ESSLLI 2010: Resource-light Morpho-syntactic Analysis of Highly Inflected Languages
Classical Approaches to Morphology

Anna Feldman & Jirka Hana
Overview

1. Intro

2. Tags

3. MA
   - Complications
   - Approaches
   - Linguistic App.
   - Engineering App.
Processing morphology

1. **Lemmatization**: word → lemma
   
   \[ saw \rightarrow \{ \text{see}, \text{saw} \} \]

2. **Morphological analysis (MA)**: word → setOf(lemma + tag), ignores context
   
   \[ saw \rightarrow \{ \langle \text{see, verb.past} \rangle, \langle \text{saw, noun.sg} \rangle, \} \]

3. **Tagging**: word → tag (often also lemma), considers context
   
   \[ saw @ Peter saw her. \rightarrow \{ \langle \text{see, verb.past} \rangle \} \]

4. **Morpheme segmentation**: \textit{de-nation-al-iz-ation}

5. **Generation**: see + verb.past → saw
Applications

- Parsing/chunking (used in machine translation, grammar correction, etc.)
- Text Generation
- Search and information retrieval. One usually searches for a lexeme not for a particular form.
- Text-to-speech synthesis.
  \( \text{read}_{\text{present}} \) [rid] vs. \( \text{read}_{\text{past}} \) [r̂d]
  Russian: \( \text{snèga}_{\text{noun.masc.sg.gen}} \) ‘snow’ vs. \( \text{snegà}_{\text{noun.masc.pl.nom/acc}} \)
- Spell checking
- (Computer assisted) language learning.
Creation/Acquisition

1. manually provided rules
2. use machine learning
   1. supervised – deduced from an annotated corpus
   2. unsupervised – deduced from plain text
3. hybrid
Tags and tagsets

- **(morphological) tag** – a symbol encoding morphological properties of a word
- **tagset** – set of tags, depends on language and application
Tags and tagsets

- **(morphological) tag** – a symbol encoding morphological properties of a word
- **tagset** – set of tags, depends on language and application

- Penn Tagset: about 40 tags; **VBD** – verb in past tense
- Czech Positional Tagset: about 4000 tags; **VpNS---XR-AA---**
## Tagsets for English: Penn Treebank

<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>VBN</td>
<td>Verb, past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
<td>VBP</td>
<td>Verb, non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td>wildest</td>
<td>VBZ</td>
<td>Verb, 3sg pres</td>
<td>eats</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>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Carolinas</td>
<td>#</td>
<td>Pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both</td>
<td>&quot;</td>
<td>Left quote</td>
<td>(’ or “)</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>’s</td>
<td>”</td>
<td>Right quote</td>
<td>(’ or ”)</td>
</tr>
<tr>
<td>PP</td>
<td>Personal pronoun</td>
<td>I, you, he</td>
<td>(</td>
<td>Left parenthesis</td>
<td>[ , ( , { , &lt;</td>
</tr>
<tr>
<td>PP$</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>
<tr>
<td>RP</td>
<td>Particle</td>
<td>up, off</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Czech positional tagset

<table>
<thead>
<tr>
<th>Position</th>
<th>Name</th>
<th>Description</th>
<th>Example</th>
<th>Example</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>
Morphological analysis

MA: form $\rightarrow$ set(lemma $\times$ set(tag))
Morphological analysis

MA: form $\rightarrow$ set(lemma $\times$ set(tag))

English: $\text{her} \rightarrow \{ (\text{she}, \{\text{PP}\}) ,
(\text{her}, \{\text{PP$}\}) \}$

Czech: $\text{ženou} \rightarrow \{ (\text{žena} '\text{woman}', \{\text{noun fem sing inst}\}) ,
(\text{hnát} '\text{hurry}', \{\text{verb pres pl 3rd}\}) \}$

$\text{ženy} \rightarrow \{ (\text{žena} '\text{woman}', \{\text{noun fem sing gen},
\text{noun fem pl nom},
\text{noun fem pl acc},
\text{noun fem pl voc}\}) \}$
Complications

- Stem internal (non-concatenative) alternations:
  German: $\textit{Stuhl} \rightarrow \textit{St"{u}hl-e}$, $\textit{Vater} \rightarrow \textit{V"{a}ter}$

- Irregularities.
  English: $\textit{goose} \rightarrow \textit{geese}$, $\textit{sheep} \rightarrow \textit{sheep}$
  Russian plural: $\textit{knig-a} \rightarrow \textit{knig-i}$, $\textit{stol} \rightarrow \textit{stol-y}$, but $\textit{kofe} \rightarrow \textit{kofe}$

- Phonological/graphemic alternations:
  English: $\textit{knife} \rightarrow \textit{knive-s}$, $\textit{city} \rightarrow \textit{citi-es}$

- Homonymy:
  English -$s$ – 3rd person singular of verbs vs. plural of nouns;
  Czech -$a$ / -$e$ (see Table next slide).
Table: Homonymy of the a ending in Czech

<table>
<thead>
<tr>
<th>form</th>
<th>lemma</th>
<th>gloss</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>měst-a</td>
<td>město</td>
<td>town</td>
<td>noun neut sg gen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NP1 (5) noun neut pl nom (voc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NP4 noun neut pl acc</td>
</tr>
<tr>
<td>tém-a</td>
<td>téma</td>
<td>theme</td>
<td>noun neut sg nom (voc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NS4 noun neut sg acc</td>
</tr>
<tr>
<td>žen-a</td>
<td>žena</td>
<td>woman</td>
<td>noun fem sg nom</td>
</tr>
<tr>
<td>pán-a</td>
<td>pán</td>
<td>man</td>
<td>noun masc anim sg gen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS4 noun masc anim sg acc</td>
</tr>
<tr>
<td>ostrov-a</td>
<td>ostrov</td>
<td>island</td>
<td>noun masc inanim sg gen</td>
</tr>
<tr>
<td>předsed-a</td>
<td>předseda</td>
<td>president</td>
<td>noun masc anim sg nom</td>
</tr>
<tr>
<td>vidě-l-a</td>
<td>vidět</td>
<td>see</td>
<td>verb past fem sg</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>verb past neut pl</td>
</tr>
<tr>
<td>vidě-n-a</td>
<td></td>
<td></td>
<td>verb passive fem sg</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>verb passive neut pl</td>
</tr>
<tr>
<td>vid-a</td>
<td>dv-a</td>
<td>two</td>
<td>verb transgressive masc sg</td>
</tr>
<tr>
<td>dv-a</td>
<td>dv-a</td>
<td>two</td>
<td>numeral masc sg nom</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>numeral masc sg acc</td>
</tr>
</tbody>
</table>
The table below shows the ending -e and noun cases in Czech.

<table>
<thead>
<tr>
<th>case</th>
<th>form</th>
<th>lemma</th>
<th>gender</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>nom</td>
<td>kuř-e</td>
<td>kuře</td>
<td>neuter</td>
<td>chicken</td>
</tr>
<tr>
<td>gen</td>
<td>muž-e</td>
<td>muž</td>
<td>masc.anim.</td>
<td>man</td>
</tr>
<tr>
<td>dat</td>
<td>mouš-e</td>
<td>moucha</td>
<td>feminine</td>
<td>fly</td>
</tr>
<tr>
<td>acc</td>
<td>muž-e</td>
<td>muž</td>
<td>masc.anim.</td>
<td>man</td>
</tr>
<tr>
<td>voc</td>
<td>pan-e</td>
<td>pán</td>
<td>masc.anim.</td>
<td>mister</td>
</tr>
<tr>
<td>loc</td>
<td>mouš-e</td>
<td>moucha</td>
<td>feminine</td>
<td>fly</td>
</tr>
<tr>
<td>inst</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Different Approaches

Two different ways to address phonological/graphemic variations and complex paradigm systems when designing a morphological analyzer:

1. A linguistic approach.
   A phonological component accompanying the simple concatenative process of attaching an ending

2. An engineering approach.
   - No (or very rudimentary) phonological component
   - Phonological changes and irregularities are factored into endings and a higher number of paradigms
## Approaches: Comparison

<table>
<thead>
<tr>
<th>woman</th>
<th>owl</th>
<th>draft</th>
<th>iceberg</th>
<th>vapor</th>
<th>fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>žen-a</td>
<td>sov-a</td>
<td>skic-a</td>
<td>kr-a</td>
<td>pár-a</td>
<td>mouch-a</td>
</tr>
<tr>
<td>žen-y</td>
<td>sov-y</td>
<td>skic-i</td>
<td>kr-y</td>
<td>pár-y</td>
<td>mouch-y</td>
</tr>
<tr>
<td>žen-ě</td>
<td>sov-ě</td>
<td>skic-e</td>
<td>kř-e</td>
<td>pář-e</td>
<td>mouš-e</td>
</tr>
</tbody>
</table>

### Linguistic approach

\[
\text{žen} + \begin{cases} 
  a \\
  y \\
  ě \\
  0
\end{cases} \quad \text{sov} + \begin{cases} 
  a \\
  y \\
  ě \\
  0
\end{cases} \quad \text{skic} + \begin{cases} 
  a \\
  y \\
  e \\
  0
\end{cases} \quad \text{kr} + \begin{cases} 
  a \\
  y \\
  e \\
  0
\end{cases} \quad \text{pár} + \begin{cases} 
  a \\
  y \\
  e \\
  0
\end{cases} \quad \text{mouch} + \begin{cases} 
  a \\
  y \\
  e \\
  0
\end{cases}
\]

### Engineering approach

\[
\text{žen} + \begin{cases} 
  a \\
  y \\
  ě \\
  0
\end{cases} \quad \text{sov} + \begin{cases} 
  a \\
  y \\
  ě \\
  0
\end{cases} \quad \text{skic} + \begin{cases} 
  a \\
  i \\
  e \\
  0
\end{cases} \quad \text{k} + \begin{cases} 
  r\text{a} \\
  r\text{y} \\
  ř\text{e} \\
  e\text{r}
\end{cases} \quad \text{p} + \begin{cases} 
  á\text{r}á \\
  á\text{r}y \\
  áře \\
  ar
\end{cases} \quad \text{m} + \begin{cases} 
  ou\text{cha} \\
  ou\text{chy} \\
  ou\text{še} \\
  u\text{ch}
\end{cases}
\]
Phonological component accompanying the simple concatenative process of attaching an ending;
Linguistic Approach

- Phonological component accompanying the simple concatenative process of attaching an ending;

- Advantages:
  - Small set of paradigms and morphemes
  - Captures linguistics generalizations
Linguistic Approach

- Phonological component accompanying the simple concatenative process of attaching an ending;

- Advantages:
  - Small set of paradigms and morphemes
  - Captures linguistics generalizations

- Problems:
  - Requires a lot of linguistic work and expertise
  - For many languages, the linguistic knowledge is not precise enough
  - It is usually not straightforward to translate even a precisely formulated linguistic description of a morphology into the representation recognized by such a system
Morphology analyzed by finite-state automata/transducers.

It is by far the most popular approach in the field.


Two-level morphology (Koskenniemi 1983, 1984)
What is finite state automaton (FSA)?

- Introduced by (Kleene 1956).
- A kind of directed graph:
  - Nodes are called states
  - Each edge is labeled with an accepted string (possibly empty)
  - One node is called the start state
  - One or more nodes are called stopping (or accepting) states

- Recognize/generate regular languages, i.e., languages specified by regular expressions.
An example

- Regular expression: `colou?r`
- Finite state machine:
Some properties of finite state machines

- Recognition problem can be solved in linear time (independent of the size of the automaton).
- There is an algorithm to transform each automaton into a unique equivalent automaton with the least number of states.
A finite state automaton is deterministic iff it has
- no $\epsilon$ (empty) transitions and
- for each state and each symbol there is at most one applicable transition.

Every non-deterministic automaton can be transformed into a deterministic one:
- Define new states representing a disjunction of old states for each non-determinacy which arises.
- Define arcs for these states corresponding to each transition which is defined in the non-deterministic automaton for one of the disjuncts in the new state names.
Finite State Transducers

- Translate strings from one language to strings from another language
- Like a FSA, but each edge is associated with two strings.
Two-level morphology

- Uses 2 levels
  - lexical/underlying/deep forms
  - surface forms
  - one-one correspondence between symbols

  country y 0 + s
  countries 0 s
Two-level morphology

- Uses 2 levels
  - lexical/underlying/deep forms
  - surface forms
  - one-one correspondence between symbols

  country 0 + s
  countries 0 s

- Two components
  - Linked lexicons – sets of (underlying forms of) morphemes
  - Phonological rules – relate lexical and surface forms
Two-level morphology (cont.)

- Linked lexicons

prefixes → roots → suffixes
Two-level morphology (cont.)

- Linked lexicons

  \[ \begin{array}{cccc}
  \Rightarrow & \Rightarrow & \Rightarrow \\
  \text{prefixes} & \text{roots} & \text{suffixes} \\
  \end{array} \]

- Rules
  - relate underlying and surface forms
  - applied simultaneously
  - Form: lexical symbol : surface symbol operator context
    - operators: ⇔, ⇒, ⇐, ⇓
    - context - left / right context
  - \( y:i \leftrightarrow \_ 0:e \) (y – ie)
Two-level morphology (cont.)

- Linked lexicons

- Rules
  - relate underlying and surface forms
  - applied simultaneously
  - Form: lexical symbol : surface symbol operator context
    - operators: ⇔, ⇒, ⇐, ⇐
    - context - left / right context
  - \( y:i ⇔ 0:e \ (y - ie) \)

- All this can be compiled into one big FST.
Engineering approach

- No (or very rudimentary) phonological component
- Phonological changes and irregularities are factored into endings and a higher number of paradigms. Therefore the terms *stem* and *ending* have slightly different meanings than they traditionally do. A stem is the part of the word that does not change within its paradigm, and the ending is the part of the word that follows such a stem.
Advantages:
- high speed;
- simple implementation;
- straightforward morphology specification;

Problems:
- high number of paradigms (e.g. around 500 for Czech);
- Impossibility to capture even the simplest and most regular phonological changes and so predict the behavior of new lexemes;
- in theory, incapable of capturing some languages

(Hajič 2004) for Czech; (Mikheev and Liubushkina 1995) for Russian