Tutorial on Universal Dependencies

Introduction

Marie-Catherine de Marneffe[1] **Joakim Nivre**[2] Daniel Zeman[3]

[1]FNRS, Université catholique de Louvain, Belgium

[2]Department of Linguistics and Philology, Uppsala University, Sweden

[3]Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia

What is Universal Dependencies?

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

`https://universaldependencies.org/introduction.html`

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

`https://universaldependencies.org/introduction.html`

A project — but also an annotation scheme, a data repository and a community

The UD Annotation Scheme — Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual NLP and linguistic research
- Build on common usage and existing de facto standards
- Complement – not replace – language-specific schemes

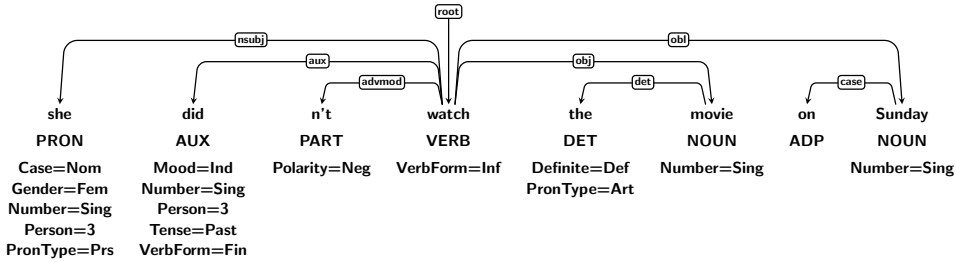
The UD Philosophy

- 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
- Universal taxonomy with language-specific elaboration
  - Languages select from a universal pool of categories
  - Allow language-specific extensions

Design Principles

- Dependency
  - 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

The UD Annotation Scheme



- Part-of-speech tags
- Morphological features
- Syntactic dependencies

The UD Data Repository

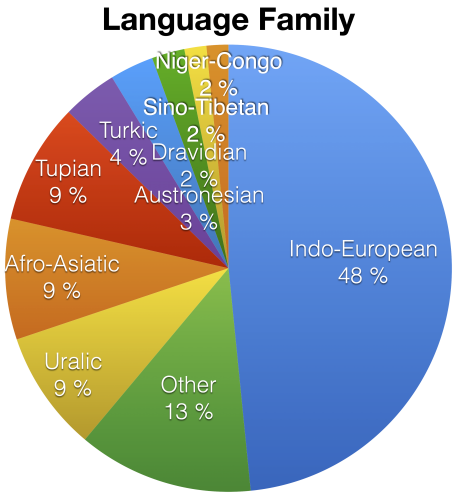
- UD v2.12 in numbers:
  - 245 treebanks
  - 141 languages
  - 30 language families
  - 1.8 million sentences
  - 30.6 million words
  - 554 contributors

Current UD Languages

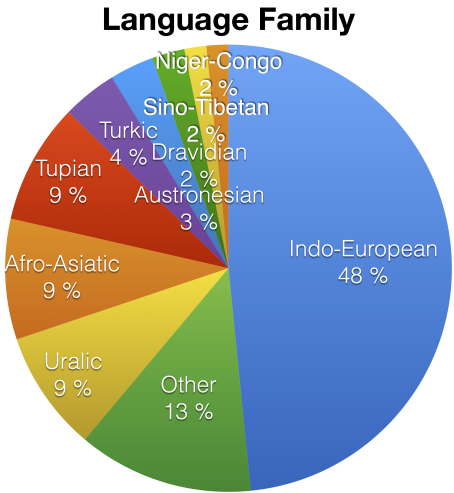
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

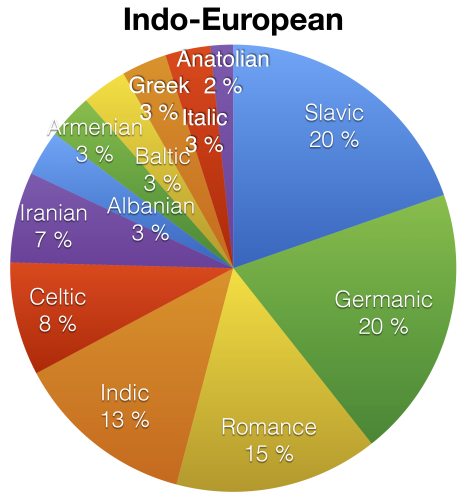
	Abaza	1	<1K		Northwest Caucasian
	Afrikaans	1	49K		IE, Germanic
	Akkadian	2	25K		Afro-Asiatic, Semitic
	Akuntsu	1	1K		Tupian, Tupari
	Albanian	1	<1K		IE, Albanian
	Amharic	1	10K		Afro-Asiatic, Semitic
	Ancient Greek	2	416K		IE, Greek
	Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
	Apurina	1	<1K		Arawakan
	Arabic	3	1,042K		Afro-Asiatic, Semitic
	Armenian	2	94K		IE, Armenian
	Assyrian	1	<1K		Afro-Asiatic, Semitic
	Bambara	1	13K		Mande
	Basque	1	121K		Basque
	Beja	1	<1K		Afro-Asiatic, Cushitic
	Belarusian	1	305K		IE, Slavic
	Bengali	1	<1K		IE, Indic
	Bhojpuri	1	6K		IE, Indic
	Bororo	1	<1K		Bororoan
	Breton	1	10K		IE, Celtic
	Bulgarian	1	156K		IE, Slavic
	Buryat	1	10K		Mongolic
	Cantonese	1	13K		Sino-Tibetan
	Catalan	1	553K		IE, Romance
	Cebuano	1	1K		Austronesian, Central Philippine
	Chinese	6	287K		Sino-Tibetan
	Chukchi	1	6K		Chukotko-Kamchatkan
	Classical Chinese	1	433K		Sino-Tibetan
	Coptic	1	55K		Afro-Asiatic, Egyptian
	Croatian	1	199K		IE, Slavic
	Czech	5	2,247K		IE, Slavic
	Danish	1	100K		IE, Germanic
	Dutch	2	306K		IE, Germanic
	English	10	726K		IE, Germanic
	Erzya	1	20K		Uralic, Mordvin
	Estonian	2	528K		Uralic, Finnic
	Faroese	2	50K		IE, Germanic
	Finnish	4	397K		Uralic, Finnic
	French	7	635K		IE, Romance
	Frisian Dutch	1	3K		Code switching
	Galician	2	164K		IE, Romance
	German	4	3,810K		IE, Germanic

The UD Data Repository



The UD Data Repository

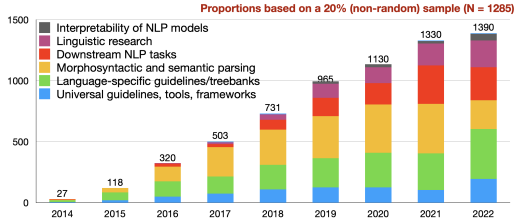




The UD Community — A Big Tent

- Universal guidelines group — universal guidelines, data validation and releases
- Treebank developers — treebanks and language-specific documentation
- Treebank users — research in NLP and linguistics based on UD resources
- Anyone can join!

The UD Community — A Literature Survey



- Categories of research in a sample of 277 publications from 2022 (top 20%):
  - Universal guidelines, tools, frameworks (14%)
  - Language-specific guidelines/treebanks (30%)
  - Morphosyntactic and semantic parsing (17%)
  - Downstream NLP tasks (19%)
  - Interpretability of NLP models (4%)
  - Linguistic research (16%)

Outline

1. Introduction [Joakim]
2. Syntactic annotation [Marie]

   BREAK [10 min]
3. Word segmentation and morphological annotation [Dan]
4. Annotation exercise [Joakim]

   BREAK [10 min]
5. Adding a new language to UD [Dan]
6. Questions and discussion [Marie]