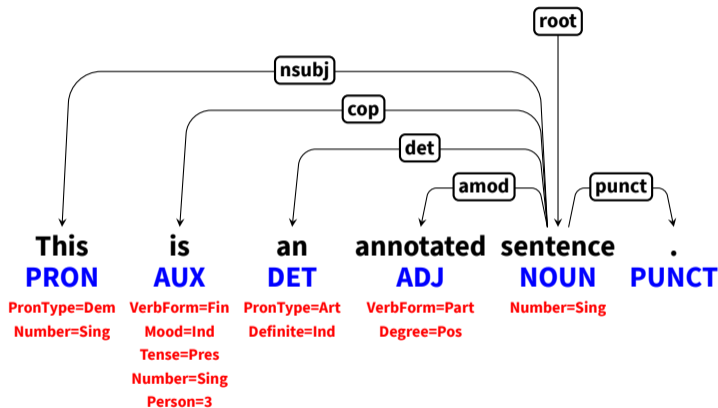


Universal Dependencies



We annotate morphology and syntax
of 90 languages
in a uniform way.

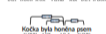
Universal Dependencies

Daniel Zeman
zeman@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia



Introduction

Universal Dependencies is an open community effort to create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework. The annotation consists in a linguistically motivated word segmentation; a morphological layer comprising lemmas, universal part-of-speech tags, and standardized morphological features; and a syntactic layer focusing on syntactic relations between predicates, arguments and modifiers.



Coverage



Language	Number of Sentences	Number of Tokens
Arabic	100,000	1,000,000
Chinese	100,000	1,000,000
English	100,000	1,000,000
French	100,000	1,000,000
German	100,000	1,000,000
Hebrew	100,000	1,000,000
Hindi	100,000	1,000,000
Japanese	100,000	1,000,000
Korean	100,000	1,000,000
Portuguese	100,000	1,000,000
Russian	100,000	1,000,000
Spanish	100,000	1,000,000
Tamil	100,000	1,000,000
Thai	100,000	1,000,000
Urdu	100,000	1,000,000
Vietnamese	100,000	1,000,000
Yiddish	100,000	1,000,000

Language	Number of Sentences	Number of Tokens
Arabic	100,000	1,000,000
Chinese	100,000	1,000,000
English	100,000	1,000,000
French	100,000	1,000,000
German	100,000	1,000,000
Hebrew	100,000	1,000,000
Hindi	100,000	1,000,000
Japanese	100,000	1,000,000
Korean	100,000	1,000,000
Portuguese	100,000	1,000,000
Russian	100,000	1,000,000
Spanish	100,000	1,000,000
Tamil	100,000	1,000,000
Thai	100,000	1,000,000
Urdu	100,000	1,000,000
Vietnamese	100,000	1,000,000
Yiddish	100,000	1,000,000

Language	Number of Sentences	Number of Tokens
Arabic	100,000	1,000,000
Chinese	100,000	1,000,000
English	100,000	1,000,000
French	100,000	1,000,000
German	100,000	1,000,000
Hebrew	100,000	1,000,000
Hindi	100,000	1,000,000
Japanese	100,000	1,000,000
Korean	100,000	1,000,000
Portuguese	100,000	1,000,000
Russian	100,000	1,000,000
Spanish	100,000	1,000,000
Tamil	100,000	1,000,000
Thai	100,000	1,000,000
Urdu	100,000	1,000,000
Vietnamese	100,000	1,000,000
Yiddish	100,000	1,000,000



Project site last updated by: Daniel Zeman, Faculty of Mathematics and Physics, Charles University, Institute of Formal and Applied Linguistics, Prague, Czechia