Tutorial on Universal Dependencies

Word segmentation and morphological annotation

Joakim Nivre\textsuperscript{1}  Daniel Zeman\textsuperscript{2}  Filip Ginter\textsuperscript{3}  Francis M. Tyers\textsuperscript{4,5}

\textsuperscript{1}Department of Linguistics and Philology, Uppsala University, Sweden
\textsuperscript{2}Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia
\textsuperscript{3}Department of Information Technology, University of Turku, Finland
\textsuperscript{4}Giela ja kultuvrra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway
\textsuperscript{5}Arvutiteaduse instituut, Tartu Ülikool, Estonia
Morphological Annotation in UD

- Tokenization / word segmentation
- Lemmatization
- Universal part-of-speech tags
- Universal features
- Language-specific features
- Errors in text
“María, I love you!” Juan exclaimed.

«¡María, te amo!», exclamó Juan.

- Classic tokenization:
  - Separate punctuation from words
  - Recognize certain clusters of symbols like “…”
  - Perhaps keep together things like `user@mail.x.edu`
Word Segmentation

Let's go to the sea.

Vámonos al mar.
   VERB?  X  NOUN  PUNCT

Vamos nos a el mar.
   VERB  PRON  ADP  DET  NOUN  PUNCT

- Syntactic word vs. orthographic word
- Multi-word tokens
- Two-level scheme:
  - Tokenization (low level, punctuation, concatenative)
  - Word segmentation (higher level, not necessarily concatenative)
• Lexicalist hypothesis:
  • Words (not morphemes) are the basic units in syntax
  • Words enter in dependency relations
  • Words are forms of lemmas and have morphological features

• Orthographic vs. syntactic word
  • Syntactically autonomous part of orthographic word
  • Contractions \((al = a + el)\)
  • Clitics \((vámonos = vamos + nos)\)
    • ¿A qué hora \textit{nos vamos} mañana?
    • \textit{Nos despertamos a las cinco.}
      “We wake up at five.”
    • \textit{Nuestro guía nos despierta a las cinco.}
      “Our guide wakes us up at five.”
He abdicated in favour of his son Baudouin.

یتنازل ʿan al-ʿarši li+ibni+hi بودوان
yatanāzalu ʿan al-ʿarši li+ibni+hi Baudouin
surrendered on the throne to son his Baudouin
We are now in Valencia.

現在我們在瓦倫西亞。
Xiànzàiwǒmenzàiwǎlúnxīyǎ.
We are now in Valencia.

現在我們在瓦倫西亞。
Xiànzàiwǒmenzàiwǎlúnxīyǎ.
We are now in Valencia.
I went to the beauty salon of Kyōdō [, Beyond-R.]
I went to the beauty salon of Kyōdō [Beyond-R.]
I went to the beauty salon of Kyōdō [Beyond-R.]
All the concrete country roads are the result of...

Tất cả đường bê tông nội đồng là thành quả ...

- Spaces delimit monosyllabic morphemes, not words.
- Multiple syllables without space occur in loanwords (bê tông).
- Spaces are allowed to occur word-internally in Vietnamese UD.
Il touche environ 100 000 sesterces par an.
Fixed Expressions

One syntactic word spans several orthographic words?

*I am still very satisfied.*
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.

• Vietnamese writing system
  • Words with spaces
  • Very restricted set of exceptions (numbers)
  • Special relations elsewhere (fixed, compound)
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.
  • Border case? Keep orthographic words (if they exist).
Word Segmentation Summary

• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
  • Hard time finding dependency relation? Don’t split!
    • Or not hard time but the relation would be compound, flat, fixed or goeswith.
  • Border case? Keep orthographic words (if they exist).

• Words with spaces
  • Vietnamese writing system
  • Very restricted set of exceptions (numbers)
  • Special relations elsewhere (fixed, compound)
## Recoverability: CoNLL-U Format

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>...</th>
<th>HEAD</th>
<th>_</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Vámonos</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>Vamos</td>
<td>ir</td>
<td>VERB</td>
<td>...</td>
<td>0</td>
<td>root</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>nos</td>
<td>nosotros</td>
<td>PRON</td>
<td>...</td>
<td>1</td>
<td>obj</td>
<td>_</td>
</tr>
<tr>
<td>3-4</td>
<td>al</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>ADP</td>
<td>...</td>
<td>5</td>
<td>case</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>el</td>
<td>el</td>
<td>DET</td>
<td>...</td>
<td>5</td>
<td>det</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>mar</td>
<td>mar</td>
<td>NOUN</td>
<td>...</td>
<td>1</td>
<td>obl</td>
<td>_</td>
</tr>
</tbody>
</table>
| 6  | .      | .      | PUNCT  | ... | 1    | punct | _ | _ | SpaceAfter=No
Recoverability: CoNLL-U Format

# text = Vámonos al mar.
# text_en = Let’s go to the sea.

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>...</th>
<th>HEAD</th>
<th>_ MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Vámonos</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>Vamos</td>
<td>ir</td>
<td>VERB</td>
<td>...</td>
<td>0</td>
<td>root</td>
</tr>
<tr>
<td>2</td>
<td>nos</td>
<td>nosotros</td>
<td>PRON</td>
<td>...</td>
<td>1</td>
<td>obj</td>
</tr>
<tr>
<td>3-4</td>
<td>al</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>ADP</td>
<td>...</td>
<td>5</td>
<td>case</td>
</tr>
<tr>
<td>4</td>
<td>el</td>
<td>el</td>
<td>DET</td>
<td>...</td>
<td>5</td>
<td>det</td>
</tr>
<tr>
<td>5-6</td>
<td>mar.</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>mar</td>
<td>mar</td>
<td>NOUN</td>
<td>...</td>
<td>1</td>
<td>obl</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>...</td>
<td>1</td>
<td>punct</td>
</tr>
</tbody>
</table>
Tokenization vs. Multi-word Tokens Summary

- Punctuation involved? Low level!
Tokenization vs. Multi-word Tokens Summary

- Punctuation involved? Low level!
- Boundary between two letters? Typically high level.
  - Exceptions: Chinese, Japanese.
Tokenization vs. Multi-word Tokens Summary

- Punctuation involved? Low level!
- Boundary between two letters? Typically high level.
  - Exceptions: Chinese, Japanese.
- Non-concatenative? High level!
<table>
<thead>
<tr>
<th>Lemmas</th>
<th>UPOS tags</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Basic or citation form (⇒ it is an existing word in most cases)
• Disambiguating ids, if available, go to MISC
• Derivational vs. inflectional morphology (if participles are ADJ, their lemma should not be infinitive)
Lemmas

within a year Algeria will become an islamic state

13  do  do  ADP  ...  LId=do-1
14  roka  rok  NOUN  ...  _
15  se  se  PRON  ...  LGloss=(zvr._zájmeno/částice)
16  Alžírsko  Alžírsko  PROP  ...  _
17  stane  stát  VERB  ...  LId=stát-2
18  islámským  islámský  ADJ  ...  _
19  státem  stát  NOUN  ...  LId=stát-1|LGloss=(státní_útvar

• Basic or citation form
• Disambiguating ids, if available, go to MISC
### Part-of-Speech Tags

<table>
<thead>
<tr>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>adposition</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>auxiliary</td>
</tr>
<tr>
<td>INTJ</td>
<td>interjection</td>
<td>coordinator</td>
</tr>
<tr>
<td>NOUN</td>
<td>com. noun</td>
<td>determiner</td>
</tr>
<tr>
<td>PROPN</td>
<td>prop. noun</td>
<td>numeral</td>
</tr>
<tr>
<td>VERB</td>
<td>verb</td>
<td>particle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pronoun</td>
</tr>
<tr>
<td></td>
<td></td>
<td>subordinator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PUNCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SYM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

- Taxonomy of 17 universal POS tags
- All languages use the same inventory
  - Not all tags have to be used by all languages
  - Need extensions? Use features!
Part-of-Speech Tags

- Traditionally a mixture of morphological, syntactic/distributional and semantic/notional criteria
- Prefer grammatical > semantic criteria
  - Language-particular definition of a category
- But the **name** of the category is universal
  - Translated words: overlapping categories, but not perfect match
    - UPOS of English *dog* is **NOUN**; so is French *chien* or Russian *собака*
- Preferably POS is encoded in lexicon, not heavily usage-dependent
  - But not for incompatible syntactic functions
    - (e.g. **PRON** vs. **SCONJ**)
## Features

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Inflectional (“Nominal”)</th>
<th>Inflectional (“Verbal”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PronType</td>
<td>Gender</td>
<td>VerbForm</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
<td>Mood</td>
</tr>
<tr>
<td>Poss</td>
<td>Number</td>
<td>Tense</td>
</tr>
<tr>
<td>Reflect</td>
<td>Case</td>
<td>Aspect</td>
</tr>
<tr>
<td>Foreign</td>
<td>Definite</td>
<td>Voice</td>
</tr>
<tr>
<td>Abbr</td>
<td>Degree</td>
<td>Evident</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polarity</td>
</tr>
</tbody>
</table>

- 21 features, each with a number of possible *values*
- Languages select relevant features
- May add language-specific features or values
### Language-Specific Features

Three types of infinitives in Finnish:

**Example:** *olla* “to be”

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>olla</em></td>
<td><em>ollessa</em></td>
<td><em>olemassa</em></td>
</tr>
<tr>
<td><em>ollen</em></td>
<td><em>olemaan</em></td>
<td><em>olemasta</em></td>
</tr>
<tr>
<td></td>
<td><em>olemalla</em></td>
<td><em>olematta</em></td>
</tr>
<tr>
<td>Joku</td>
<td>yrittää</td>
<td>piristää</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Someone</td>
<td>tries</td>
<td>to-uplift</td>
</tr>
<tr>
<td>PRON</td>
<td>VERB</td>
<td>VERB</td>
</tr>
</tbody>
</table>

- **VerbForm=Fin**
- **VerbForm=Inf**
- **Inf3**
- **Mood=Ind**
- **InfForm=1**
- **InfForm=3**
- **Tense=Pres**
- **Case=Ade**
Language-Specific Features

Joku yrittää piristää itseään värjäämällä hiuksensa
Someone tries to-uplift oneself by-staining their-hair

VerbForm=Fin Mood=Ind Tense=Pres
VerbForm=Inf

VerbForm=Inf Mood=Ind Tense=Pres
VerbForm=Inf
VerbForm=Inf Mood=Ind Tense=Pres
InfForm=1
VerbForm=Inf InfForm=3
VerbForm=Inf Case=Ade
VerbForm=Inf Case=Ade
Czech adjectives agree with nouns in gender.

velký      bratr
big         brother
ADJ         NOUN
Gender=Masc Gender=Masc

velká      sestra
big         sister
ADJ         NOUN
Gender=Fem  Gender=Fem
Possessive adjectives: agreement gender vs. lexical gender

<table>
<thead>
<tr>
<th></th>
<th>father’s</th>
<th>brother</th>
<th>mother’s</th>
<th>brother</th>
</tr>
</thead>
<tbody>
<tr>
<td>otcův</td>
<td>otcova</td>
<td>bratr</td>
<td>matčin</td>
<td>bratr</td>
</tr>
<tr>
<td>ADJ</td>
<td>ADJ</td>
<td>NOUN</td>
<td>ADJ</td>
<td>NOUN</td>
</tr>
<tr>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
</tr>
<tr>
<td>otcův</td>
<td>bratr</td>
<td>matčin</td>
<td>bratr</td>
<td></td>
</tr>
<tr>
<td>father’s</td>
<td>brother</td>
<td>mother’s</td>
<td>brother</td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>NOUN</td>
<td>ADJ</td>
<td>NOUN</td>
<td></td>
</tr>
<tr>
<td>Gender=Masc</td>
<td>Gender=Fem</td>
<td>Gender=Masc</td>
<td>Gender=Fem</td>
<td></td>
</tr>
</tbody>
</table>
Multi-valued Features (Disjunction / Parallel Application)

- Feature can have two or more values
- Interpreted as disjunction
- Example: in some languages, many pronouns function both as interrogative and relative, but some pronouns are only relative. The former will have PronType=Int,Rel
- In other cases, it is desirable to disambiguate by context. Polish którym (form of który “which”) can be Case=Ins, Loc in singular or Dat in plural but we do not want to annotate Case=Dat,Ins,Loc!
- All values of the feature/language? Omit the feature completely! Polish: Gender=Fem,Masc,Neut. Spanish: Gender=Fem,Masc
Multi-valued Features (Serial Application)

- Currently used in Turkish (language-specific values)
- Two or more morphemes in chain, affecting the same feature
- Example: Voice=CauPass (causative + passive => someone is caused to do something)
  - yanıl “be wrong”
  - yanılmışım Voice=Act “I was wrong”
  - okuru yanılttığını Voice=Cau “mislead the reader”
  - okurlar yanıltılmıştır Voice=CauPass “readers were misled”
Multi-valued Features (Serial Application)

- Currently used in Turkish (language-specific values)
- Two or more morphemes in chain, affecting the same feature
- Example: **Voice=CauPass** (causative + passive => someone is caused to do something)
  - *yanıll “be wrong”*
  - *yanılmışım Voice=Act “I was wrong”*
  - *okuru yanılttığını Voice=Cau “mislead the reader”*
  - *okurlar yanıltılmıştır Voice=CauPass “readers were misled”*
  - Hypothetical: **Voice=PassCau** (not used in Turkish) could mean “to cause something to be done by someone”
Features Apply to Individual Words

Future tense in Spanish and German: no Tense=Fut in German!

<table>
<thead>
<tr>
<th>Spanish</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dormirá</td>
<td>Er</td>
</tr>
<tr>
<td>He-will-sleep</td>
<td>He will</td>
</tr>
<tr>
<td>VERB</td>
<td>AUX</td>
</tr>
<tr>
<td>VerbForm=Fin</td>
<td>VerbForm=Fin</td>
</tr>
<tr>
<td>Mood=Ind</td>
<td>Mood=Ind</td>
</tr>
<tr>
<td>Tense=Fut</td>
<td>Tense=Pres</td>
</tr>
<tr>
<td>Number=Sing</td>
<td>Number=Sing</td>
</tr>
<tr>
<td>Person=3</td>
<td>Person=3</td>
</tr>
<tr>
<td>Gender=Masc</td>
<td></td>
</tr>
<tr>
<td>Case=Nom</td>
<td></td>
</tr>
</tbody>
</table>
### Participle Types

| некурящий | человек | начавшийся | разговор |
| nekurjaščij | čelovek | načavšijsja | razgovor |
| non-smoking | person | that-has-started | conversation |

- **ADJ**
- **NOUN**
- **VerbForm=Part**
- **Tense=Pres**
- **Gender=Masc**
- **Number=Sing**
- **Case=Nom**

- **VerbForm=Part**
- **Tense=Past**
- **Gender=Masc**
- **Number=Sing**
- **Case=Nom**

- **Sometimes features like **Tense** help distinguish participle types**
- **Not the same tense as with finite verbs (reference point)**
- **But useful because:**
  - **We use known UD primitives rather than language-specific labels such as** VerbForm=PastPart, or even ParticType=Past**
  - **Reasonably close to the grammatical meaning**
Conflicting Traditional Terminologies

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- **VerbForm=Conv** – *converb*, *transgressive*, *adverbial participle*, *gerund*
Conflicting Traditional Terminologies

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- \texttt{VerbForm=Conv} – \textit{converb}, \textit{transgressive}, \textit{adverbial participle}, \textit{gerund}
- \textit{Gerund} (\texttt{VerbForm=Ger})
  - English: close to verbal nouns (\texttt{VerbForm=Vnoun})
  - Spanish: more like present participle (\texttt{VerbForm=Part | Tense=Pres})
  - Slavic: \textit{converb} (\texttt{VerbForm=Conv})
Conflicting Traditional Terminologies

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- `VerbForm=Conv` – *converb*, *transgressive*, *adverbial participle*, *gerund*
- **Gerund** (`VerbForm=Ger`)
  - English: close to verbal nouns (`VerbForm=Vnoun`)
  - Spanish: more like present participle (`VerbForm=Part | Tense=Pres`)
  - Slavic: *converb* (`VerbForm=Conv`)
- **Aorist**
  - Ancient Greek, Turkish: neutral non-past tense (they use a language-specific value `Tense=Aor`)
  - Slavic languages: simple past tense (`Tense=Past`)
Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
Errors in Underlying Text

• Currently not covered by the guidelines
• We do not want to hide errors (learning robust parsers!)
• Possibilities:
  • Typo not involving word boundary
    • FORM = annotation; LEMMA = annotation; FEATS: Typo=Yes; MISC: Correct=annotation
• Wrongly split word: annotation
  • Goes with
• Wrongly merged words: the car
• Fix tokenization (i.e. two lines); first line MISC: SpaceAfter=No | CorrectSpaceAfter=Yes
• Sentence segmentation can be affected, too!
Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Typo not involving word boundary
  - FORM = anotation; LEMMA = annotation; FEATS: Typo=Yes; MISC: Correct=annotation
  - goeswith
  - annotation
    - X
    - X
- Wrongly split word:
Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
  - Typo not involving word boundary
    - FORM = \textit{anotation}; LEMMA = \textit{annotation}; FEATS: Typo=Yes; MISC: Correct=\textit{annotation}
  - Wrongly split word:
  - Wrongly merged words: \textit{thecar}
    - Fix tokenization (i.e. two lines); first line MISC: SpaceAfter=No | CorrectSpaceAfter=Yes
    - Sentence segmentation can be affected, too!
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo (*the car iss produced...*)
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo (*the car iss produced...*)
  - Not obvious what is correct
    - *the car is*
    - *the cars are*
Errors in Underlying Text

• Wrong morphology: *the cars is produced in Detroit*
  • Not like normal typo (*the car iss produced...*)
  • Not obvious what is correct
    • *the car is*
    • *the cars are*

• Suggestion: select which word to fix, e.g. *cars to car*

• FORM = *cars*; FEATS: Number=Plur; MISC: Correct=car | CorrectNumber=Sing
Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo (*the car iss produced...*)
  - Not obvious what is correct
    - *the car is*
    - *the cars are*

- Suggestion: select which word to fix, e.g. *cars to car*
- FORM = *cars*; FEATS: Number=Plur; MISC: Correct=car | CorrectNumber=Sing

- cs: *viděl moři* “he saw the sea”
  - Should be *moře*
  - Would be Case=Acc (disambiguated from Case=Acc,Gen,Nom,Voc)
  - This form is Case=Dat,Loc (but which one?)

- *cestoval k moři* “he traveled to the sea” Case=Dat
- *plavil se po moři* “he sailed the sea” Case=Loc
Questions?