NEXUSLINGUARUM – EUROPEAN NETWORK FOR WEB-CENTRED LINGUISTIC DATA SCIENCE

(Európska sieť pre web-centrickú lingvistickú dátovú vedu[1])

Jorge Gracia
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza

Radovan Garabík
Jazykovedný ústav Ľudovíta Štúra SAV Bratislava

Vladimír Benko
Jazykovedný ústav Ľudovíta Štúra SAV Bratislava

Jazykovedný ústav Ľ. Štúra SAV (JÚĽŠ SAV) sa podieľa na COST akcii[2] CA18209 *European network for Web-centred linguistic data science*[3], ktorej hlavným cieľom je podpora synergií v celej Európe medzi jazykovedou, informatikou, umelou inteligenciou a ostatnými prírodovednými a humanitnými odbormi s cieľom podporiť výskum a rozšíriť oblasť lingvistickej dátovej vedy. Plánované trvanie projektu NexusLinguarum je štyri roky (október 2019 – október 2023).

Lingvistickú dátovú vedu chápeme ako súčasť aktuálne sa rýchlo rozvíjajúceho odboru *data science* (dátovej vedy – slovenský termín nie je zatiaľ veľmi zaužívaný), ktorý sa zameriava na systematickú analýzu a štúdium štruktúry a vlastností lingvistických dát veľkého rozsahu spolu s metódami a technikami ich spracovania a získavania z nich nových poznatkov. Špecificky, lingvistická dátová veda poskytuje formálny základ pre analýzu, reprezentáciu, integráciu a využívanie lingvistických dát používaných pri automatickom spracovaní a analýze jazyka (na úrovni syntaxe, morfológie, terminológie atď.) a aplikovaného počítačového spracovania prirodzeného jazyka (napr. pri strojovom preklade, rozpoznávaní reči, analýze sentimentu).

Aby sa podporilo štúdium lingvistickej vedy o údajoch najúčinnejším a najproduktívnejším spôsobom, akcia podporuje vybudovanie viacjazyčného ekosystému interoperabilných jazykových údajov, na čo sa budú využívať metódy známe zo sémantického webu, počítačového spracovania prirodzeného jazyka (NLP), prelinkovaných otvorených lingvistických dát (LLOD[4]) prepojených na viacjazyčné zdroje (dvojjazyčné slovníky, viacjazyčné korpusy, terminologické databázy atď.). Takýto

[1] Oficiálny názov projektu je iba anglický – tu uvádzame náš neoficiálny preklad.
[2] Nástroj pre európsku spoluprácu vo vede a technológiách; https://cost.eu.
[3] Stránka akcie: https://nexuslinguarum.eu/.
[4] Linguistic Linked Open Data.

digitálny a používateľský ekosystém by mohol uľahčiť prekonávanie jazykových bariér v Európe (a prípadne aj inde) a podporiť elektronické obchodovanie a kultúrnu výmenu medzi krajinami s rôznymi jazykmi a napomôcť aj technologickej podpore menšinových jazykov, v súčasnosti značne obmedzenej.

Medzi hlavné ciele projektu NexusLinguarum patria:

- navrhovanie a schvaľovanie postupov a štandardov pre prepájanie údajov a služieb týkajúcich sa viacerých jazykov;
- organizovanie aktivít na podporu spolupráce a komunikácie medzi vedeckými komunitami, ako sú workshopy, semináre a konferencie;
- zhromažďovanie a analyzovanie prípadov možností použitia lingvistickej dátovej vedy a vývoj prototypov nástrojov pre niektoré ukážkové prípady.

Ďalej je v projekte plánované vypracovanie vzorového študijného programu pre celoeurópske vysokoškolské štúdium, ktorý by mohol pomôcť pri vzniku novej generácie výskumných pracovníkov v tejto oblasti a priniesol tak lingvistickú dátovú vedu do interdisciplinárneho akademického prostredia.

V súčasnosti je do projektu zapojených 42 krajín (37 krajín akcie COST, tri susedné krajiny a dve z medzinárodných partnerských krajín). Doteraz sa do pracovných skupín zapojilo 191 členov a tento počet neustále rastie, keďže sieť je otvorená pre nových účastníkov. Účastníci tvoria široko zameranú skupinu odborníkov z rôznych vedných oblastí – z informatiky, sémantického webu, umelej inteligencie, lingvistiky, humanitných vied atď.

V rámci projektu NexusLinguarum existuje päť pracovných skupín (WG); štyri technické a jedna riadiaca:

WG1 – Prepojené jazykové zdroje založené na dátach. Táto pracovná skupina si kladie za cieľ tvorbu odporúčaných postupov pre vývoj, vytváranie, zlepšovanie, diagnostiku, opravu a obohatenie zdrojov LLOD.

WG2 – Prelinkované NLP založené na dátach. Táto pracovná skupina sa zameriava na uplatňovanie metód lingvistickej dátovej vedy vrátane LLOD, cielené na obohatenie spracovania prirodzeného jazyka so zámerom využiť rastúce množstvo lingvistických (otvorených) dát dostupných na webe.

WG3 – Podpora lingvistickej dátovej vedy. Cieľom tejto pracovnej skupiny je podporiť štúdium lingvistických dát využívaním dátovo-analytických metód v kombinácii s LLOD a prelinkovaným počítačovým spracovaním jazyka založeným na dátach.

WG4 – Prípady použitia a aplikácie. Táto pracovná skupina sa zameriava na skúmanie prípadov použitia a praktických aplikácií príslušných technológií.

WG5 – Manažment a diseminácia. Táto pracovná skupina sa stará o riadenie celej akcie, o jej zviditeľňovanie a monitoruje aktivity prepájajúce rôzne pracovné skupiny.

JÚĽŠ SAV sa zapája hlavne do WG3 a WG4, kde ako jedna z hlavných vedeckovýskumných lingvistických inštitúcií na Slovensku a vedúca inštitúcia v NLP takto zužitkuje dlhoročné skúsenosti a vedecké poznatky v oblasti klasickej lingvistiky, NLP a počítačovej lexikografie. Keďže projekt kladie dôraz na využitie LLOD v lingvistike a dátovej vede, aj prínos JÚĽŠ SAV sa prispôsobuje tomuto zameraniu, s dôrazom na tvorbu a aplikácie LLOD v slovenskom jazykovednom prostredí a na využitie moderných metód neurónových sietí a umelej inteligencie aplikovaných na počítačové spracovanie prirodzeného jazyka a na použitie LLOD v slovenskej lexikografii. Aktuálne sa zdá, že lingvistika stojí na prahu ďalšej technologickej revolúcie (po pomerne nedávnej revolúcii spôsobenej dostupnosťou veľkých textových korpusov), tentoraz vyvolanej rozmachom a predpokladanou dostupnosťou – v budúcnosti pravdepodobne aj pre slovenčinu – veľkých jazykových modelov založených na deep learning metódach. Z tohto dôvodu je dôležité, aby slovenská jazykoveda aspoň nestratila z dohľadu moderné lingvistické metódy, postupy a trendy.

Text vychádza z publikácie: Jorge Gracia: *NexusLinguarum: European network for Web-centred linguistic data science*. In: K Lexical News, č. 28, 2020, s. 21 – 29.