Morphology

- Introduction
- Morphology
- Morphological Analysis (MA)
- Using FS techniques in MA
- Automatic learning of the morphology of a language
Morphology

- Morphology
  - Structure of a word as a composition of morphemes
  - Related to word formation rules
- Functions
  - Flexion
  - Derivation
  - Composition
- Result of morphologic analysis
  - Morphosyntactic categorization (POS)
    - e.g. Parole tagset (VMIP1S0), more than 150 categories for Spanish
    - e.g. Penn Treebank tagset (VBD), about 30 categories for English
  - Morphological features
    - Number, case, gender, lexical functions
Morphology 3

• Morphologic analysis
  • Decompose a word into a concatenation of morphemes
  • Usually some of the morphemes contain the meaning
    • One (root or stem) in flexion and derivation
    • More than one in composition
  • The other (affixes) provide morphological features

• Problems
  • Phonological alterations in morpheme concatenation
  • Morphotactics
    • Which morphemes can be concatenated with which others
• Problems
  • Affixes
    • Suffixes, prefixes, infixes, interfixes
  • flexive Affixes ≠ derivative Affixes
  • Derivation implies sometimes a semantic change not always predictable
    • Meaning extensions
    • Lexical rules
  • A derivative suffix can be followed by a flexive suffix
    • love => lover => lovers
  • Flexion does not change POS, sometimes derivation does
  • Flexion affects other words in the sentence
    • agreement
• Morphotactics
  • Word formation rules
  • Valid combinations between morphemes
    • Simple concatenation
    • Complex models root/pattern
    • Regularity language dependent

• Phonological alterations (Morphophonology)
  • Changes when concatenating morphemes
  • Source: Phonology, morphology, orthography
  • variable in number and complexity
  • e.g. vocalic harmony
Morphology 6

- 1 morpheme:
  - evitar
- 2 morphemes:
  - evitable = evitar + able
- 3 morphemes:
  - inevitable = in + evitar + able
- 4 morphemes:
  - inevitabilidad = in + evitar + able + idad
Morphology 7

Flexive Morphology

- **number**
  - house  houses
  - cheval  chevaux
  - casa  casas

- **verbal form**
  - walk  walkes  walked  walking
  - amo  amas  aman  ...

- **gender**
  - niño  niña
Morphology

- **Form**
  - Without change: barcelonés
  - Prefix: inevitable
  - Suffix: importantísimo

- **Source**
  - verb => adjective: tardar => tardío
  - verb => noun: sufrir => sufrimiento
  - noun => noun: actor => actorazo
  - noun => adjective: atleta => atlético
  - adjective => adjective: rojo => rojizo
  - adjective => adverb: alegre => alegremente
Morphological Analysis

Types of morphological analyzers

Formaries
- Dictionaries of word forms
  + efficiency
  + Languages with few variants (e.g. English)
  + extensibility
  + Possibility of building and maintenance from a morphological generator
    - Languages with high flexive variation
    - derivation, composition

- FS techniques
  - FSA
    - 1 level analyzers
  - FST
    - > 1 level analyzers
Morphological Analysis 2

2 levels morphological analyzers

• General model for languages with morpheme concatenation
• Independence between lingware and analyzer
• Valid for analysis and generation
• Distinction between lexical and superficial levels
• Parallel rules for morphophonology
• Simple implementation
Morphological Analysis

• Morphological rules
  • Define the relations between characters (surface) and morphemes and map strings of characters and the morphemic structure of the word.

• Spelling rules
  • Perform at the level of the letters forming the word. Can be used to define the valid phonological alterations.

• Ritchie, Pulman, Black, Russell, 1987
Morphological Analysis

- input:
  - form
- output
  - lemma + morphological features

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>cat + N + sg</td>
</tr>
<tr>
<td>cats</td>
<td>cat + N + pl</td>
</tr>
<tr>
<td>cities</td>
<td>city + N + pl</td>
</tr>
<tr>
<td>merging</td>
<td>merge + V + pres_part</td>
</tr>
<tr>
<td>caught</td>
<td>(catch + V + past) or (catch + V + past_part)</td>
</tr>
</tbody>
</table>
### Morphological Analysis

#### Morphotactics

<table>
<thead>
<tr>
<th>reg_noun</th>
<th>irreg_pl_noun</th>
<th>irreg_sg_noun</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>fox</td>
<td>sheep</td>
<td>sheep</td>
<td>-s</td>
</tr>
<tr>
<td>cat</td>
<td>mice</td>
<td>mouse</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram:

```
0 -> 1 -> 2
reg_noun
irreg_pl_noun
irreg_sg_noun
plural (-s)
```
fog
cat
dog
donkey
mouse
mice

Letter Transducers
### Morphological Analysis

#### Upper Level
- **Lexic**: cat + N
- **Cat + N + Pl**: cat + N + pl

#### Lower Level
- **Surface**: cat
- **Cats**

Diagram:

```
c:c  a:a  t:t  +N:ε  +pl:s
```
Morphological Analysis

• As a recognizer
  • From a pair of input strings (one lexical and the other superficial) and answers if one is transduction of the other.

• As a generator
  • generated pairs of strings

• As a translator
  • From a superficial string generates its lexical transduction

Using FST
Morphological Analysis

<table>
<thead>
<tr>
<th>reg_noun</th>
<th>irreg_pl_noun</th>
<th>irreg_sg_noun</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>fox</td>
<td>sheep</td>
<td>sheep</td>
<td>s</td>
</tr>
<tr>
<td>cat</td>
<td>mouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>goose</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Diagram:**
- **0** (reg_noun) → **1** (irreg_sg_noun) → **2** (irreg_pl_noun) → **3** (irreg_pl_noun) → **4** (reg_noun) → **5** (irreg_sg_noun) → **6** (irreg_sg_noun) → **2** (plugin)
- **1** (irreg_sg_noun) → **2** (irreg_pl_noun) → **3** (irreg_pl_noun) → **4** (reg_noun) → **5** (irreg_sg_noun) → **6** (irreg_sg_noun) → **2** (plugin)

**Annotations:**
- `+N:ε`
- `+sg:ε`
- `+pl:ε`
<table>
<thead>
<tr>
<th>morphotactics</th>
<th>spelling rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>lexical level</strong></td>
<td>f o x +N +pl</td>
</tr>
<tr>
<td><strong>intermediate level</strong></td>
<td>f o x ^ s</td>
</tr>
<tr>
<td><strong>superficial level</strong></td>
<td>f o x e s</td>
</tr>
</tbody>
</table>
Morphological Analysis

fog  cat  dog  donkey  mouse  mice

+sg:ε  +u:ε  +N:ε  +pl:ε

+sg:ε  +pl:~s

+sg:ε  +pl:ε

NLP Morphology
## Morphological Analysis

### Spelling rules

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>consonant doubling</td>
<td>single letter consonant doubled before -ing/-ed</td>
<td>beg/begging</td>
</tr>
<tr>
<td>e deletion</td>
<td>silent e dropped before -ing/-ed</td>
<td>make/making</td>
</tr>
<tr>
<td>e insertion</td>
<td>e added after -s,-z,-x,-ch,-sh before -s</td>
<td>watch/watches</td>
</tr>
<tr>
<td>y replacement</td>
<td>-y changes to -ie before -s, to i before -ed</td>
<td>try/tries</td>
</tr>
<tr>
<td>k insertion</td>
<td>verbs ending with voyel +c add -k</td>
<td>panic/panicked</td>
</tr>
</tbody>
</table>
Morphological Analysis

Spelling rules: e-insertion

\[
\varepsilon:e \iff [xsz]^:\varepsilon ___ s# \\
\Rightarrow \quad \text{decomposition} \quad \Rightarrow \quad / \iff \\
\varepsilon:e \Rightarrow [xsz]^:\varepsilon ___ s# \\
\varepsilon:e / \iff [xsz]^:\varepsilon ___ s#
\]
epenthesis

\[ + : e \quad \leftrightarrow \quad \{ < \{ s:s \ c:c \} \ h:h > \ s:s \ x:x \ z:z \} \quad \rightarrow \quad s:s \]

context

C: \{...\}
V: \{a,e,i,o,u,y\}
C2: \{...\}
=: whatever

Example: \text{box} + s \quad \text{box} e s
### e-deletion

<table>
<thead>
<tr>
<th></th>
<th>e</th>
<th>+</th>
<th>ed</th>
</tr>
</thead>
<tbody>
<tr>
<td>mov</td>
<td>e</td>
<td>+</td>
<td>ed</td>
</tr>
<tr>
<td>mov</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>agre</td>
<td>e</td>
<td>+</td>
<td>ed</td>
</tr>
<tr>
<td>agre</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
e : 0 \iff = :C2 \quad \text{---} \quad <+:0 \ V:= >
\]
\[
\text{or}<C:C \ V:V> \quad \text{---} \quad <+:0 \ e:e >
\]
\[
\text{or}<c:c \ g:g> \quad \text{---} \quad <+:0 \ {e:e \ i:i} >
\]
\[
\text{or}1:0 \quad \text{---} \quad +:0
\]
\[
\text{or}c:c \quad \text{---} \quad <+:0 \ a:0 \ t:t \ b:b>
\]
### a-deletion

\[
\text{a : 0 } \iff \langle \text{c:c e:0 +:0} \rangle \quad \cdots \quad t:t
\]

<table>
<thead>
<tr>
<th>redu</th>
<th>c</th>
<th>e</th>
<th>+</th>
<th>a</th>
<th>t</th>
<th>ion</th>
</tr>
</thead>
<tbody>
<tr>
<td>redu</td>
<td>c</td>
<td></td>
<td></td>
<td>a</td>
<td>t</td>
<td>ion</td>
</tr>
<tr>
<td>redu</td>
<td>c</td>
<td></td>
<td></td>
<td>t</td>
<td>t</td>
<td>ion</td>
</tr>
</tbody>
</table>

... contexto izdo foco contexto ...

dcho
Morphological Analysis

Lexicon-FST

spelling rules

superficial level

lexical level

intermediate level
Lexicon-FST

\[ FST_1 \ldots \ FST_n \]

\[ FST_A = FST_1 \land \ldots \land FST_n \]

Lexicon-FST

\[ FSTA \]

Lexicon-FST

composition

intersection
Automatic morphology learning

• Problem
  • Paradigm stem + affixea
  • Obtaining the stems
  • Classification of stems into models
  • Learning part of the morphology (e.g. derivational)

• Two approaches
  • No previous morphologic knowledge is available
    • Goldsmith, 2001
    • Brent, 1999
    • Snover, Brent, 2001, 2002
  • Morphologic knowledge can be used
    • Oliver at al, 2002
Automatic morphology learning

- Automatic morphological analysis
  - Identification of borders between morphemes
    - Zellig Harris
      - \{prefix, suffix\} conditional entropy
  - bigrams and trigrams with high probability of forming a morpheme
  - Learning of patterns or rules of mapping between pairs of words
  - Global approach (top-down)
    - Goldsmith, Brent, de Marcken
Automatic morphology learning

- Goldsmith’s system based on MDL (Minimum Description Length)
  - Initial Partition: word -> stem + suffix
    - split-all-words
      - A good candidate to {stem, suffix} splitting in a word has to be a good candidate in many other words
  - MI (mutual information) strategy
    - Faster convergence
- Learning Signatures
  - {signatures, stem, suffixes}
- MDL
Automatic morphology learning

- Semi-automatic morphological analysis
  - Oliver, 2004
  - Starts with a set of manually written morphological rules
    - TL:TF:Desc
      - lemma ending
      - form ending
      - POS
  - Lists of nonflexive classes, closed classes and irregular words
- Corpora
  - Serbo-Croatian 9 Mw
  - Russian 16 Mw