• Verbs
  – Participles (between verbs and nouns / adjectives / adverbs)

• Adjectives
  – Modify nouns

• Adverbs
  – Modify verbs, adjectives or adverbs
Review: Common Minors

- Adpositions
  - Prepositions
  - Postpositions
  - Circumpositions
- Conjunctions
  - Subordinators
  - Coordinators
- Interjections
- Particles (often “garbage can category”)

Review: Pronouns vs. Determiners

• In some tagsets clear (but context-dependent) definition:
  – Pronouns replace noun phrases
    • *I, you, he, she, it, we, they, who, something…*
    • *This is unbelievable!*
    • *Yours is better.*
  – Determiners modify noun phrases (so they include traditional possessive “pronouns”)
    • *This book is John’s.*
    • *Your book is better.*

• Some traditional grammars (and tagsets) refer to all the above as pronouns (e.g. Czech)
BulTreeBank Tagset (bg): The Broadest Sense of Pronouns

- **Subcategories:**
  - personal: аз, ти, мой
  - possessive: мой, моя, твой, негов
  - demonstrative: този, тоя
  - interrogative: кой, коя, кое
  - relative: който, що
  - collective: всеки, всякой
  - indefinite: един, някой
  - negative: никой, никакъв

- **Referential type:**
  - entity: кой, коя, кое
  - attribute: какъв, каква, какво
  - possession: чий, чия, чие
  - quantity: колко, доколко
  - location: къде, где, докъде
  - time: кога, докога, откога
  - manner: как
  - cause: защо
Numerals vs. Adjectives

• Many tagsets distinguish *cardinal numbers*
  – While some (Danish) take them as special class of adjectives

• Ordinal numbers
  – Sometimes separate POS
  – Sometimes special class of adjectives
  – Sometimes “normal” adjectives (undistinguished)
Prague Dependency Treebank (CS): The Greatest Variety of Numerals

- **Cardinal:** jeden, dva, tři, čtyři, pět
- **Adjectival ordinal:** první, druhý, třetí, čtvrtý, pátý
- **Adverbial ordinal:** poprvé, podruhé
- **Multiplicative:** jedenkrát, dvakrát
- **Generic (N sets of):** jedny, dvoje, troje, čtvery, patery
- **Generic (N sorts of):** dvojí, trojí, čtverý, paterý
- **Generic (N-tuple):** dvé, tré, čtvero, patero
  - But noun n-tuple: dvojice, trojice
- **Fraction:** polovina, třetina, čtvrťina
- **Number Arabic digits:** 1, 2, 3, 4, 5
- **Number Roman:** I, II, III, IV, V

**Pronominal type:**

- **Interrogative / relative:** kolik, kolikátý, pokolikáté, kolikrát, kolikery, kolikerý, kolikero
- **Indefinite:** několik, několikátý, poněkolikáté, několikrát, několikery, několikerý, několikero, mnoho, málo
- **Demonstrative:** tolik, tolikátý, potolikáté, tolikrát, tolikery, tolikery, tolikero
Some Endemic Classes

• Existential *there* in English
• Infinitival marker: English *to*, German *zu*, Swedish *att*
• Predeterminer: English *both the boys, all the people*
• Response particle: *yes, no, thanks*
• Negative particle: *not, n’t, Arabicَلاَ, lā*
• Separable verbal prefix: German *vorstellen ⇒ stellen Sie sich vor*
• Adjectival particle: German *am besten, zu groß*
• Measure word: Chinese 一个人 = *yí gè rén* = “one (piece) man”
Various Other Classes

• Foreign words (foreign-language quotations, names of books etc.; not loanwords!)
  – The police confiscated illegal copies of the banned *Mein Kampf* by Adolf Hitler.
  – Could be subclassified as foreign nouns, verbs etc.
  – POS and features need not be the same as in the source language!
    • German *Burg* is feminine. If embedded in Czech it will be treated as masc.

• Abbreviations
  – Could be subclassified as abbreviated nouns, verbs etc.

• Parts of multi-token idioms

• Numbers (*123*)
• Symbols (*$*, *€*)
• Punctuation (*, . – “ ”*)
Clitics

- Clitic is a
  - Syntactically independent word
  - Phonologically / orthographically dependent morpheme
- es: despiértate = wake yourself; démelo = give me it;
- ru: защищаться = zaščiščat’ja = to defend oneself
- de: zum = zu dem = to the; am = an dem = on the;
- fr: du = de le = of the
- cs: proň = pro něj = for him; oč = o co = for what; tys = ty jsi = you have; žes = že jsi = that you have; scvrnkls = scvrnkl jsi = you flicked off; přišel’ = nebot’ přišel = because he came
- ar: وبالالفالوجة = wabiāl fālūjah = wa/CONJ + bi/PREP + AlfAlwp/NOUN_PROP = and in al-Falujah
Features of Nouns and Adjectives

- Gender / animateness (lexical for nouns, agreemental for adjectives) or class (Bantu languages)
- Number (singular, dual, plural)
- Case (en: 2 for pronouns; cs: 7; fi: 14)
- Definiteness (ro: poiană = a meadow, poiana = the meadow)
- Negativeness (cs: schopný = able, neschopný = unable; schopnost = ability, neschopnost = inability)
- Degree of comparison (positive, comparative, superlative, absolutive)
# Noun Classes in Swahili

<table>
<thead>
<tr>
<th>Class</th>
<th>SG</th>
<th>PL</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (humans)</td>
<td>m + tu</td>
<td>wa + tu</td>
<td>person</td>
</tr>
<tr>
<td>3 (thin objects)</td>
<td>m + ti</td>
<td>mi + ti</td>
<td>tree</td>
</tr>
<tr>
<td>5 (paired things)</td>
<td>ji + cho</td>
<td>ma + cho</td>
<td>eye</td>
</tr>
<tr>
<td>7 (instrument)</td>
<td>ki + tu</td>
<td>vi + tu</td>
<td>thing</td>
</tr>
<tr>
<td>11 (extended body parts)</td>
<td>u + limi</td>
<td>n + dimi</td>
<td>tongue</td>
</tr>
</tbody>
</table>
Features of Verbs

• Form: infinitive, participle, gerund, transgressive, supine, finite
• Mood: indicative, imperative, subjunctive, jussive, conditional
• Tense / aspect: present, past, future; continuous; aorist, imperfect, perfect, pluperfect
• Voice: active, passive
• Person: 1st, 2nd, 3rd, honorific registers
• Number: singular, dual, plural
• Gender of participles: masculine, feminine, neuter
• Negativeness: dělat = to do, nedělat = not to do
Other Features

- Case of adpositions (subcategorization, not inflection)
  - What case must the governed noun phrase be in?
- Possessor’s gender and number
  - **cs**: *jejímu psovi* = *to her dog*: feminine possessor, masculine possessed
  - **cs**: *jehož kráva* = *whose (“of which guy”) cow*: singular masculine possessor, singular feminine possessed
  - **cs**: *jejíž kráva* = *whose (“of which woman”) cow*: singular feminine possessor, singular feminine possessed
  - **cs**: *jejichž kráva* = *whose (“of which people”) cow*: plural possessor, singular possessed
Tagsets

- Tag as a set of feature (category) values … \((k_1,k_2,\ldots,k_n)\)
- Simple list of tags
  \[ T = \{t_i\}_{i=1..n} \]
- 1-1 mapping between tags and feature value space
  \[ T \leftrightarrow (K_1,K_2,\ldots,K_n) \]
- English
  - Penn TreeBank (45 tags), Brown Corpus (87), Claws c5 (62), London-Lund (197)
- Czech
  - Prague Dependency Treebank (4294; new positional, old compact), Multext-East (1458; Orwell 1984 parallel corpus), Ajka/Desam (MU Brno), Prague Spoken Corpus (over 10000!)
    - \(<i1>0</i1><i2>1</i2><i3>1</i3><i4>1</i4><i5>1</i5><i6>1</i6><i7>_</i7><i8>_</i8><i9>_</i9><i10>_</i10><i11>_</i11>\)
Czech Positional Tags of PDT

AGFS3----1A----

part of speech
subpos
poss
gender
number
negat
voice
gender
number
case
person
tense
degree
style
Parts of Speech in PDT

- N noun (podstatné jméno)
- A adjective (přídavné jméno)
- P pronoun (zájmeno)
- C numeral (číslovka)
- V verb (sloveso)
- D adverb (příslovce)
- R preposition (předložka)
- J conjunction (spojka)
- T particle (částice)
- I interjection (citoslovce)
- Z special (e.g. punctuation) (zvláštní, např. interpunkce)
- X unknown word (neznámé slovo)
Gender in PDT

<table>
<thead>
<tr>
<th>M</th>
<th>masculine animate</th>
<th>Y</th>
<th>M or I</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>masculine inanimate</td>
<td>T</td>
<td>I or F</td>
</tr>
<tr>
<td>F</td>
<td>feminine</td>
<td>W</td>
<td>I or N</td>
</tr>
<tr>
<td>N</td>
<td>neuter</td>
<td>H, Q</td>
<td>F or N</td>
</tr>
<tr>
<td>X</td>
<td>unknown</td>
<td>Z</td>
<td>M, I or N</td>
</tr>
</tbody>
</table>
### Number in PDT

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S</strong></td>
<td>singular (<em>jednotné</em>)</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>dual (<em>dvojné</em>)</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>plural (<em>množné</em>)</td>
</tr>
<tr>
<td><strong>X</strong></td>
<td>unknown (<em>neznámé</em>)</td>
</tr>
</tbody>
</table>
## Case in PDT

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nominative</td>
</tr>
<tr>
<td>2</td>
<td>genitive</td>
</tr>
<tr>
<td>3</td>
<td>dative</td>
</tr>
<tr>
<td>4</td>
<td>accusative</td>
</tr>
<tr>
<td>5</td>
<td>vocative</td>
</tr>
<tr>
<td>6</td>
<td>locative</td>
</tr>
<tr>
<td>7</td>
<td>instrumental</td>
</tr>
<tr>
<td>X</td>
<td>unknown</td>
</tr>
</tbody>
</table>
Degree, Negativeness and Person

- Degree of comparison of adjectives and adverbs:
  - 1 (positive), 2 (comparative), 3 (superlative)

- Negativeness (for V, A, D and N):
  - A (affirmative), N (negative)

- Person of verbs and pronouns:
  - 1, 2, 3
Mood, Tense and Voice

- Changes relevance of other categories (such as person and number) ⇒ in a sense, these are (sub-)parts of speech
- Tense: present (P), past (M), future (F)
- Voice: active (A), passive (P)
- Mood: indicative (N), imperative (R), conditional (C – e.g. bych)
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>other variant, less frequent</td>
</tr>
<tr>
<td>2</td>
<td>other variant, very rare, archaic or literary</td>
</tr>
<tr>
<td>3</td>
<td>very archaic or colloquial variant</td>
</tr>
<tr>
<td>5</td>
<td>colloquial, tolerated in both spoken and written discourse</td>
</tr>
<tr>
<td>6</td>
<td>colloquial, inappropriate in written discourse</td>
</tr>
<tr>
<td>7</td>
<td>colloquial like 6 but less preferred by speakers</td>
</tr>
<tr>
<td>9</td>
<td>special usage (e.g. after some prepositions)</td>
</tr>
<tr>
<td>1. <strong>CC</strong></td>
<td>coord. conj.</td>
</tr>
<tr>
<td>2. <strong>CD</strong></td>
<td>cardinal number</td>
</tr>
<tr>
<td>3. <strong>DT</strong></td>
<td>determiner</td>
</tr>
<tr>
<td>4. <strong>EX</strong></td>
<td>existential <em>there</em></td>
</tr>
<tr>
<td>5. <strong>FW</strong></td>
<td>foreign word</td>
</tr>
<tr>
<td>6. <strong>IN</strong></td>
<td>preposition or subord. conjunction</td>
</tr>
<tr>
<td>7. <strong>JJ</strong></td>
<td>adjective</td>
</tr>
<tr>
<td>8. <strong>JJR</strong></td>
<td>adj, comparative</td>
</tr>
<tr>
<td>9. <strong>JJS</strong></td>
<td>adj, superlative</td>
</tr>
<tr>
<td>10. <strong>LS</strong></td>
<td>list item marker</td>
</tr>
<tr>
<td>11. <strong>MD</strong></td>
<td>modal</td>
</tr>
<tr>
<td>12. <strong>NN</strong></td>
<td>noun, singular/mass</td>
</tr>
<tr>
<td>13. <strong>NNS</strong></td>
<td>noun, plural</td>
</tr>
<tr>
<td>14. <strong>NNP</strong></td>
<td>proper noun, sing.</td>
</tr>
<tr>
<td>15. <strong>NNPS</strong></td>
<td>proper noun, pl.</td>
</tr>
<tr>
<td>16. <strong>PDT</strong></td>
<td>predeterminer</td>
</tr>
<tr>
<td>17. <strong>POS</strong></td>
<td>possessive ending</td>
</tr>
<tr>
<td>18. <strong>PRP</strong></td>
<td>personal pronoun</td>
</tr>
<tr>
<td>19. <strong>PRP$</strong></td>
<td>poss. pronoun</td>
</tr>
<tr>
<td>Tag</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>adv, comparative</td>
</tr>
<tr>
<td>RBS</td>
<td>adv, superlative</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
</tr>
<tr>
<td>SYM</td>
<td>symbol</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base (do)</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past (did)</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund or pres. participle (doing)</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle (done)</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, non-3\textsuperscript{rd} pers. sing. present (do)</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3\textsuperscript{rd} pers. sing. present (does)</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-det. (which?)</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun (who)</td>
</tr>
<tr>
<td>WPS</td>
<td>possessive wh-pronoun (whose)</td>
</tr>
<tr>
<td>WP$</td>
<td>possessive wh-pronoun (whose)</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-adv. (where)</td>
</tr>
</tbody>
</table>

...