**Panel 18**

**Základní argumenty ve světových jazycích optikou Universal Dependencies**

Daniel Zeman (Univerzita Karlova, Matematicko-fyzikální fakulta)

Universal Dependencies (UD) (de Marneffe et al., 2021)[[1]](#footnote-1) představují morfosyntakticky orientovaný přístup k lingvistické anotaci, založené na binárních závislostních relacích mezi větnými členy. Závislosti se anotují primárně mezi plnovýznamovými slovy, zatímco funkční slova se považují za nositele morfosyntaktických rysů, které typicky „patří“ k některému plnovýznamovému slovu. Tento přístup v UD zvyšuje paralelismus mezi syntaktickými strukturami v různých jazycích. Cílem projektu UD byla v jeho počátcích harmonizace existujících korpusů a usnadnění vícejazyčného počítačového zpracování jazyků. Postupně se ale UD proměnily v soustavně se zvětšující sbírku korpusů, obohacovanou o data z dříve nepokrytých jazykových rodin, a staly se tak neocenitelným zdrojem i pro srovnávací jazykovědu a typologii (Futrell et al., 2015; Alzetta et al., 2018; Levshina, 2019; atd.). Tato výhoda ještě více vynikne, když se jazykověda spojí s jazykovými technologiemi a využije parser natrénovaný na datech UD k analýze dalších dat.

Zatím poslední verze, UD 2.14, se může pochlubit pokrytím 161 jazyků z 31 rodin (popř. 154 jazyků a 28 rodin, jestliže pomineme zvláštní „rodiny“ pro kreolštiny, znakové jazyky a střídání kódů). Tato sbírka je nicméně stále poměrně nevyvážená: Největší rodina, indoevropská, je reprezentována 74 jazyky, zatímco 18 rodin má jen po 1 jazyku. Ještě méně vyvážené jsou UD co do velikosti dat: indoevropské jazyky mají 24 mil. slov, uralské 1 mil., přes 100 tis. slov najdeme ještě v dalších 7 rodinách (afroasijská, turkická, sinotibetská, korejská, japonská, austronéská, baskická), zatímco na druhém konci škály je 9 rodin, které zatím nepřekročily laťku 10 tis. slov. Toto omezení je třeba mít na paměti a UD není možné považovat za proporční odraz trendů ve světových jazycích. I tak jde ale o zajímavý zdroj dat pro srovnání jazyků, které už pokryty jsou.

Jednou z klíčových vlastností anotačního schématu UD je, že rozlišuje tzv. základní argumenty (*core arguments*, S, A, P) od nepřímých rozvití (*oblique dependents*). Tím se UD liší od jiných závislostních korpusů, ve kterých se rozlišují argumenty od adjunktů, popř. přímo sémantické role. Vychází se z předpokladu, že většina jazyků používá specifické strategie pro kódování a specifická gramatická pravidla cílící na základní argumenty (Andrews, 2007; Zeman, 2017), čímž je opodstatněno jejich zachycení pomocí vyhrazených typů závislostí. Strategie pro kódování základních argumentů pracují s prostředky z poměrně malé množiny (slovosled, pádové morfémy, morfologie slovesa), nejsou však ve všech jazycích stejné. Přesná kritéria pro rozlišení základních argumentů od nepřímých rozvití musí být stanovena pro každý jazyk zvlášť.

Díky normalizované sadě značek používané v UD můžeme nyní studovat strategie kódování základních argumentů v jednotlivých jazycích. V případě konkurujících si strategií v rámci jednoho jazyka pak můžeme kvantifikovat jejich využití. UD například obsahuje 10 jazyků s ergativními morfologickými rysy: baskičtinu, gruzínštinu, abazinštinu, hindštinu, urdštinu, gudžarátštinu, marátštinu, bhódžpurštinu, warlpiri a jupičtinu. Mnoho dalších jazyků v UD vykazuje nominativně-akuzativní morfologii a nemálo jazyků pády nemá. Kolem 100 jazyků má anotace slovesné morfologie, která odráží rysy (např. osobu, číslo nebo rod) podmětu; 17 z nich promítá do téhož slovesa rysy dvou nebo více základních argumentů (podmětu a předmětu/ů).

Data z UD jsou také potvrzením, že slovosled hraje při kódování základních argumentů v některých jazycích větší roli než v jiných. Namísto striktní klasifikace jazyků jako SVO, SOV atd. můžeme pozorovat trendy a vyhodnotit, jak silně se ten který jazyk přiklání ke konkrétnímu slovoslednému typu. Obr. 1 je příkladem takového pozorování a ukazuje, že některé jazykové rodiny silně tíhnou ke konkrétnímu typu, zatímco jiné jsou rozprostřenější nebo zcela nevyhraněné, pokud jde o pořadí argumentů.



Obrázek 1: Podíl pořadí SV a VS (osa *y*) vs. pořadí OV a VO (osa *x*). S představuje jak jmenné podměty (nsubj), tak vedlejší věty podmětové (csubj). O představuje přímé předměty (obj) i vedlejší věty předmětové (ccomp). V je pozice hlavního slovesného uzlu (nikoli AUX). Klauze bez podmětu se počítají jako střed mezi SV a VS; podobně klauze bez předmětu táhnou jazyk k prostředku na ose OV-VO. Jazyky jsou reprezentovány svými kódy podle ISO 639. Vybrané rodiny jsou barevně zvýrazněné: indoevropská modře, uralská modrozeleně, afroasijská oranžově, turkická červeně, tupíjská fialově, sinotibetská purpurově, austronéská tyrkysově a drávidská olivově.

**Bibliografie**

Chiara Alzetta, Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal Dependencies and quantitative typological trends. Proc. of LREC 2018.

Avery D. Andrews. 2007. The major functions of the noun phrase. In Timothy Shopen (ed.): *Language Typology and Syntactic Description*, pages 132–223, Cambridge University Press.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. Proc. of Natl. Academy of Sciences, 112(33):10336–10341

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Daniel Zeman. 2017. Core arguments in Universal Dependencies. Proc. of DepLing 2017, p. 287–296.

1. <https://universaldependencies.org/> [↑](#footnote-ref-1)