SLOVKO 2023.

POČÍTAČOVÉ SPRACOVANIE PRIRODZENÉHO JAZYKA A KORPUSOVÁ LINGVISTIKA[1]

Beáta Kmeťová – Adriána Žáková

Jazykovedný ústav Ľudovíta Štúra SAV, v. v. i.
Panská 26, Bratislava
E-mail: beata.kmetova@korpus.juls.savba.sk, adriana.zakova@korpus.juls.savba.sk

V dňoch 18. – 20. októbra 2023 sa v priestoroch hotela Devín v Bratislave uskutočnila dvanásta medzinárodná konferencia SLOVKO 2023. Podujatie zamerané na *spracovanie prirodzeného jazyka a korpusovú lingvistiku* organizovalo oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra SAV, v. v. i. Konferenciu slávnostne otvorila vedúca oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra SAV, v. v. i., Jana Levická, riaditeľka Jazykovedného ústavu Ľudovíta Štúra SAV, v. v. i., Gabriela Múcsková a podpredseda Slovenskej akadémie vied, v. v. i., Miroslav Morovics.

Na konferencii odzneli dve plenárne prednášky a 35 prezentácií v anglickom, českom a slovenskom jazyku. Na podujatí vystúpili odborníci z Česka, Luxemburska, Nemecka, Poľska, Slovenska a Švédska. Tlačené príspevky z konferencie boli vydané v prvom tohtoročnom čísle *Jazykovedného časopisu*,[2] ktorý mali účastníci k dispozícii už na podujatí. Tematické číslo je dostupné aj v elektronickej podobe v archíve Jazykovedného časopisu.[1] Texty príspevkov sú publikované v anglickom jazyku. V správe informujeme o príspevkoch z konferencie vrátane tých, ktoré nie sú uvedené v Jazykovednom časopise (autori Daniel Kluvanec, Adrian Jan Zasina, Václav Cvrček a Masako Fidlerová). Okrem toho uvádzame aj príspevky, ktoré na podujatí neboli odprezentované, ale sú publikované v uvedenom čísle Jazykovedného časopisu (príspevok M. Zumríka). Mená autorov z krajín, kde sa používa iná grafická sústava ako latinská, uvádzame v latinke.

Príspevky sú v správe usporiadané do štyroch tematických okruhov rovnako, ako sú usporiadané v Jazykovednom časopise. Do prvého tematického celku sú zaradené štúdie, ktoré sa zameriavajú na korpusovo podporovaný a riadený výskum

[1] Správa vznikla v rámci projektu *Tvorba a rozvoj Slovenského národného korpusu (V. etapa),* ktorý finančne podporuje Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky, Ministerstvo kultúry Slovenskej republiky, Slovenská akadémia vied, v. v. i., a Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, v. v. i.

[2] *SLOVKO 2023. Natural Language Processing and Corpus Linguistics*. 2023, roč. 74, č. 1, 401 s. Prizvané editorky: Katarína Gajdošová, Adriána Žáková.

(corpus-based and corpus-driven research). Druhý tematický okruh tvoria príspevky venované jazykovej akvizícii, tvorbe a využívaniu jazykových zdrojov. V treťom celku sa autori venovali témam z oblasti budovania a tvorby korpusov. Samostatnú oblasť tvoria príspevky štvrtého okruhu, ktorý obsahuje štúdie z oblasti počítačového spracovania prirodzeného jazyka a digitálnych humanitných a spoločenských vied.

Prvý tematický celok venovaný korpusovo podporovanému a riadenému výskumu otvorila **Renata Bronikowska** (Ústav poľského jazyka Poľskej akadémie vied vo Varšave), ktorá predstavila vo svojej prednáške *Verbification of Feminine Forms of Adjectives można 'possible', niemożna 'impossible' and niepodobna 'impossible' – Corpus-based Approach* výskum predikatívnych lexém v poľštine, ktoré čerpala z dát elektronického korpusu textov zo 17. a 18. storočia. Výskum ukázal, že dve zo skúmaných adjektív *można* a *niemożna* svoj proces verbifikácie ukončili už v 17. storočí, ale transformácia adjektíva *niepodobna* stále prebieha.

Autori **Jaroslav David**, **Michal Místecký** a **Tereza Klemensová** (Filozofická fakulta Ostravskej univerzity v Ostrave) opísali v príspevku *Appellativization of Proper Names – In the Perspective of Corpus Analysis* výskum procesu tvorby apelatív z vlastných mien v češtine. Na konferencii sa Michal Místecký venoval apelativizovaným lexémam vznikajúcim z mien známych osobností ako *Masaryk, Beneš, Hitler, Kafla, Stalin* a iní. Analýzou dát českého korpusu SYNv11 autori zistili, že najčastejšími novovzniknutými lexémami sú adjektíva vytvorené súčasne skladaním a sufixáciou alebo súčasne sufixáciou a prefixáciou.

V prezentácii s názvom *The Economy of Czech Exchange in the Slovak Marketplace of Austria after the Fall of Hungary* informoval **Martin Diweg-Pukanec** (Filozofická fakulta Univerzity Konštantína Filozofa v Nitre) o výskume dĺžky slov vo výpovediach zachytených v historických textoch zo 16. storočia, ktoré sú k dispozícii v Kremnickom archíve. V štúdii uviedol, že priemerná dĺžka slov hovoriacich pochádzajúcich z rôznych sociálnych vrstiev sa líšila; čím bol hovoriaci menej vzdelaný, tým vyššiu tendenciu jazykovej ekonómie jeho výpovede vykazujú.

Prehľad vývoja a zmien v oblasti spracovania jazyka predovšetkým po príchode softvérových nástrojov koncom 80. rokov 20. storočia ponúkol v plenárnej prednáške s názvom *From Solitary Corpus Analysis to Collective Insight: A Glimpse into Translation and Lexicography*[3] **Łukasz Grabowski** (Filologická fakulta Opolskej univerzity v Opole). Autor poukázal nielen na rozmach NLP nástrojov, ale aj na rozmanitosť štatistických metód, techník na vizualizáciu dát a počítačových modelov na spracovanie jazyka. Upozornil na to, že korpusový lingvista sa vzhľadom na neustále sa zväčšujúci objem dát a inovatívnych techník stáva nielen počítačovým

[3] Názov príspevku v Jazykovednom časopise je iný: *Statistician, Programmer, Data Scientist? Who Is, or Should Be, a Corpus Linguist in the 2020s?*

programátorom, ale aj štatistikom a dátovým vedcom. Spomenul aj možné „vzdialenie“ sa korpusovej lingvistiky od asistovaného výskumu umelej inteligencie (AI-assisted research), no zároveň vyjadril svoju vieru v ich synergiu v budúcnosti.

Jakob Horsch (Fakulta jazykov a literatúry Katolíckej univerzity v Eichstätte – Ingolstadte) predstavil vo svojej prípadovej štúdii *Corroborating Corpus Data with Elicited Introspection Data: A Case Study* výskum anglickej, prevažne elidovanej hypotaktickej spojky *that* (*že*, resp. *ktorý*). Korpusové dáta autor pre extrémne nízku frekvenciu javu doplnil elicitovanými introspektívnymi dátami z tzv. testu odhadu veľkosti (Magnitude Estimation Test; MET). V analýze ukázal, že spojka *that* plní svoju funkciu uvádzania vedľajšej vety, preto nemôže byť z vety vždy vynechaná.

V korpusovej štúdii *Dative Ambiguity in Russian: A Corpus Induced Study* predstavila **Edyta Jurkiewicz-Rohrbacher** (Inštitút slavistických štúdií Regensburskej univerzity v Regensburgu aj Inštitút slavistických štúdií Hamburskej univerzity v Hamburgu) problematické značkovanie datívnych predikatívnych syntagiem v ruskom jazyku spôsobené neprítomnosťou zhody medzi podmetom a prísudkom, čo pri syntaktickej analýze spôsobuje nesprávnu identifikáciu podmetu. Ako autorka uviedla, správne priradenie syntaktickej funkcie často vyplýva práve z kontextu.

Na používanie a frekvenciu adjektív zakončených na *-al* a *-ell*, príp. *-ial* a *-iel* sa v prezentácii *The Competition of German Adjectival Suffixes* zameral **Filip Kalaš** (Fakulta aplikovaných jazykov Ekonomickej univerzity v Bratislave). V príspevku ponúkol prehľad najfrekventovanejších kolokátov týchto adjektív získaných z korpusu Araneum Germanicum III Maximum. Z výsledkov jeho štúdie vyplýva, že kolokáty adjektív zakončených na *-ell/-iel* sa vďaka svojej vyššej lexikálnej spájateľnosti vyskytujú častejšie v špecializovaných textoch, napr. v textoch z oblasti astronómie, matematiky a fyziky.

Marie Kopřivová (Filozofická fakulta Karlovej univerzity v Prahe) predstavila štúdiu českých paremiologických jednotiek *Proverbs in Contemporary Czech. Corpus Probe into Written Texts* (v spoluautorstve s **Kateřinou Šichovou** z Centra pre české štúdie – Bohemicum na Regensburskej univerzite v Regensburgu). Autorky informovali o novovzniknutom paremiologickom minime PM2023, zozname najfrekventovanejších prísloví získaných zo štyroch reprezentatívnych písaných korpusov češtiny. Získaný zoznam prísloví autorky porovnali s existujúcimi prácami o parémiách a prienikom vytvorili zoznam najčastejších prísloví. V ďalšom výskume sa plánujú zamerať na výber prísloví vhodných pri tvorbe učebníc a na porovnanie s paremiologickým minimom nemčiny.

Extrakcii a analýze kľúčových slov z podkorpusu náboženských textov pochádzajúcich zo 17. a 18. storočia sa v príspevku *Keywords in Religious Literature of 17th and 18th Centuries in Light of the Data from the Electronic Corpus of 17th- and*

18th-century Polish Texts venovala **Madalena Majdak** (Inštitút poľského jazyka Poľskej akadémie vied vo Varšave). Autorka porovnávala náboženskú lexiku získanú z historického korpusu náboženských textov s lexikou referenčného korpusu poľštiny. V prezentácii konštatovala, že použitá logaritmická metóda vierohodnosti (z angl. log-likelihood) bola na získanie kľúčových slov zvolená vhodne a identifikované dáta budú použité v ďalšom lexikálnom výskume.

Marie Mikulová (Matematicko-fyzikálna fakulta Karlovej univerzity v Prahe) predniesla na konferencii príspevok *Expressing Measure in Czech (A Corpus-based Study),* v ktorom sa zamerala predovšetkým na teoreticky podložený a korpusovo overený opis výrazov s významom miery v češtine, napr. *akorát*, *víceméně*, *nadmíru*, *částečně*. Analýza rozsiahleho korpusového materiálu odhalila rôznorodosť štruktúr a relatívnu náročnosť ich klasifikácie. Prednášajúca uviedla, že výsledný popis je dobre využiteľný v sémanticko-syntaktickej anotácii Pražského závislostného korpusu, ako aj v sémanticky orientovanom opise jazykov.

V príspevku *Adverbs Derived from Adjectival Present Participles in Polish, Slovak and Czech: A Comparative Corpus-based Study* analyzovala **Aksana Schillová** (Ústav pre jazyk český Akadémie vied Českej republiky v Prahe), aký slovotvorný typ prísloviek v češtine je ekvivalentný typu prísloviek utvorených od adjektivizovaných príčastí prítomných v poľštine a slovenčine. Porovnávací výskum uskutočnila na korpusoch Araneum Polonicum Minus, Araneum Slovacum VI Minus Beta a Araneum Bohemicum IV Minus.

Barbora Štěpánková (Matematicko-fyzikálna fakulta Karlovej univerzity v Prahe) sa v prezentácii príspevku *The Epistemic Marker určitě in the Light of Corpus Data*, ktorý vznikol v spoluautorstve s **Janou Šindlerovou** a **Luciou Polákovou** (Filozofická fakulta Karlovej univerzity v Prahe), zamerala na prostriedky vyjadrovania epistemickej modality a evidenciality v češtine. Cieľom ich pilotnej štúdie bola analýza častice *určitě*, typicky považovanej za signál vysokej istoty. Na základe svojich zistení v paralelnom korpuse InterCorp v15 navrhujú alternatívnu metódu na rozlišovanie medzi jej významovými odtieňmi v rôznych výpovediach.

Kvantitatívnou charakteristikou najčastejších substantívnych lem v jazyku slovenských súdnych rozhodnutí v trestných záležitostiach sa v príspevku *Comparative Lexical Analysis of Nouns Lemmas in Slovak Judicial Decisions* zaoberal **Miroslav Zumrík** (Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, v. v. i., v Bratislave). Texty rozhodnutí porovnával so vzorkami slovenských zákonov, odborných textov z oblasti marketingu, neprekladovej beletrie a s vyváženým korpusom prim-10.0--public-vyv. Výskumnou otázkou naďalej ostáva, či sa substantívny slovník daného žánru právnych textov naozaj vyznačuje lexikálnou chudobnosťou, aká sa administratívnemu a právnemu štýlu tradične prisudzuje.

Druhý tematický celok venovaný jazykovej akvizícii, tvorbe a využívaniu jazykových zdrojov otvárali **Cristina Fernández-Alcaina, Eva Fučíková, Jan Hajič** a **Zdeňka Urešová** (Matematicko-fyzikálna fakulta Karlovej univerzity v Prahe) prednáškou *Spanish Synonyms as Part of a Multilingual Event-Type Ontology*. Autori informovali o aktuálnom rozširovaní ontologického slovníka SynSemClass, ktorý obsahuje ekvivalentné skupiny slovies v paralelných textoch a integruje ich do valenčných lexikónov a iných externých zdrojov. Najnovšia verzia valenčného slovníka tak bude obohatená o španielske texty s 1 400 slovesných synonymami.

Michaela Mošaťová a **Petra Švancarová** (Filozofická fakulta Univerzity Komenského v Bratislave) ponúkli v príspevku *Errors in the Congruent Attribute Among Students Learning Slovak as a Foreign Language (Learner Corpus-based)*, v spoluautorstve s **Katarínou Gajdošovou** z Jazykovedného ústavu Ľudovíta Štúra SAV, v. v. i., v Bratislave, prehľad chýb študentov učiacich sa slovenčinu ako cudzí jazyk. Kvalitatívne analyzovaný materiál pochádza z pilotnej verzie korpusu err-korp-pilot. Autorky skúmali najčastejšie typy chýb v kongruentnom atribúte, ako sú nesprávna zhoda s gramatickým rodom a pádom nadradeného podstatného mena, a ponúkli interpretáciu príčin vzniku týchto chýb. Z výskumu vyplýva, že chyby sú zvyčajne spôsobené najmä transferom z pôvodného jazyka, prílišným zovšeobecňovaním pravidla v cieľovom jazyku alebo neznalosťou gramatického pravidla.

Veronika Kolářová, Václava Kettnerová a **Jiří Mírovský (**Matematicko-fyzikálna fakulta Karlovej univerzity v Prahe) v prednáške *Through Derivational Relations to Valency of Non-verbal Predicates in the NOMVALLEX Lexicon* predstavili nové vlastnosti valenčného slovníka substantív a adjektív NomVallex a možnosti jeho využitia v korpusovom výskume valencie. V štúdii sa autori zamerali na valenciu deverbálnych adjektív a rozdielov ich derivačných typov.

Pracovníci Filozofickej fakulty Ostravskej univerzity v Ostrave **Michaela Nogolová, Michaela Hanušková**, **Miroslav Kubát** a **Radek Čech** sa v príspevku *Linear Dependency Segments in Foreign Language Acquisition: Syntactic Complexity Analysis in Czech Learners' Texts* zamerali na analýzu syntaxe pri osvojovaní si cudzieho jazyka v dátach akvizičného korpusu českého jazyka CzeSL-SGT. Autori venujú pozornosť syntaktickým jednotkám, ktoré nazvali segmenty lineárnej závislosti (Linear Dependency Segment; LDS), využiteľné pri korpusovom výskume komplexnosti syntaxe. Z výskumu vyplýva, že čím vyššia je jazyková úroveň učiaceho sa študenta, tým sú jednotky LDS kratšie.

V online prezentácii *Differences in Spoken Language Processing in General Corpora (ORAL, ORTOFON) and in a Specialized Corpus (DIALEKT) and Their Reflection in the Mapka Application* predstavila **Martina Waclawičová** (Filozofická fakulta Karlovej univerzity v Prahe) všeobecné korpusy hovoreného českého jazyka ORAL a ORTOFON, špecializovaný korpus DIALEKT a interaktívnu apliká-

ciu Mapka, ktorá slúži ako doplnkový materiál k týmto korpusom. Medzi hlavné funkcie aplikácie patrí detailné zobrazenie územného, predovšetkým nárečového členenia. Ponúka prehľad charakteristických jazykových javov v nárečovej oblasti, obsahuje taktiež funkciu vyhľadávania a možnosť vytvoriť si vlastnú mapu. Od roku 2023 je k dispozícii nová verzia aplikácie Mapka, obohatená o podrobné opisy nárečových javov a ukážky z uvedených korpusov.

Adrian Jan Zasina (Filozofická fakulta Karlovej univerzity v Prahe) v prezentácii *Can Low-proficiency Level Learners Produce Diverse Texts? A Multidimensional Approach to Czech as a Foreign Language* ponúkol analýzu funkčnej variability textov v korpuse písaných textov študentov češtiny pochádzajúcich z Poľska. Počas výskumu si všímal intra- a extratextuálne charakteristiky vybraných typov textov (neformálny list, opis, argumentácia a rozprávanie príbehu) študentov učiacich sa češtinu. V závere prezentácie uviedol, že aj študenti s nižšou znalosťou češtiny dokážu produkovať texty rôznych žánrov. Výsledky svojho výskumu plánuje porovnávať s výskumom variability textov u študentov s inými materinskými jazykmi, napr. kórejčinou.

Pracovníci viacerých výskumných inštitúcií, **Daniel Zeman** (Matematicko-fyzikálna fakulta Karlovej univerzity v Prahe), **Pavel Kosek** a **Martin Březina** (Filozofická fakulta Masarykovej univerzity v Brne) a **Jiří Pergler** (Ústav pre jazyk český Akadémie vied Českej republiky, v. v. i., v Prahe), predstavili v príspevku *Morphosyntactic Annotation in Universal Dependencies for Old Czech* prvé kroky pri príprave syntakticky anotovaného korpusu textov češtiny zo 14. storočia v koncepcii Universal Dependencies. Na pilotné testovanie, ktoré zahŕňa aj úpravu anotácie pre javy vyskytujúce sa v historickej, ale nie modernej češtine, vybrali drážďanskú a olomouckú verziu Evanjelia podľa Matúša. V príspevku zaznelo viacero zaujímavých postrehov týkajúcich sa použiteľnosti syntaktického analyzátora modernej češtiny na staročeské dáta.

Tretí tematický blok zahŕňal príspevky z oblasti tvorby a budovania korpusov. **Ilia Afanasev, Olga Lyashevskaya, Stefan Rebrikov, Yana Shishkina, Igor Trofimov** a **Natalia Vlasova** sa v príspevku *The Effect of (Historical) Language Variation on the East Slavic Lects Lemmatisers Performance* venovali porovnaniu modelov na lematizáciu textov v historických východoslovanských jazykoch, a to modely BERT-based a BART-large. BART-large model mal menšie problémy pri morfologickom značkovaní dát v historických textoch, Rubic (BERT-based) model dokázal ľahšie prekonať synchrónne regionálne jazykové variácie.

Hana Skoumalová a **Vladimír Petkevič** (Filozofická fakulta Karlovej univerzity v Prahe) ponúkli v príspevku *Annotation of Analytic Verbs Forms in Czech – Complex Cases* prehľad teoretických a praktických problémov pri automatizovanej morfologickej a morfosyntaktickej anotácii analytických slovies v písanom korpuse

súčasnej češtiny. Hana Skoumalová okrem iného uviedla, že anotácia zlyháva pri značkovaní spojenia slovesa *byť* s doplnkom (najmä adjektívom) a problematické sú tiež trpné príčastia.

V príspevku *CapekDraCor: A New Contribution to the European Programmable* **Petr Pořízka** (Filozofická fakulta Univerzity Palackého v Olomouci) predstavil projekt DraCor so svojou výskumne orientovanou koncepciou programovateľných korpusov vytváraných na kvantitatívne analýzy v rámci počítačovej literárnej vedy. DraCor je jedinečná otvorená infraštruktúra na analýzu európskej drámy, ktorá v súčasnosti obsahuje 15 korpusov z rôznych období v 10 rôznych jazykoch. Autor predstavil korpus CapekDraCor, pozostávajúci zo všetkých hier Karla a Josefa Čapka, a informoval o spôsobe spracovania dát s ohľadom na špecifiká jeho štruktúry.

Problematike rôznorodosti systémov a nástrojov na anotáciu paralelného korpusu InterCorp sa v štúdii *The InterCorp Parallel Corpus with a Uniform Annotation for All Languages* venoval **Alexander Rosen** (Filozofická fakulta Karlovej univerzity v Prahe). Autor predstavil integrovanú morfologickú a morfosyntaktickú anotačnú schému univerzálnych závislostí (Universal Dependencies), aplikovanú na InterCorp, ktorá okrem morfologickej umožňuje tiež syntaktickú analýzu paralelných textov.

Na problémy s lematizáciou a anotáciou textov v korpusoch historickej ruštiny sa vo svojej prednáške *Multiple Interpretation and Fragmented Texts Within a Historical Corpus: The Case of Old East Slavic Vernacular Writing* zameral **Dmitri Sitchinava** (Ústav slavistiky Postupimskej univerzity v Postupime). Autor ponúkol prehľad rozmanitých dokumentov historickej ruštiny – texty písané na brezovej kôre, dobové nápisy, popisky a pod. Keďže korpusy obsahujú často iba fragmenty pôvodných textov alebo nápisov, autor upozornil na nejednotnú alebo chýbajúcu anotáciu týchto textov a v prezentácii navrhol spôsoby možného značkovania predstavených dobových dokumentov.

Trojica autorov **Lucie Benešová, Martin Stluka** (Filozofická fakulta Karlovej univerzity v Prahe) a **Klára Pivoňková** (Filozofická a Prírodovedecká fakulta Karlovej univerzity v Prahe) sa v príspevku *Lemmatization of the DIA1900 Diachronic Corpus* podelili o skúsenosti s procesom lematizácie pripravovaného nového diachrónneho korpusu DIA1900, ktorý mapuje češtinu 2. polovice 19. storočia. Autori spoločne predstavili koncepciu lematizácie a prípravy korpusu, ako aj problémy, s ktorými sa pri jeho vytváraní stretávajú. Na záver uviedli etalón možností automatizovanej anotácie.

Štvrtý tematický okruh konferencie zahŕňal štúdie z oblasti počítačového spracovania prirodzeného jazyka a digitálnych humanitných a spoločenských vied. Patrí doň aj príspevok autorov **Martina Braxatorisa** (Ústav slovenskej literatúry Slovenskej akadémie vied, v. v. i., v Bratislave) a **Anity Braxatorisovej** (Filozofická fakul-

ta Univerzity sv. Cyrila a Metoda v Trnave), ktorí v príspevku *Use of Computer and Corpus Tools in the Research of a 19th Century German-language Manuscript Book of Notes and Extracts* predstavili možnosti využitia počítačových a korpusových nástrojov pri interpretácii textov špecifického žánru, a to nemeckojazyčnej knihy poznámok, výpiskov a záznamov Samuela Ferjenčíka (1793 – 1855). Výsledky ich výskumu odhalili viaceré manipulácie s východiskovými autorskými textami, najmä zámeny v textoch, ktorým autori štúdie venovali osobitnú pozornosť. Ďalší výskum plánujú zamerať na nástroje, ktoré majú potenciál výrazne uľahčiť výskum intertextových vzťahov aj pri iných dokumentoch.

Metodológiu získavania kolokácií a asociácií v rôznych diskurzoch predstavil **Václav Cvrček** (Filozofická fakulta Karlovej univerzity v Prahe). V štúdii *Associations and Collocations in Corpora,* ktorá vznikla v spolupráci s **Masako Fidlerovou** (Brownova univerzita v Providence v Spojených štátoch amerických), porovnávali autori metódu analýzy kolokácií (CA; Collocation Analysis) s metódou analýzy nákupného koša (MBA; Market Basket Analysis), ktorou sa získavajú asociácie vyskytujúce sa často spolu v istých typoch textov. Z porovnania vyplynulo, že metóda MBA odhalila viac termínov v textoch napríklad z oblasti geografie, migrácie a počasia. Asociácie sa častejšie vyskytujú v textoch zameraných na politiku a ekonomiku.

Jazykovým poruchám a ich analýze sa vo svojej prezentácii *Lexical Diversity and Language Impairment* venovala **Nataliia Časnochová Zozuk** (Fakulta prírodných vied a informatiky Univerzity Konštantína Filozofa v Nitre). Prezentovala výskum na projekte EWA (Early Warning of Alzheimer), ktorého cieľom bolo potvrdiť hypotézu, že jazykové prejavy ľudí s neurodegeneratívnymi ochoreniami majú menšiu lexikálnu rôznorodosť ako prejavy zdravých ľudí. Hlavným prínosom projektu bolo vyvinutie mobilnej aplikácie EWA, ktorá umožňuje detekciu ochorenia prostredníctvom analýzy rečových prejavov testovanej osoby.

V príspevku *Text Vectorization Techniques Based on Wordnet* predstavili **Dávid Držík** a **Kirsten Šteflovič** (Fakulta prírodných vied a informatiky Univerzity Konštantína Filozofa v Nitre) využitie synsetov, zoznamov synoným z anglickej lexikálnej databázy slov WordNet, pri použití Text Data Augmentation (TDA) – nahrádzaní slov vo vete synonymami. Na základe klasifikačnej úspešnosti autori hodnotili, či je lepší originálny alebo novovytvorený korpus pomocou synoným. Svojimi výsledkami demonštrovali, že rozšírenie korpusu touto metódou vedie k vylepšenej vektorovej reprezentácii slov. V budúcom výskume sa plánujú zamerať na ďalšie TDA metódy. Okrem náhrady synonymami chcú skúmať, či sa zlepší klasifikácia aj pri spätnom preklade, prípadne pri odstránení plnovýznamových slov.

V prezentácii *When Is a Crisis Really a Crisis? Using NLP and Corpus Linguistic Methods to Reveal Differences in Migration Discourse across Czech Media* sa **Irene Elmerot** (Fakulta humanitných vied Štokholmskej univerzity v Štokholme)

venovala analýze variability jazykových spojení v rôznych typoch médií s využitím nástrojov spracovania prirodzeného jazyka a korpusovej lingvistiky na získanie obrazu o diskurze migrácie v českých médiách v rokoch 2015 – 2023. Poukázala na značné rozdiely v používaní termínov a kolokácií slov (ne)legálny, (i)migrant a pod., v oblasti migrácie.

Maroš Harahus, Daniel Hládek, Ján Staš a **Matúš Pleva** (Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach) predstavili v príspevku *Slovak Language Models for Basic Preprocessing Tasks in Python* nové jazykové modely spracovania prirodzeného jazyka s použitím programovacieho jazyka Python a knižnice spaCy. Prezentovaný jazykový model dokáže text tokenizovať, parsovať, rozpoznať v ňom slovné druhy a pomenované entity.

Autori **Richard Holaj** (Filozofická fakulta Masarykovej univerzity v Brne) a **Petr Pořízka** (Filozofická fakulta Univerzity Palackého v Olomouci) prezentovali v príspevku *ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech* výskumný projekt, ktorý pristupuje inovatívne k e-learningovým aplikáciám. Autori zhromaždili údaje, uskutočnili pilotný projekt a vyvinuli prvú verziu systému atribútov založených na označovaní izolovaných zvukov reči. Vo svojej prezentácii predstavili aj anotačný nástroj Anophone, nový anotačný systém založený na celoslovnom značkovaní dvoj- až štvorslabičných slov. Pomocou nástroja docielili zlepšenie modelu na rozpoznávanie reči aj jeho vyššiu úspešnosť.

V plenárnej prednáške s názvom *eTranslation as a EU Flagship Use Case of Natural Language Processing* informoval **Daniel Kluvanec** (Generálne riaditeľstvo pre preklad Európskej komisie v Luxemburgu) o prekladovej platforme eTranslation, najväčšom projekte programu Európskej únie Digitálna Európa na podporu viacjazyčnosti v jednotlivých krajinách EÚ. *eTranslation* ako platformu na automatický preklad môžu bezplatne využívať všetky európske inštitúcie a orgány verejnej správy v celej EÚ. Autor sa v prezentácii bližšie zameral na technický opis prekladovej služby – predstavil používané techniky, veľké jazykové modely, a technológie umelej inteligencie, ktoré sú pri budovaní obsiahlej platformy nevyhnutné. Informácie o prekladovej platforme sú dostupné na webstránke.[2]

Hana Žižková a **Jakub Machura** (Filozofická fakulta Masarykovej univerzity v Brne) s **Adamom Frémundom** a **Janom Švecom** (Fakulta aplikovaných vied Západočeskej univerzity v Plzni) v príspevku *Is It Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction* predstavili nástroj RoBERTa na automatické vkladanie interpunkcie do textu. Autori sa zaoberali možnosťou zvýšenia presnosti nástroja RoBERTa-commas pri vkladaní vybraného interpunkčného znamienka – čiarky. Vo svojom výskume sa orientovali na frázy vo vokatíve. Hana Žižková v závere konštatovala, že pri pretrénovaní nástroja hrá dôležitú úlohu výber a štruktúra trénovacích dát.

Autori **Ján Staš, Daniel Hládek** (Fakulta elektrotechniky a informatiky Technickej univerzity v Košiciach) a **Tomáš Koctúr** (Deutsche Telekom IT & Telecommunications Slovakia s.r.o.) v príspevku *Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0* opisujú proces budovania prvého rozsiahleho korpusu na automatické odpovedanie SQuAD-sk pre slovenčinu. Dáta automaticky preložili z pôvodného anglického datasetu SQuAD v2.0 pomocou voľne dostupného frameworku Marian a s využitím anglicko-slovenského modelu Helsinky-NLP Opus. Súbor dát následne použili na trénovanie slovenského systému odpovedí na otázky. Dosiahnuté skóre pre vyladený model mBERT ukázal porovnateľné výsledky odpovedí na otázky s nedávno publikovanými strojovo preloženými súbormi údajov SQuAD pre iné európske jazyky.

Michal Škrabal (Filozofická fakulta Karlovej univerzity v Prahe) a **Karel Pioreck**ý (Ústav pre českú literatúru Akadémie vied Českej republiky v Prahe) prezentovali na konferencii príspevok *Corpus of Contemporary Poetry – A Gauntlet to Pick Up to Slovak Colleagues*. Cieľom ich výskumu bolo vytvoriť dátovú základňu a softvérové nástroje, ktoré umožnia skúmať českú poéziu publikovanú medzi rokmi 1990 až 2020, a to v širokej škále jej mediálnych podôb (printovej a webovej) a na veľkom súbore textov. S týmto cieľom sa od roku 2015 buduje Korpus současné české poezie, ktorý obsahuje približne 35,5 miliónov slov. Je určený vedcom, literárnym kritikom, študentom literatúry aj širokej verejnosti.

Prehľad historického vývoja ne/znelosti spoluhlások *r* a *l* v češtine ponúkli autori **Markéta Ziková**, **Martin Březina**, **Radek Čech** a **Pavel Kosek** (Filozofická fakulta Masarykovej univerzity v Brne) v príspevku *Syllabic Consonants in Historical Czech and How to Identify Them*. Pomocou parsera autori identifikovali v českých historických textoch zo 14. a 15. storočia potenciálne segmenty podľa zvučnosti. Zmenu *r* a *l* z neznelej na znelú spoluhlásku ovplyvňuje ich pozícia v slove – koncové spoluhlásky majú vyššiu tendenciu neznelosti.

V závere medzinárodnej konferencie SLOVKO 2023 sa všetkým prednášajúcim, zúčastneným, diskutujúcim, organizátorom i recenzentom príspevkov poďakovala Jana Levická, vedúca oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV, v. v. i. Vyzdvihla vysokú aktívnu účasť na konferencii, príjemnú a priateľskú atmosféru počas celého podujatia a zaželala všetkým pozitívne prínosy z rokovaní pre ďalšie projekty. Zároveň vyjadrila želanie, aby sa v rovnakom duchu spolupráce mohla uskutočniť aj plánovaná konferencia SLOVKO 2025.

Internetové zdroje

[1] https://www.juls.savba.sk/ediela/jc/2023/1/jc23-01.pdf (cit. 16. 11. 2023)
[2] https://commission.europa.eu/resources-partners/etranslation_en (cit. 16. 11. 2023)