Universal Dependencies
A Framework for Morphosyntactic Annotation

Joakim Nivre
Uppsala University
Department of Linguistics and Philology

Based on collaborative work with Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Natalia Silveira, Slav Petrov, Sampo Pyysalo, Sebastian Schuster, Reut Tsarfaty, Francis Tyers, Daniel Zeman and many others
Introduction
Introduction

Linguistic annotation is tremendously useful
  • Computational linguistics use it for machine learning and evaluation
  • Corpus linguistics use it for studying complex linguistic phenomena
Introduction

Linguistic annotation is tremendously useful
  • Computational linguistics use it for machine learning and evaluation
  • Corpus linguistics use it for studying complex linguistic phenomena

Linguistic annotation is available for many languages
  • Multilingual evaluation to test generality of computational models
  • Cross-lingual learning to support under-resourced languages
  • Empirically grounded linguistic typology and comparative linguistics
Introduction

Linguistic annotation is tremendously useful
- Computational linguistics use it for machine learning and evaluation
- Corpus linguistics use it for studying complex linguistic phenomena

Linguistic annotation is available for many languages
- Multilingual evaluation to test generality of computational models
- Cross-lingual learning to support under-resourced languages
- Empirically grounded linguistic typology and comparative linguistics

Linguistic annotation guidelines vary across languages 🚨
A cat chases rats and mice.
A cat chases rats and mice

En katt jagar råttor och möss

Toutefois, les filles adorent les desserts.
En katt jagar råttor och möss

En kat jager rotter og mus

A cat chases rats and mice
A cat chases rats and mice.
A cat chases rats and mice.

En katt jagar råttor och möss.

En kat jager rotter og mus.

Toutefois, les filles adorent les desserts.
Why is this a problem?
Why is this a problem?

- Hard to compare empirical results across languages
Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
Why is this a problem?

• Hard to compare empirical results across languages
• Hard to usefully do cross-lingual structure learning
• Hard to evaluate cross-lingual learning
Why is this a problem?

• Hard to compare empirical results across languages
• Hard to usefully do cross-lingual structure learning
• Hard to evaluate cross-lingual learning
• Hard to build and maintain multilingual systems
Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to start annotation projects for new languages
Toutefois, les filles adorent les desserts.

toutefois, le fille adorer les dessert.
Universal Dependencies

http://universaldependencies.org

Toutefois, les filles adorent les desserts.

ADV  PUNCT  DET  NOUN  VERB  DET  NOUN  PUNCT
Definite=Def  Gender=Fem  Number=Plur  Number=Plur  Person=3  Tense=Pres  Number=Plur  Number=Plur

Part-of-speech tags Google
Toutefois, les filles adorent les desserts.

Definite=Def  Gender=Fem  Number=Plur  Person=3  Tense=Pres

ADV  PUNCT  DET  NOUN  VERB  DET  NOUN  PUNCT
Universal Dependencies

http://universaldependencies.org

Brief history of UD:

• Kick-off meeting at EACL in Gothenburg, April 2014
• Guidelines v1, October 2014
• Treebank releases every 6 months (v1.0–v1.4)
• Guidelines v2, December 2016
• Latest treebank release v2.4, May 2019

Open community effort – anyone can contribute!
**Release v2.4 (May, 2019):**

- 83 languages
- 146 treebanks
- 22.4 million words
- 323 contributors
Release v2.4 (May, 2019):
- 83 languages
- 146 treebanks
- 22.4 million words
- 323 contributors

Universal Guidelines Group

Release and Documentation Task Force

Chief Cat Herder
Plan for the Talk

The annotation framework:
- Basic design principles
- Word segmentation
- Morphological annotation
- Syntactic annotation

Treebanks and other resources

Applications in NLP and linguistics
Annotation Framework
Goals and Requirements
Goals and Requirements

Cross-linguistically consistent morphosyntactic annotation
Goals and Requirements

Cross-linguistically consistent morphosyntactic annotation

Support multilingual research in NLP and linguistics
  • Meaningful linguistic analysis within and across languages
  • Morphosyntactic parsing in monolingual and cross-lingual settings
  • Useful information for downstream language understanding tasks
Goals and Requirements

Cross-linguistically consistent morphosyntactic annotation

Support multilingual research in NLP and linguistics

• Meaningful linguistic analysis within and across languages
• Morphosyntactic parsing in monolingual and cross-lingual settings
• Useful information for downstream language understanding tasks

Build on common usage and existing de facto standards
Goals and Requirements

Cross-linguistically consistent morphosyntactic annotation

Support multilingual research in NLP and linguistics
  • Meaningful linguistic analysis within and across languages
  • Morphosyntactic parsing in monolingual and cross-lingual settings
  • Useful information for downstream language understanding tasks

Build on common usage and existing de facto standards

Complement – not replace – language-specific schemes
The UD Philosophy
Maximize parallelism – but don’t overdo it

- Don’t annotate the same thing in different ways
- Don’t make different things look the same
- Don’t annotate things that are not there
The UD Philosophy

Maximize parallelism – but don’t overdo it
  •  Don’t annotate the same thing in different ways
  •  Don’t make different things look the same
  •  Don’t annotate things that are not there

Universal taxonomy with language-specific elaboration
  •  Languages select from a common pool of categories
  •  Allow language-specific extensions
Design Principles
Design Principles

Dependency – syntactic relations

- Widely used in practical NLP systems
- Available in treebanks for many languages
Design Principles

Dependency – syntactic relations
  • Widely used in practical NLP systems
  • Available in treebanks for many languages

Lexicalism
  • Basic annotation units are words – syntactic words
  • Words have morphological properties
  • Words enter into syntactic relations
Design Principles

Dependency – syntactic relations
• Widely used in practical NLP systems
• Available in treebanks for many languages

Lexicalism
• Basic annotation units are words – syntactic words
• Words have morphological properties
• Words enter into syntactic relations

Recoverability
• Transparent mapping from input text to word segmentation
Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words
Word Segmentation

What is a word?
  • Single part-of-speech tag
  • Real syntactic relation

Two-level segmentation
  • Represent orthographic tokens in addition to syntactic words
Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

<table>
<thead>
<tr>
<th>Text</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>aux</td>
<td>à les</td>
</tr>
</tbody>
</table>
What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

<table>
<thead>
<tr>
<th>Text</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>aux</td>
<td>à les</td>
</tr>
<tr>
<td>dámelo</td>
<td>da me lo</td>
</tr>
</tbody>
</table>
Word Segmentation

What is a word?
- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation
- Represent orthographic tokens in addition to syntactic words

<table>
<thead>
<tr>
<th>Text</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>aux</td>
<td>à les</td>
</tr>
<tr>
<td>dáme lo</td>
<td>da me lo</td>
</tr>
<tr>
<td>השמהשמש</td>
<td>שֶׁמַּהְשֶׁמָּה</td>
</tr>
</tbody>
</table>
Word Segmentation

What is a word?
  • Single part-of-speech tag
  • Real syntactic relation

Two-level segmentation
  • Represent orthographic tokens in addition to syntactic words

<table>
<thead>
<tr>
<th>Text</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>aux</td>
<td>à les</td>
</tr>
<tr>
<td>dámeló</td>
<td>da me lo</td>
</tr>
<tr>
<td>וישמעהנש</td>
<td>יְשֵׁם הָשָׁמֶשׁ</td>
</tr>
<tr>
<td>大阪国際会議場</td>
<td>大阪 国際 会議場</td>
</tr>
</tbody>
</table>
Morphology

Le chat chasse les chiens.
Morphology

Le chat chasse les chiens.
le chat chasser le chien.

• Lemma representing the semantic content of the word
Morphology

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class
### Morphology

- **Lemma:** representing the semantic content of the word
- **Part-of-speech (POS) Tag:** representing its grammatical class

<table>
<thead>
<tr>
<th></th>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADP</td>
<td></td>
<td>PUNCT</td>
</tr>
<tr>
<td>ADV</td>
<td>AUX</td>
<td></td>
<td>SYM</td>
</tr>
<tr>
<td>INTJ</td>
<td>CCONJ</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>NOUN</td>
<td>DET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROPN</td>
<td>NUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>PART</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRON</td>
<td>SCONJ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Le chat chasse les chiens.
- le chat chasser le chien.

**Example:**

- Osaka *welcomes* the participants of COLING.
- Napoli *accoglie* i partecipanti di CLiC-it.

- Le *chat* *chas* les chiens.
- Le *chat* *chass* le chien.

**Table:**

<table>
<thead>
<tr>
<th>DET</th>
<th>NOUN</th>
<th>PROPN</th>
<th>VERB</th>
<th>DET</th>
<th>PROPN</th>
<th>NOUN</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>le</em></td>
<td><em>chat</em></td>
<td><em>chauv</em></td>
<td><em>chass</em></td>
<td><em>le</em></td>
<td><em>chat</em></td>
<td><em>chauv</em></td>
</tr>
</tbody>
</table>

**Tense:**

- Tense = Present
- Aspect = Imperfect
- Polite = Humble

**Mood:**

- Mood = Indicative
- Mood = Subjunctive
- Mood = Imperative

**Case:**

- Case = Nominative
- Case = Accusative
- Case = Genitive
- Case = Dative
- Case = Vocative

**Number:**

- Number = Singular
- Number = Plural

**Gender:**

- Gender = Masculine
- Gender = Feminine
- Gender = Neuter

**Person:**

- Person = 1st
- Person = 2nd
- Person = 3rd

**Pronoun Type:**

- Pronoun Type = Personal
- Pronoun Type = Possessive
- Pronoun Type = Reflexive
- Pronoun Type = Interrogative
- Pronoun Type = Relative

**Verb Form:**

- Verb Form = Infinitive
- Verb Form = Gerund
- Verb Form = Participle
- Verb Form = Question
- Verb Form = Exclamatory
Morphology

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form
Morphology

<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>NounClass</td>
<td>Tense</td>
</tr>
<tr>
<td>Reflex</td>
<td>Number</td>
<td>Aspect</td>
</tr>
<tr>
<td>Foreign</td>
<td>Case</td>
<td>Voice</td>
</tr>
<tr>
<td>Abbr</td>
<td>Definite</td>
<td>Evident</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
<td>Polarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clusivity</td>
</tr>
</tbody>
</table>

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form
The cat could have chased all the dogs down the street.
The cat could have chased all the dogs down the street.

- Content words are related by syntactic relations
• Content words are related by syntactic relations
• Function words attach to the content word they modify
• Content words are related by syntactic relations
• Function words attach to the content word they modify
• Punctuation attach to head of phrase or clause
The dog was chased by the cat.
The dog was chased by the cat.

Hunden jagades av katten.

Pes byl honěn kočkou.
The dog was chased by the cat.
Syntactic Relations
Syntactic Relations

Taxonomy of 37 universal syntactic relations

- Three types of structures: nominals, clauses, modifiers
- Core arguments vs. other dependents (not arguments vs. adjuncts)
- Language-specific subtypes
Syntactic Relations

Taxonomy of 37 universal syntactic relations

- Three types of structures: nominals, clauses, modifiers
- Core arguments vs. other dependents (not arguments vs. adjuncts)
- Language-specific subtypes

Basic and enhanced representations

- Basic representation forms a (possibly non-projective) tree
- Additional relations in the enhanced representation
# Syntactic Relations

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Clause</th>
<th>Modifier Word</th>
<th>Function Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Predicate Dep</td>
<td>nsubj</td>
<td>csubj</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>obj</td>
<td>ccomp</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iobj</td>
<td>xcomp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Core Predicate Dep</td>
<td>obl</td>
<td>advcl</td>
<td>advmod*</td>
<td>aux</td>
</tr>
<tr>
<td></td>
<td>vocative</td>
<td>advcl</td>
<td>discourse</td>
<td>cop</td>
</tr>
<tr>
<td></td>
<td>expl</td>
<td></td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td></td>
<td>dislocated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal Dep</td>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
<td>det</td>
</tr>
<tr>
<td></td>
<td>appos</td>
<td>acl</td>
<td>amod</td>
<td>clf</td>
</tr>
<tr>
<td></td>
<td>nummod</td>
<td>acl</td>
<td>amod</td>
<td>case</td>
</tr>
<tr>
<td>Coordination</td>
<td>conj</td>
<td>fixed</td>
<td>parataxis</td>
<td>orphan</td>
</tr>
<tr>
<td></td>
<td>cc</td>
<td>flat</td>
<td>list</td>
<td>goeswith</td>
</tr>
<tr>
<td></td>
<td></td>
<td>compound</td>
<td></td>
<td>reparandum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MWE</td>
<td>Loose</td>
<td>Special</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>fixed</td>
<td>parataxis</td>
<td>orphan</td>
<td>punct</td>
</tr>
<tr>
<td></td>
<td>flat</td>
<td>list</td>
<td>goeswith</td>
<td>root</td>
</tr>
<tr>
<td></td>
<td>compound</td>
<td></td>
<td>reparandum</td>
<td>dep</td>
</tr>
</tbody>
</table>
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

<table>
<thead>
<tr>
<th>Universal</th>
<th>Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>acl:relcl</td>
</tr>
</tbody>
</table>
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

<table>
<thead>
<tr>
<th>Universal</th>
<th>Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>acl:relcl</td>
</tr>
<tr>
<td>compound</td>
<td>compound:prt</td>
</tr>
</tbody>
</table>
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

<table>
<thead>
<tr>
<th>Universal</th>
<th>Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>acl:relcl</td>
</tr>
<tr>
<td>compound</td>
<td>compound:prt</td>
</tr>
<tr>
<td>nmod</td>
<td>nmod:poss</td>
</tr>
</tbody>
</table>
A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

<table>
<thead>
<tr>
<th>Universal</th>
<th>Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>acl:relcl</td>
</tr>
<tr>
<td>compound</td>
<td>compound:prr</td>
</tr>
<tr>
<td>nmod</td>
<td>nmod:poss</td>
</tr>
<tr>
<td>flat</td>
<td>flat:name</td>
</tr>
</tbody>
</table>
Resources
UD Guidelines

- Basic principles
  - Tokenization and word segmentation
  - Morphology
  - Syntax
  - Enhanced dependencies
  - CoNLL-U format and its extensions

- Annotation guidelines
  - Nominials
  - Complex clauses
  - Comparative constructions – working group materials
  - Other constructions

- Documentation of tags, features and relations
  - POS tags (single document)
  - Features (single document)
  - Layered features
  - Language-specific features
  - Syntactic relations (single document)
  - Relation subtypes
  - Conversion from other tagsets to UD tags and features

This is the online documentation of UD guidelines v2 (2016-12-01). For earlier versions, see archived v1 guidelines and changes from v1 to v2.

Download

- The next release (v2.3) is scheduled for November 15, 2019 (data freeze on November 1).
- Version 2.3 treebanks are available at http://hdl.handle.net/11234/1-2988. 146 treebanks, 83 languages, released May 15, 2019.
- Version 2.2 treebanks are archived at http://hdl.handle.net/11234/1-2837. 122 treebanks, 71 languages, released July 1, 2018.
- Version 2.1 treebanks are archived at http://hdl.handle.net/11234/1-2515. 102 treebanks, 60 languages, released November 15, 2017.
- Version 2.0 treebanks are archived at http://hdl.handle.net/11234/1-1933. 70 treebanks, 50 languages, released March 1, 2017.
- Version 1.1 treebanks are archived at http://hdl.handle.net/11234/1-2194. 81 treebanks, 49 languages, released May 18, 2017.
- Version 1.0 treebanks are archived at http://hdl.handle.net/11234/1-1827. 64 treebanks, 47 languages, released November 15, 2016.
- Version 1.3 treebanks are archived at http://hdl.handle.net/11234/1-1699. 54 treebanks, 40 languages, released May 15, 2016.
- Version 1.2 treebanks are archived at http://hdl.handle.net/11234/1-1548. 37 treebanks, 33 languages, released November 15, 2015.
- Version 1.1 treebanks are archived at http://hdl.handle.net/11234/1-1468. 10 treebanks, 10 languages, released January 15, 2016.
- In general, we intend to have regular treebank releases every six months. The v2.0 and v2.2 releases were brought forward because of their usage in the CoNLL-2017 and 2018 Multilingual Parsing Shared Tasks.

Upcoming UD Languages

http://universaldependencies.org
Language Family

- Niger-Congo
- Austronesian
- Turkic
- Dravidian
- Sino-Tibetan
- Afro-Asiatic: 7
- Uralic: 9
- Other: 10
- Indo-European: 45
Treebank Size
## Ancient Treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>#Treebanks</th>
<th>#Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkadian</td>
<td>1</td>
<td>1k</td>
</tr>
<tr>
<td>Ancient Greek</td>
<td>2</td>
<td>416k</td>
</tr>
<tr>
<td>Classical Chinese</td>
<td>1</td>
<td>55k</td>
</tr>
<tr>
<td>Coptic</td>
<td>1</td>
<td>25k</td>
</tr>
<tr>
<td>Gothic</td>
<td>1</td>
<td>55k</td>
</tr>
<tr>
<td>Latin</td>
<td>3</td>
<td>582k</td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>1</td>
<td>57k</td>
</tr>
<tr>
<td>Old French</td>
<td>1</td>
<td>170k</td>
</tr>
<tr>
<td>Old Russian</td>
<td>2</td>
<td>164k</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>1</td>
<td>1k</td>
</tr>
</tbody>
</table>
Documentation and Tools

Documentation of guidelines and treebanks
  • Universal annotation guidelines
  • Language documentation
  • Treebank documentation

Tools for search, validation, annotation and parsing
  • Treebank search tools
  • Data validation tools
  • Annotation and parsing tools
Applications
## Syntactic Parsing

### Straka et al. (2015) All

<table>
<thead>
<tr>
<th>Language</th>
<th>POS</th>
<th>PM</th>
<th>FM</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Haverinen et al., 2013b)</td>
<td>94.3</td>
<td>90.5</td>
<td>89.0</td>
<td>81.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Stanford Dependencies (SD)</td>
<td>96.3</td>
<td>93.4</td>
<td>90.3</td>
<td>80.1</td>
<td>84.1</td>
</tr>
<tr>
<td>Universal Dependencies (UD)</td>
<td>96.0</td>
<td>93.1</td>
<td>90.5</td>
<td>81.0</td>
<td>85.0</td>
</tr>
<tr>
<td>Pure Universal Dependencies (Pure UD)</td>
<td>96.0</td>
<td>93.1</td>
<td>90.5</td>
<td>81.5</td>
<td>84.7</td>
</tr>
</tbody>
</table>

### Haverinen et al. (2015) Finnish

<table>
<thead>
<tr>
<th>Language</th>
<th>POS</th>
<th>PM</th>
<th>FM</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Haverinen et al., 2013b)</td>
<td>94.3</td>
<td>90.5</td>
<td>89.0</td>
<td>81.4</td>
<td>85.2</td>
</tr>
<tr>
<td>Stanford Dependencies (SD)</td>
<td>96.3</td>
<td>93.4</td>
<td>90.3</td>
<td>80.1</td>
<td>84.1</td>
</tr>
<tr>
<td>Universal Dependencies (UD)</td>
<td>96.0</td>
<td>93.1</td>
<td>90.5</td>
<td>81.0</td>
<td>85.0</td>
</tr>
<tr>
<td>Pure Universal Dependencies (Pure UD)</td>
<td>96.0</td>
<td>93.1</td>
<td>90.5</td>
<td>81.5</td>
<td>84.7</td>
</tr>
</tbody>
</table>

### Johannsen et al. (2015) Danish

<table>
<thead>
<tr>
<th>Language</th>
<th>LAS</th>
<th>UAS</th>
<th>LA</th>
<th>LAS</th>
<th>UAS</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDT Dev</td>
<td>85.20</td>
<td>89.38</td>
<td>90.83</td>
<td>84.59</td>
<td>89.46</td>
<td>90.61</td>
</tr>
<tr>
<td>CDT Test</td>
<td>84.38</td>
<td>88.70</td>
<td>90.17</td>
<td>84.11</td>
<td>89.44</td>
<td>90.69</td>
</tr>
<tr>
<td>UD-DANISH Dev</td>
<td>81.87</td>
<td>84.51</td>
<td>92.10</td>
<td>65.87</td>
<td>81.57</td>
<td>75.71</td>
</tr>
<tr>
<td>UD-DANISH Test</td>
<td>81.56</td>
<td>84.64</td>
<td>92.00</td>
<td>63.87</td>
<td>80.91</td>
<td>74.54</td>
</tr>
</tbody>
</table>

### Øvrelid et al. (2016) Norwegian

<table>
<thead>
<tr>
<th>Language</th>
<th>LAS</th>
<th>UAS</th>
<th>LA</th>
<th>LAS</th>
<th>UAS</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev Gold</td>
<td>90.15</td>
<td>88.50</td>
<td>92.51</td>
<td>91.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dev Auto</td>
<td>86.73</td>
<td>83.91</td>
<td>89.99</td>
<td>87.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Gold</td>
<td>90.55</td>
<td>88.54</td>
<td>92.76</td>
<td>91.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Auto</td>
<td>86.76</td>
<td>83.86</td>
<td>90.13</td>
<td>87.16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Schuster and Manning (2016) English

<table>
<thead>
<tr>
<th>Genre</th>
<th>LAS</th>
<th>UAS</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question-answers</td>
<td>92.0</td>
<td>95.4</td>
<td>93.7</td>
</tr>
<tr>
<td>Email</td>
<td>91.4</td>
<td>95.8</td>
<td>92.7</td>
</tr>
<tr>
<td>Newsgroups</td>
<td>93.1</td>
<td>96.8</td>
<td>94.0</td>
</tr>
<tr>
<td>Business reviews</td>
<td>92.5</td>
<td>95.9</td>
<td>93.9</td>
</tr>
<tr>
<td>Weblogs</td>
<td>94.5</td>
<td>97.1</td>
<td>95.7</td>
</tr>
<tr>
<td>Entire corpus</td>
<td>92.6</td>
<td>96.1</td>
<td>93.9</td>
</tr>
</tbody>
</table>

### Silveira and Manning (2015) English

<table>
<thead>
<tr>
<th>Genre</th>
<th>full</th>
<th>partial</th>
<th>simple</th>
</tr>
</thead>
<tbody>
<tr>
<td>auxhead</td>
<td>84.37</td>
<td>84.84</td>
<td>84.43</td>
</tr>
<tr>
<td>casehead</td>
<td>84.13%</td>
<td>84.91%</td>
<td>84.86%</td>
</tr>
<tr>
<td>cophead</td>
<td>84.28%</td>
<td>84.53%</td>
<td>84.03%</td>
</tr>
<tr>
<td>markhead</td>
<td>84.27%</td>
<td>84.89%</td>
<td>85.00%</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td>84.69%</td>
</tr>
</tbody>
</table>
Off the Shelf – UDPipe

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, C#, and as a web service. Third-party R CRAN package also exists.

UDPipe is a free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distributed under the CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using Semantic Versioning.

Description of the available methods is available in the API Documentation and the models are described in the UDPipe User's Manual.

Service

The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – explicit written permission of the authors is required for any commercial exploitation of the system. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

**Model:**
- UD 2.0 (description)
- CoNLL17 Baseline UD 2.0 (description)
- UD 1.2 (description)

**Actions:**
- Tag and Lemmatize
- Parse

french-ud-2.0-170801
Cross-Lingual Parsing

- Train a parser for language X on data from language Y
- Usually motivated by a low-resource scenario
- Annotation projection, transfer learning and translation

Reddy et al. (2017) Universal Semantic Parsing

(a) The dependency tree for Disney won an Oscar for the movie Frozen in the Universal Dependencies formalism.

(b) The binarized s-expression for the dependency tree.

(c) The composed lambda-calculus expression.

Gotham and Haug (2018) Glue Semantics for Universal Dependencies
Language Typology

Futrell et al. (2015) Quantifying Word Order Freedom in Dependency Corpora

Conditional entropy
  • Head direction entropy
  • Relation order entropy

Hypothesis test
  • Free word order $\Rightarrow$ Case marking
  • Case marking $\Rightarrow$ Free word order
Language Typology


<table>
<thead>
<tr>
<th>Feature</th>
<th>Languages</th>
<th>Types</th>
<th>Tokens</th>
<th>Most common</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj,Obj,V</td>
<td>342</td>
<td>85.4%</td>
<td>85.7%</td>
<td>SOV: 43.3%</td>
</tr>
<tr>
<td>Subj,V</td>
<td>376</td>
<td>89.4%</td>
<td>90.4%</td>
<td>SV: 79.8%</td>
</tr>
<tr>
<td>Obj,V</td>
<td>387</td>
<td>96.4%</td>
<td>96.4%</td>
<td>VO: 54.8%</td>
</tr>
<tr>
<td>Adp, NP</td>
<td>329</td>
<td>94.8%</td>
<td>95.1%</td>
<td>Prep: 50.4%</td>
</tr>
<tr>
<td>Adj, Noun</td>
<td>334</td>
<td>85.9%</td>
<td>88.0%</td>
<td>AdjN: 68.9%</td>
</tr>
</tbody>
</table>

- Bible translations in 986 languages
- Word alignment and high-precision annotation projection
- Word order statistics and prediction of WALS features
Language Typology

Yu et al. (2019) Dependency Length Minimization vs. Word Order Constraints

![Graph showing dependency length minimization vs. sentence length and freedom of head direction.]
Conclusion
Conclusion

Universal Dependencies

- Framework for morphosyntactic annotation
- Designed to support cross-linguistically consistent annotation
- Treebanks for 83 languages and counting …
Conclusion

Universal Dependencies

- Framework for morphosyntactic annotation
- Designed to support cross-linguistically consistent annotation
- Treebanks for 83 languages and counting …

Future challenges

- Get more languages and more annotated data
- Improve cross-language consistency and documentation
Conclusion

Universal Dependencies

• Framework for morphosyntactic annotation
• Designed to support cross-linguistically consistent annotation
• Treebanks for 83 languages and counting …

Future challenges

• Get more languages and more annotated data
• Improve cross-language consistency and documentation

You can help!
Thanks to all UD contributors!