Universal Dependencies

Joakim Nivre Uppsala University joakim.nivre@lingfil.uu.se **Daniel Zeman** Charles University in Prague **Filip Ginter** University of Turku ginter@cs.utu.fi **Francis Tyers** University of Tromsø fty000@post.uit.no

Abstract

Universal Dependencies (UD) is a project that seeks to develop cross-linguistically consistent treebank annotation for many languages. This tutorial gives an introduction to the UD framework and resources, from basic design principles to annotation guidelines and existing treebanks. We also discuss tools for developing and exploiting UD treebanks and survey applications of UD in NLP and linguistics.

1 Tutorial Description

Multilingual research on syntax and parsing has for a long time been hampered by the fact that annotation schemes vary enormously across languages, which has made it virtually impossible to perform sound comparative evaluations and cross-lingual learning experiments. The Universal Dependencies (UD) project (www.universaldependencies.org) seeks to tackle this problem by developing cross-linguistically consistent treebank annotation for many languages, aiming to capture similarities as well as idiosyncracies among typologically different languages (e.g., morphologically rich languages, prodrop languages, and languages featuring clitic doubling). The goal is not only to support comparative evaluation and cross-lingual learning but also to facilitate multilingual natural language processing and enable comparative linguistic studies. To serve all these purposes, the framework needs to have a solid linguistic foundation and at the same time be transparent and accessible to nonspecialists.

The syntactic annotation in UD is based on dependency, which is widely used in contemporary NLP, both for treebank annotation and as a parsing representation. It is also based on lexicalism, the idea that words are the basic units of grammatical annotation. Words have morphological properties and enter into syntactic relations, which is what the UD annotation is primarily meant to capture. This is reflected in the annotation, which consists of a morphological and a syntactic layer. However, it is important to note that syntactic wordhood does not always coincide with whitespaceseparated orthographic units, and another important design consideration is that there should be a transparent relation between the original textual representation and the linguistically motivated word segmentation. This is referred to in UD as the recoverability principle.

The morphological annotation consists of three levels of information: a lemma, a part-of-speech tag as well as a set of features which encode lexical and grammatical properties associated with the word form. Part-of-speech tags are drawn from a fixed inventory of 17 tags, which are assumed to be universal although not all categories have to be used in all languages. Features encode traditional grammatical categories like case, number, gender, tense, etc. in a standardized way, but it is possible to add new features and values if needed.

The syntactic annotation consists in labeled dependencies between words, using an inventory of 37 grammatical relations. The organization of this taxonomy distinguishes between three types of structure: nominals, clauses and modifier words. The scheme also makes a distinction between core arguments (e.g., subject and object) and other dependents, but does not attempt to distinguish complements vs. adjuncts. Each word depends either on another word in the sentence or on a notional root of the sentence, following three principles: content words are related by dependency relations; function words attach to the content word they further specify; and punctuation attaches to the head of the phrase or clause in which it appears. Giving priority to dependency relations between content words increases the probability of finding parallel structures across languages, since function words in one language often correspond to morphological inflection (or nothing at all) in other languages.

The UD initiative was launched in the fall of 2014 as an open community effort and has met with great interest and seen a very rapid growth. The latest official release in December 2016 contained 64 treebanks for 47 languages, and there is ongoing work on over 50 languages. In addition to contemporary standard languages, it has been applied to a number of classical languages, to Swedish sign language, and to English as a second language. Treebanks vary in size, from a few thousand to millions of words. All treebanks have the mandatory part-of-speech tags and dependencies, and most of them also have morphological features and lemmas. The UD resources have quickly been established as the de facto standard for crosslingual parsing experiments, but they have also been used in other types of NLP research as well as more linguistically oriented studies, for example, of word order typology. In the spring of 2017, there will be a CoNLL shared task on end-to-end UD parsing, which will coincide with the first release under v2 of the UD guidelines.

The first half of the tutorial will introduce the UD framework and resources, starting with the motivation and basic design principles, continuing with an overview of the annotation scheme, focusing on consistent annotation of linguistic constructions across typologically different languages, and ending with a survey of existing treebanks and their documentation. The second half of the tutorial will be more user-oriented and will cover tools for developing, maintaining and exploiting UD treebanks, procedures for adding new languages to UD as well as contributing new annotations for existing languages, ending with a survey of representative applications of UD in NLP and linguistics.

2 Outline

- 1. Introduction to UD (15 min)
 - Motivation
 - History
 - Basic principles
- 2. Treebank annotation (30 min)
 - Basic clauses
 - Nominal phrases

- Complex clauses
- Multiword expressions
- Ellipsis
- 3. Morphology (15 min)
 - Word segmentation
 - Part-of-speech tags
 - Morphological features
- 4. Syntax (15 min)
 - Clauses, nominals and modifiers
 - Core arguments vs. other dependents
 - Universal and language-specific relations
- 5. Resources and infrastructure (15 min)
 - CoNLL-U format
 - Treebanks
 - Documentation

BREAK

- 6. Tools for UD (20 min)
 - Annotation, conversion and validation
 - Treebank query tools
 - Morphological and syntactic processing
- 7. Adding new languages to UD (20 min)
 - The basic todo-list
 - Case study: Turkic languages
- 8. Extending UD annotations (20 min)
 - Enhanced dependencies
 - Standoff annotation
 - Parallel corpora
- 9. Using UD resources (20 min)
 - NLP applications
 - Linguistic research
- 10. Conclusion and outlook (10 min)
 - Where do we go from here?

3 About the Presenters

Joakim Nivre is professor of computational linguistics at Uppsala University, Sweden. His research focuses on morphological and syntactic processing of natural language, often from a multilingual perspective. He is one of the main developers of the widely used MaltParser system and a co-founder and current director of the Universal Dependencies project. He has previously given tutorials on dependency parsing at ACL-COLING in Sydney in 2006 (with Sandra Kbler) and at EACL in Gothenburg in 2014 (with Ryan McDonald). He is currently the vice-president of the Association for Computational Linguistics.

Daniel Zeman is a senior researcher and lecturer in mathematical linguistics at Charles University, Prague. His research interests include syntactic parsing, computational morphology and machine translation. He is the main author of the morphological part of the UD guidelines, and he has actively participated in converting more than 15 treebanks to the UD standard.

Filip Ginter is an assistant professor at the University of Turku, Finland. His research interests include syntactic parsing, treebank development, large-scale corpora, and event extraction from scientific literature. He has contributed to the tools and the technical infrastructure of the UD project, as well as the development of the UD guidelines.

Francis Tyers is a postdoctoral fellow in Sámi and Russian language technology at UiT in Troms, Norway. His research area is language technology for lesser-resourced languages, and especially languages with rich morphology such as the Turkic and Uralic languages. He has been involved in the development of rule-based morphological analysers, machine translation systems and treebanks for a number of languages, including Kazakh, Tatar, Tuvan, Kyrgyz and Turkish. He is secretary and co-founder of the ACL SIG on Uralic languages (SIGUR).