Universal Feature-based Morphological Trees

Federica Gamba, Abishek Stephen, Zdeněk Žabokrtský
Outline

An Overview

Exploited Resources

Workflow

Results
An Overview

Exploited Resources

Workflow

Results
An Overview

Exploited Resources

Workflow

Results

Comparability of Trees I

I

nsubj

PRON

will

aux

AUX

go

root

VERB

through

case

ADP

forest

obl

NOUN

a

det

DET

Půjdu

root

VERB

lesem

obl

NOUN
Comparability of Trees II

An Overview

Exploited Resources

Workflow

Results
Features and Morphs

An Overview  Exploited Resources  Workflow  Results
Outline - Exploited Resources

An Overview

Exploited Resources

Workflow

Results
• **UniSegments** (Žabokrtský et al., 2022): collection of harmonized versions of 17 segmentation resources covering 32 languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>DeriNet</td>
</tr>
<tr>
<td>English</td>
<td>MorphoLex</td>
</tr>
<tr>
<td>French</td>
<td>Demonette</td>
</tr>
<tr>
<td>Italian</td>
<td>DerIvaTario</td>
</tr>
<tr>
<td>Latin</td>
<td>WordFormationLatin</td>
</tr>
<tr>
<td>Catalan</td>
<td>MorphyNet</td>
</tr>
<tr>
<td>Finnish</td>
<td>MorphyNet</td>
</tr>
<tr>
<td>German</td>
<td>MorphyNet</td>
</tr>
<tr>
<td>Hungarian</td>
<td>MorphyNet</td>
</tr>
<tr>
<td>Portuguese</td>
<td>MorphyNet</td>
</tr>
</tbody>
</table>
UniMorph and SIGMORPHON data

- **UniMorph** (McCarthy et al., 2020): collection of morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema.
UniMorph and SIGMORPHON data

- **UniMorph** (McCarthy et al., 2020): collection of morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema.

- **SIGMORPHON**: manually annotated Czech dataset made available for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022).
Universal Dependencies

- **UD** (de Marneffe et al., 2021): selected treebanks from version 2.12.

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>PUD, PDT</td>
</tr>
<tr>
<td>English</td>
<td>PUD, GUM</td>
</tr>
<tr>
<td>Finnish</td>
<td>PUD, TDT</td>
</tr>
<tr>
<td>French</td>
<td>PUD, GSD</td>
</tr>
<tr>
<td>German</td>
<td>PUD, GSD</td>
</tr>
<tr>
<td>Italian</td>
<td>PUD, ISDT</td>
</tr>
<tr>
<td>Portuguese</td>
<td>PUD, Bosque</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan</td>
<td>AnCora</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Szeged</td>
</tr>
<tr>
<td>Latin</td>
<td>ITTB</td>
</tr>
</tbody>
</table>
Outline - Workflow

An Overview

Exploited Resources

Workflow

Results
0. In **SIGMORPHON** data?
   
   0.1 N: continue to 1
   0.2 Y: segment and quit
Manipulation of Nodes I

0. In **SIGMORPHON** data?
   0.1 N: continue to 1
   0.2 Y: segment and quit

1. **Lemma** segmented in UniSegments?
   1.1 N: cs. *rok* \(\rightarrow\) *rok*; continue to 3
   1.2 Y: cs. *prokonzul* ‘proconsul’ \(\rightarrow\) *pro* + *konzul*; continue to 2
2. **Inflected** form of a segmented lemma?
   
   2.1 N: cs. *prokonzul, rok*
   
   2.2 Y: S2.1 Form in **UniMorph**?
2. **Inflected** form of a segmented lemma?

   2.1 N: cs. *prokonzul, rok*
   2.2 Y: S2.1 Form in *UniMorph*?

      2.2.1 N: approximation of inflectional ending by string comparison
      en. *shortened* → *short* + *en* (US) + *ed* (string comparison)
      2.2.2 Y: ca. *culturals*: *cultur* + *al* (US) + *s* (UM)
2. **Inflected** form of a segmented lemma?
   2.1 N: cs. *prokonzul, rok*
   2.2 Y: S2.1 Form in UniMorph?
      2.2.1 N: approximation of inflectional ending by string comparison
             en. *shortened → short + en (US) + ed* (string comparison)
      2.2.2 Y: ca. *culturals: cultur + al (US) + s (UM)*

3. Unsegmented lemma, form inflected in UM?
   3.1 N: la. *caelum → caelum*; no splitting
   3.2 Y: fr. *travaillait → travailler + ait*
An Overview

Exploited Resources

Workflow

Results
Feature Extraction I

- *gyerekek* ‘children’: Number=Plur|Case=Nom

- *gyerek + ek → gyerek*: Number=Plur|Case=Nom
  - *ek*: Number=Plur|Case=Nom

- *gyerek*: Number=Plur
  - *gyerek*: Case=Nom

- *ek*: Number=Plur
  - *ek*: Case=Nom

- **ΔP scores** (Jenkins and Ward, 1965) as a measure of cue validity, i.e. measuring how strongly two events are linked.

\[
\Delta P_{forward} = P(m|f) - P(m|\neg f) \tag{1}
\]

\[
\Delta P_{backward} = P(f|m) - P(f|\neg m) \tag{2}
\]
## Feature Extraction II

<table>
<thead>
<tr>
<th>Morph</th>
<th>Feature</th>
<th>ΔP forward</th>
<th>ΔP backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>ek</td>
<td>Case=Nom</td>
<td>-0.006</td>
<td>-0.146</td>
</tr>
<tr>
<td>ek</td>
<td>Number=Sing</td>
<td>-0.033</td>
<td>-0.431</td>
</tr>
<tr>
<td>ek</td>
<td>Person=3</td>
<td>0.031</td>
<td>0.427</td>
</tr>
<tr>
<td>ek</td>
<td>Definite=Ind</td>
<td>0.026</td>
<td>0.328</td>
</tr>
<tr>
<td>ek</td>
<td>PronType=Ind</td>
<td>0.064</td>
<td>0.099</td>
</tr>
<tr>
<td>ek</td>
<td>Mood=Ind</td>
<td>0.030</td>
<td>0.340</td>
</tr>
<tr>
<td>ek</td>
<td>Tense=Pres</td>
<td>0.032</td>
<td>0.344</td>
</tr>
<tr>
<td>ek</td>
<td>VerbForm=Fin</td>
<td>0.028</td>
<td>0.333</td>
</tr>
<tr>
<td>ek</td>
<td>Voice=Act</td>
<td>0.028</td>
<td>0.333</td>
</tr>
<tr>
<td>ek</td>
<td>Number=Plur</td>
<td><strong>0.163</strong></td>
<td><strong>0.531</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morph</th>
<th>Number=Plur</th>
<th>ΔP forward</th>
<th>ΔP backward</th>
</tr>
</thead>
<tbody>
<tr>
<td>tunk</td>
<td>1</td>
<td><strong>0.033</strong></td>
<td><strong>0.972</strong></td>
</tr>
<tr>
<td>ok</td>
<td>7</td>
<td><strong>0.232</strong></td>
<td><strong>0.852</strong></td>
</tr>
<tr>
<td>ak</td>
<td>5</td>
<td>0.165</td>
<td>0.690</td>
</tr>
<tr>
<td>ek</td>
<td>5</td>
<td>0.163</td>
<td>0.531</td>
</tr>
<tr>
<td>ai</td>
<td>1</td>
<td>0.033</td>
<td>0.972</td>
</tr>
</tbody>
</table>
Conforming to UD

• Lemma:
  • info about morpheme in US (if available).
    la. *averto* ‘to turn away’ → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  • else, lemma = form.
Conforming to UD

- **Lemma:**
  - info about morpheme in US (if available).
  - la. *averto* 'to turn away' → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.

- **POS:**
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): X.

- **Features:**
  - **Prefixes:**
    - nmod:morph if NOUN/PROPN, else advmod:morph.
  - **Suffixes:**
    - aux:morph for VERBs and AUXs.
    - case:morph for NOUNs, PROPNs, ADJs, DETs, PRONs, ADVs, NUMs, very rare ADPs.
    - else dep:morph.
Conforming to UD

- **Lemma:**
  - info about morpheme in US (if available).
    1a. *averto* ‘to turn away’ → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.

- **POS:**
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): X.

- **Features:** feature-based alignment.
Conforming to UD

• **Lemma:**
  - info about morpheme in US (if available).
    la. *averto* ‘to turn away’ → *a + verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.

• **POS:**
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): X.

• **Features:** feature-based alignment.

• **Deprel:**
  - Prefixes: `nmod:morph` if NOUN/PROPN, else `advmod:morph`.
  - If single root: deprel of the manipulated node; `conj:morph` for the second (or +).
  - Suffixes:
    - `aux:morph` for VERBs and AUXs.
    - `case:morph` for NOUNs, PROPNs, ADJs, DETs, PRONs, ADVs, NUMs, very rare ADPs.
    - else `dep:morph`.

An Overview  Exploited Resources  Workflow  Results
Outline - Results

An Overview

Exploited Resources

Workflow

Results
Morphological Trees II

An Overview  Exploited Resources  Workflow  Results
Morphological Trees III

[Diagram of a morphological tree with nodes labeled with Finnish words and their parts of speech: <root>, löydetävissä, root, VERB, Pelien, obl, NOUN, on, aux:pass, AUX, yhtäläisyksia, obj, NOUN, punct, PUNCT, elämiemme, conj, NOUN, vällä, case, ADP, ja, cc, CCONJ, jokapäivä, amod, ADJ, sten, case:morph, X]
### CoNLL-U Representation

<table>
<thead>
<tr>
<th></th>
<th>There</th>
<th>there</th>
<th>PRON</th>
<th>EX</th>
<th>2</th>
<th>expl</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>are</td>
<td>be</td>
<td>VERB</td>
<td>VBP</td>
<td>Mood=Ind</td>
<td>Tense=Pres</td>
</tr>
<tr>
<td>3-4</td>
<td>parall</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>parallel</td>
<td>parallel</td>
<td>NOUN</td>
<td>NNS</td>
<td>2</td>
<td>nsubj</td>
</tr>
<tr>
<td>4</td>
<td>s</td>
<td>s</td>
<td>X</td>
<td></td>
<td>Number=Plur</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>to</td>
<td>to</td>
<td>TO</td>
<td>_</td>
<td>_</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>draw</td>
<td>draw</td>
<td>VERB</td>
<td>VB</td>
<td>VerbForm=Inf</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>here</td>
<td>here</td>
<td>ADV</td>
<td>RB</td>
<td>PronType=Dem</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>between</td>
<td>between</td>
<td>ADP</td>
<td>IN</td>
<td>_</td>
<td>9</td>
</tr>
<tr>
<td>9-10</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>9</td>
<td>game</td>
<td>game</td>
<td>NOUN</td>
<td>NNS</td>
<td>3</td>
<td>nmod</td>
</tr>
<tr>
<td>10</td>
<td>s</td>
<td>s</td>
<td>X</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>11</td>
<td>and</td>
<td>and</td>
<td>CCONJ</td>
<td>CC</td>
<td>_</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>our</td>
<td>we</td>
<td>PRON</td>
<td>PRP$</td>
<td>Number=Plur</td>
<td>Person=1</td>
</tr>
<tr>
<td>13-14</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>13</td>
<td>every</td>
<td>every</td>
<td>ADJ</td>
<td>JJ</td>
<td>_</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>day</td>
<td>day</td>
<td>X</td>
<td>_</td>
<td>Degree=Pos</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>lives</td>
<td>life</td>
<td>NOUN</td>
<td>NNS</td>
<td>Number=Plur</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>.</td>
<td>.</td>
<td>FUNCT</td>
<td>.</td>
<td>_</td>
<td>2</td>
</tr>
</tbody>
</table>
To Sum Up

• Novel data structure:
  • Integration of the morphological internal structure of words into a UD-like sentence representation.
  • To enhance comparability of languages that express comparable meaning through different grammatical strategies.
  • Focus on cross-lingual correspondence of morphs.

• Case study of 10 languages, leading to a prototype of methodology to manipulate UD treebanks.

• Existing segmentation resources employed:
  • Approach that ties the quality of our data to that of the employed resources.
  • Some limitations observed.
References


