Tagsets, Corpora Annotation

ESSLLI 2013: Computational Morphology

Jirka Hana & Anna Feldman
Overview

- **Tagsets**
  - Types of tagsets
  - Tagset size
  - Harmonization of tags

- **Corpora annotation**
Tags and tagsets

- *(Morphological)* tag is a symbol encoding (morphological) properties of a word.
- Tagset is a set of tags.
The size of a tagset depends on a particular application as well as on language properties.

1. Penn tagset (Am. English): 36 tags; VBD – verb in past tense
2. The Lancaster-Oslo-Bergen Corpus (LOB) (Br. English): 132 tags
3. Czech positional tagset: about 4000 tags; VpNS---XR-AA--- (verb, participle, neuter, singular, any person, past tense, active, affirmative)
Types of tagsets

There are many ways to classify morphological tagsets. For our purposes, we distinguish the following three types:

1. **atomic** (*flat* in (Cloeren 1993)) – tags are atomic symbols without any formal internal structure (e.g., the Penn TreeBank tagset, (Marcus et al 1993)).

2. **structured** – tags can be decomposed into subtags each tagging a particular feature.

   - compact: Czech Compact tagsets (Hajic 2004)
   - positional – e.g., Czech Positional tagset (Hajic 2004), MULTEXT-East (Erjavec 2004, 2009, 2010)
### Penn Treebank Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordin. Conjunction</td>
<td>and, but, or</td>
<td>SYM</td>
<td>Symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>one, two, three</td>
<td>TO</td>
<td>‘to’</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>a, the</td>
<td>UH</td>
<td>Interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>Existential</td>
<td>‘there’ there</td>
<td>VB</td>
<td>Verb, base form</td>
<td>eat</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td>mea culpa</td>
<td>VBD</td>
<td>Verb, past tense</td>
<td>ate</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/sub-conj</td>
<td>of, in, by</td>
<td>VBG</td>
<td>Verb, gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>yellow</td>
<td>VBP</td>
<td>Verb, 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
<td>VBN</td>
<td>Verb, past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td>wildest</td>
<td>VBP</td>
<td>Verb, non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td>1, 2, One</td>
<td>WDT</td>
<td>Wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
<td>WP</td>
<td>Wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>llama</td>
<td>WP$</td>
<td>Possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td>llamas</td>
<td>WRB</td>
<td>Wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>IBM</td>
<td>$</td>
<td>Dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Carolinas</td>
<td>#</td>
<td>Pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both</td>
<td>“</td>
<td>Left quote</td>
<td>(‘ or “)</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>‘s</td>
<td>”</td>
<td>Right quote</td>
<td>(’ or ”)</td>
</tr>
<tr>
<td>PP</td>
<td>Personal pronoun</td>
<td>I, you, he</td>
<td>(</td>
<td>Left parenthesis</td>
<td>[ , ( , { ,&lt; ]</td>
</tr>
<tr>
<td>PP$</td>
<td>Possessive pronoun</td>
<td>your, one’s</td>
<td>)</td>
<td>Right parenthesis</td>
<td>[ ] ) , } , &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>quickly, never</td>
<td>,</td>
<td>Comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>faster</td>
<td>.</td>
<td>Sentence-final punc</td>
<td>( . !? )</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>fastest</td>
<td>:</td>
<td>Mid-sentence punc</td>
<td>( : ; ... ‘-)</td>
</tr>
</tbody>
</table>
Structured tagsets

- Any tagset capturing morphological features of richly inflected languages is necessarily large.

- A natural way to make them manageable is to use a structured system.

- In such a system, a tag is a composition of tags each coming from a much smaller and simpler atomic tagset tagging a particular morpho-syntactic property (e.g., gender or tense).
Structured tagset: benefits

1. Learnability
2. Systematic description
3. Decomposability
4. Systematic evaluation

It is trivial to view a structured tagset as an atomic tagset (e.g., by assigning a unique natural number to each tag), while the opposite is not true.
Structured tagsets: MULTEXT-East

MULTEXT-East Tagset V.4 (Erjavec 2010):

- 13 languages: English, Romanian, Russian, Czech, Slovene, Resian, Croatian, Serbian, Macedonian, Bulgarian, Persian, Finno-Ugric, Estonian, Hungarian.

- Positions’ interpretations vary across different parts of speech.
  - For instance, for nouns, position 2 is Gender, whereas for verbs, position 2 is VForm, whose meaning roughly corresponds to the mood.

- For example:
  - Ncmsn noun, common, masculine, singular, nominative;
  - Ncmsa--n noun, common, masculine, singular, accusative, indefinite, no clitic, inanimate.
Structured > Positional tagsets: Czech Positional Tagset

- Tags are sequences of values encoding individual morphological features.
- All tags have the same length, encoding all the features distinguished by the tagset.
- Features not applicable for a particular word have a N/A value.
- The – value meaning N/A or not-specified is possible for all positions except the first two (POS and SubPOS).
- SubPOS generally determines which positions are specified (with very few exceptions).
### Czech positional tagset (cont.)

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>Description</th>
<th>Example</th>
<th>Example Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>POS</td>
<td>part of speech</td>
<td>V</td>
<td>verb</td>
</tr>
<tr>
<td>2</td>
<td>SubPOS</td>
<td>detailed part of speech</td>
<td>p</td>
<td>past participle</td>
</tr>
<tr>
<td>3</td>
<td>gender</td>
<td>gender</td>
<td>N</td>
<td>neuter</td>
</tr>
<tr>
<td>4</td>
<td>number</td>
<td>number</td>
<td>S</td>
<td>singular</td>
</tr>
<tr>
<td>5</td>
<td>case</td>
<td>case</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>6</td>
<td>possgender</td>
<td>possessor’s gender</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>7</td>
<td>possnumber</td>
<td>possessor’s number</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>8</td>
<td>person</td>
<td>person</td>
<td>X</td>
<td>any</td>
</tr>
<tr>
<td>9</td>
<td>tense</td>
<td>tense</td>
<td>R</td>
<td>past tense</td>
</tr>
<tr>
<td>10</td>
<td>grade</td>
<td>degree of comparison</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>11</td>
<td>negation</td>
<td>negation</td>
<td>A</td>
<td>affirmative</td>
</tr>
<tr>
<td>12</td>
<td>voice</td>
<td>voice</td>
<td>A</td>
<td>active voice</td>
</tr>
<tr>
<td>13</td>
<td>reserve1</td>
<td>unused</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>14</td>
<td>reserve2</td>
<td>unused</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>15</td>
<td>var</td>
<td>variant, register</td>
<td>--</td>
<td>basic variant</td>
</tr>
</tbody>
</table>
Czech Positional Tagset: Wildcards

Wildcards are values that cover more than one atomic value.
Czech Positional Tagset: Wildcards

Wildcards are values that cover more than one atomic value.

<table>
<thead>
<tr>
<th>Atomic values:</th>
<th>Wildcard values:</th>
</tr>
</thead>
<tbody>
<tr>
<td>F feminine</td>
<td>X M, I, F, N any of the basic four genders</td>
</tr>
<tr>
<td>I masculine inanimate</td>
<td>H F, N feminine or neuter</td>
</tr>
<tr>
<td>M masculine animate</td>
<td>T I, F masculine inanimate or feminine (plural only)</td>
</tr>
<tr>
<td>N neuter</td>
<td>Y M, I masculine (either animate or inanimate)</td>
</tr>
<tr>
<td></td>
<td>Z M, I, N not feminine (i.e., masculine animate/inanimate or neuter)</td>
</tr>
<tr>
<td></td>
<td>Q feminine (with singular only) or neuter (with plural only)</td>
</tr>
</tbody>
</table>
Czech Positional Tagset: SubPOS position

- SubPOS values do not always encode the same level of detail:
  - E.g., personal pronouns:
    - P (regular personal pronoun),
    - H (clitical personal pronoun), and
    - 5 (personal pronoun in prepositional form).
  - E.g., There are eight values corresponding to relative pronouns, four to generic numerals, etc.
- It is a trade off between complexity of the tagset and linguistic adequacy
## The Russian positional tagset

<table>
<thead>
<tr>
<th>Pos</th>
<th>Abbr</th>
<th>Name</th>
<th>Nr. of values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p</td>
<td>Part of Speech</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>s</td>
<td>SubPOS (Detailed Part of Speech)</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>g</td>
<td>Gender</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>y</td>
<td>Animacy</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>n</td>
<td>Number</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>c</td>
<td>Case</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>f</td>
<td>Possessor’s Gender</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>m</td>
<td>Possessor’s Number</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>e</td>
<td>Person</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>r</td>
<td>Reflexivity</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>t</td>
<td>Tense</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>b</td>
<td>Verbal aspect</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>d</td>
<td>Degree of comparison</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>a</td>
<td>Negation</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>v</td>
<td>Voice</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>i</td>
<td>Variant, Abbreviation</td>
<td>7</td>
</tr>
</tbody>
</table>
Tagsets for highly inflected languages are typically far bigger than those for English.

It might seem obvious that the size of a tagset would be negatively correlated with tagging accuracy: for a smaller tagset, there are fewer choices to be made, thus there is less opportunity for an error.
Tagset size and tagging accuracy

- Tagsets for highly inflected languages are typically far bigger than those for English.
- It might seem obvious that the size of a tagset would be negatively correlated with tagging accuracy: for a smaller tagset, there are fewer choices to be made, thus there is less opportunity for an error.
- (Elworthy 1995) shows this is not true.
External and Internal Criteria for Tagset Design (Elworthy 1995)

- **External criterion**: the tagset must be capable of making the linguistic distinctions required in the output corpora;
- **Internal criterion**: make the tagging as effective as possible;
Harmonizing tagsets across languages?

- e.g., *Multext*-East, CLiC-TALP (Torruella 2002)
- What are the advantages and disadvantages?
Harmonization: Pros

- Harmonized tagsets make it easier to develop multilingual applications or to evaluate language technology tools across several languages.
- Interesting from a language-typological perspective as well because standardized tagsets allow for a quick and efficient comparison of language properties.
- Convenient for researchers working with corpora in multiple languages – they do not need to learn a new tagset for each language.
Various grammatical categories and their values might have different interpretations in different languages.

- E.g., definiteness is expressed differently in various languages: determiners in English, clitics in Romanian; only pronominal adjectives in Lithuanian etc.
- E.g., plural: in Russian, only plural; in Slovenian, dual and plural.
Summary: Tagset design challenges

- Tagset size: computationally tractable? Linguistically adequate?
- Atomic or Structural? If Structural, compact or positional?
- What linguistic properties are relevant?
  - The PDT Czech tagset mixes the morpho-syntactic annotation with what might be called dictionary information, e.g., gender;
  - The Czech tagset sometimes combines several morphological categories into one.
  - The Penn Treebank tagset has many singleton tags (e.g., infinitive to, punctuation).
- Should the system be standardized and be easily adaptable for other languages?
Some annotation problems

- Truly ambiguous text.
  \[Má\m\text{m rá\d maso na šalvěji.} – M or F? (TWO GENDERS)\]
  \[Je to v kuchyni. – lemma kuchyně or kuchyň? (TWO NOMINATIVES)\]
Some annotation problems

- Truly ambiguous text.
  \[Mám rád maso na šalvěji. – M or F? (TWO GENDERS)\]
  \[Je to v kuchyni. – lemma kuchyně or kuchyň? (TWO NOMINATIVES)\]
- Expressions from other languages (company names, logos, song names, DJ’s, horse names).

\[\text{Mám rád maso na šalvěji. – M or F? (TWO GENDERS)}\]
\[\text{Je to v kuchyni. – lemma kuchyně or kuchyň? (TWO NOMINATIVES)}\]
\[\text{Expressions from other languages (company names, logos, song names, DJ’s, horse names).}\]
Some annotation problems

- Truly ambiguous text.
  *Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)
  *Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
Some annotation problems

- Truly ambiguous text.
  
  *Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)
  
  *Je to v kuchyni.* – lemma *kuchyně* or *kuchyň?* (TWO NOMINATIVES)

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)

- Text with errors, text of foreigners, text which is hard to understand.
  
  *Tak že mislim, že kdy by byl sebe svim dítě, . . .* (MANY ERRORS)
Some annotation problems

- Truly ambiguous text.
  \[\text{Mám rád maso na šalvěji.} - \text{M or F? (TWO GENDERS)}\]
  \[\text{Je to v kuchyni.} - \text{lemma kuchyně or kuchyň? (TWO NOMINATIVES)}\]

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (\textit{La Manche} – fem in French, masc in Czech)

- Text with errors, text of foreigners, text which is hard to understand.
  \[\text{Tak že mislim, že kdy by byl sebe svim dítě, …} \text{(MANY ERRORS)}\]

- Diachronic corpus? Lemmas change over time.
  \[\text{kóň, kuoň, kůň. \text{‘horse’ ALL HORSE BUT AT DIFF TIME}}\]
Some annotation problems

- Truly ambiguous text.
  
  \textit{Mám rád maso na šalvěji.} – M or F? (TWO GENDERS)
  
  \textit{Je to v kuchyni.} – lemma \textit{kuchyně} or \textit{kuchyň}? (TWO NOMINATIVES)

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (\textit{La Manche} – fem in French, masc in Czech)

- Text with errors, text of foreigners, text which is hard to understand.
  
  \textit{Tak že mislim, že kdy by byl sebe svým dítě, …} (MANY ERRORS)

- Diachronic corpus? Lemmas change over time.
  
  \textit{kón}, \textit{kuoň}, \textit{kůň}. ‘horse’ ALL HORSE BUT AT DIFF TIME

- Clitics
  
  \textit{Jste-li pojištěni} … ‘If you are insured …’
Some annotation problems

- Truly ambiguous text.
  
  \[Mám rád maso na šalvěji. – M or F? (TWO GENDERS)\]
  \[Je to v kuchyni. – lemma kuchyně or kuchyň? (TWO NOMINATIVES)\]

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? \(\text{(La Manche – fem in French, masc in Czech)}\)

- Text with errors, text of foreigners, text which is hard to understand.
  \[\text{Tak že mislim, že kdy by byl sebe svim dítě, … (MANY ERRORS)}\]

- Diachronic corpus? Lemmas change over time.
  \[kóň, kuoň, kůň. ‘horse’ ALL HORSE BUT AT DIFF TIME\]

- Clitics
  
  \[\text{Jste-li pojištěni … ‘If you are insured …’ }\]
  \[\text{Tys to viděl? ‘You haven’t seen it?’}\]
Some annotation problems

- Truly ambiguous text.
  
  \textit{Mám rád maso na šalvěji.} – M or F? (TWO GENDERS)
  
  \textit{Je to v kuchyni.} – lemma \textit{kuchyně} or \textit{kuchyň}? (TWO NOMINATIVES)

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (\textit{La Manche} – fem in French, masc in Czech)

- Text with errors, text of foreigners, text which is hard to understand.
  
  \textit{Tak že mislim, že kdy by byl sebe svim dítě, . . .} (MANY ERRORS)

- Diachronic corpus? Lemmas change over time.
  
  \textit{kón}, \textit{koun}, \textit{kûn}. ‘horse’ ALL HORSE BUT AT DIFF TIME

- Clitics
  
  \textit{Jste-li pojištěni} ... ‘If you are insured ...’
  
  \textit{Tys to viděl?} ‘You haven’t seen it?’
  
  \textit{Náms to nedal.} ‘You did not give it to us’
Some annotation problems

- Truly ambiguous text.
  
  *Mám rád maso na šalvěji. – M or F? (TWO GENDERS)*
  
  *Je to v kuchyni. – lemma kuchyně or kuchyň? (TWO NOMINATIVES)*

- Expressions from other languages (company names, logos, song names, DJ’s, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)

- Text with errors, text of foreigners, text which is hard to understand.
  
  *Tak že mislim, že kdy by byl sebe svím dítě, … (MANY ERRORS)*

- Diachronic corpus? Lemmas change over time.
  
  *kůň, kuoň, kůň. ‘horse’ ALL HORSE BUT AT DIFF TIME*

- Clitics
  
  *Jste-li pojištěni ... ‘If you are insured ...’*
  
  *Tys to viděl? ‘You haven’t seen it?’*
  
  *Náms to nedal. ‘You did not give it to us’*
Cohen’s kappa (Cohen 1960)

- The most popular measure of agreement between two annotators.
- Takes into account (somewhat) the possibility of chance agreement.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

- $Pr(a)$ - the relative observed agreement
- $Pr(e)$ - the hypothetical probability of chance agreement
  $$Pr(e) = \sum_t \frac{t_a \times t_b}{N}$$
  - $t_a$ – number of tags $t$ assigned by annotator $a$
  - $N$ – number of all tags
- Weighted kappa – gives different weights to different errors.
(Variant of) Kendall’s tau - the minimal number of operations necessary to turn one annotation into the other.

There are other measures.

High agreement is important but it is not everything:

- One can use a tagset with a single tag.
- The annotation manual can be purely formal (Tag all sentence initial words as topics).
- On the other hand, if iaa is below the accuracy of a tagger ...