LREC-COLING 2024: združena mednarodna konferenca računalniškega jezikoslovja ter jezikovnih virov in evalvacije

Mojca BRGLEZ,[1] Matej KLEMEN,[2] Luka TERČON,[1, 2] Špela VINTAR[1]

1 Filozofska fakulteta, Univerza v Ljubljani

2 Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

1. Uvod

Konec maja, natančneje 22.–25. 5., je v piemontskem Torinu potekala megakonferenca LREC-COLING 2024. Letošnja konferenca je bila edinstvena, saj je združila sicer samostojni mednarodni konferenci LREC in COLING, ki sta poprej na vsaki dve leti potekali v istem letu, odslej pa se bosta vsako leto izmenjevali. Prva je usmerjena v jezikovne vire in evalvacijo in je bila že 14. zapovrstjo. Konferenco organizira združenje ELRA[1], prej Evropsko združenje za jezikovne vire (European Language Resources Association), od letos preimenovano v regijsko neomejeno Združenje za jezikovne vire ELRA (ELRA Language Resources Association). Druga konferenca, COLING, se osredotoča na računalniško jezikoslovje in je pod okriljem Mednarodnega odbora za računalniško jezikoslovje ICCL[2] (International Committee on Computational Linguistics) potekala že tridesetič. Dogodek je bil umeščen v kongresni

1 https://www.elra.info/

2 https://ufal.mff.cuni.cz/iccl

Brglez, M. et al.: LREC-COLING 2024: združena mednarodna konferenca računalniškega jezikoslovja ter jezikovnih virov in evalvacije. Slovenščina 2.0, 12(1): 95–105.

1.19 Recenzija, prikaz knjige, kritika / Review, book review, critique

DOI: https://doi.org/10.4312/slo2.0.2024.1.95-105

center kompleksa Lingotto, ki je nekoč služil kot tovarna avtomobilov FIAT. Konferenca je potekala hibridno, del udeležencev je dogajanje spremljal in prispevke predstavljal prek spleta. Sicer, morda zaradi že tako velikega števila prispevkov in dogajanja, ni bilo čutiti velikega mešanja med vsebinami/udeleženci v živo in tistimi na daljavo. Glavni del programa je zajemal vrsto predstavitev v obliki posterja ali govorne predstavitve. Na začetku in koncu konference sta se odvili uvodna in zaključna slovesnost, ki sta bili namenjeni predvsem formalnim vidikom. Poleg glavnega, vsebinskega dela konference sta bila organizirana tudi dva dogodka, namenjena mreženju in spoznavanju italijanske glasbe in hrane.

Na uvodni slovesnosti je programski odbor predstavil analizo sprejetih člankov. Na glavni del konference je bilo prijavljenih skupno kar 3.471 prispevkov, od tega je bilo sprejetih 44 % ali 1.554 prispevkov. Letošnji delež sprejetih prispevkov se je tako gibal med bolj restriktivno politiko COLING in bolj permisivno LREC. Države z največjim zastopanjem na konferenci so bile Kitajska, ZDA, Nemčija, Francija in Japonska. Slovenija je po drugi strani slavila po največjem deležu sprejetih prispevkov, saj je bilo pri prijavi uspešnih 12 od 15 prispevkov oziroma kar 80 %. Glede obiskanosti ocenjujemo, da se je konference v živo udeležilo približno 2000 ljudi, skupaj z udeleženci na daljavo pa bi številka znala znašati vsaj dvakrat toliko. Zbornik konference z rekordno dolžino 18.801 strani je na voljo na https://aclanthology.org/2024.lrec-main.

Poleg poročila in splošnih obvestil je predsednik odbora ELRE Simon Krek na uvodni slovesnosti podelil tudi nagrado Antonio Zampolli za izjemni prispevek k napredku jezikovnih tehnologij in evalvacije jezikovnih tehnologij na področju človeških jezikovnih tehnologij (»Outstanding Contributions to the Advancement of Language Resources and Language Technology Evaluation within Human Language Technologies«). Prejemnik letošnje nagrade je Nizar Habash, redni profesor na Univerzi New York v Abu Dhabiju, ki se prvenstveno ukvarja z obdelavo arabskega jezika in arabskih narečij. Kot prejemnik nagrade je svoje delo predstavil v okviru posebnega predavanja na konferenci.

Zaključna slovesnost je bila namenjena predvsem zahvali vsem, ki so pomagali pri organizaciji megakonference: lokalnim organizator-

jem, vsebinskim koordinatorjem, prostovoljcem, recenzentom, avtorjem in obiskovalcem. Temu so sledila še vabila na naslednji, ločeni izvedbi konferenc LREC in COLING ter vabila na ostale prihajajoče jezikovnotehnološke konference.

2. Obkonferenčna srečanja

Pred začetkom glavne konference in dan po njej so potekali številni seminarji in kolokviji, skupaj se je zvrstilo kar 36 dogodkov. V nadaljevanju na kratko opišemo zgolj nekaj izbranih, ki smo se jih udeležili avtorji prispevka.

Med seminarji smo se najprej udeležili *Navigating the Modern Evaluation Landscape: Considerations in Benchmarks and Frameworks for Large Language Models,* na katerem so Leshem Choshen, Ariel Gera, Yotam Perlitz, Michal Shmueli-Scheuer in Gabriel Stanovsky temeljito predstavili metode za evalvacijo jezikovnih modelov in izzive, ki jih predstavljajo novejši generativni modeli. Z bolj dinamičnimi oblikami generiranih vsebin in novimi možnostmi uporabe namreč starejše metrike za vrednotenje, kot so natančnost, priklic, BLEU in podobne, po eni strani niso več ustrezne za vrednotenje vse bolj kreativnih vsebin, po drugi pa ne naslavljajo drugih, nejezikovnih zmogljivosti, ki jih nudijo jezikovni modeli, npr. jezikovni model kot pogovorni partner ali kot vršilec dejanj v digitalnem svetu. Še posebej veliko zanimanja je med poslušalci vzbudila debata o ugotovitvah številnih novih raziskav, ki razkrivajo, kako občutljivi so sodobni generativni jezikovni modeli na obliko poziva (angl. *prompt*), ki služi kot navodilo za ustvarjanje želenega rezultata. Že najmanjše variacije v obliki lahko namreč privedejo do velikih razlik v rezultatih, ki jih isti modeli dosežejo na različnih evalvacijskih lestvicah.

Na seminarju z naslovom *Towards a Human-Computer Collaborative Scientific Paper Lifecycle* so avtorji Qingyun Wang, Carl Edwards, Heng Ji in Tom Hope predstavili dobre in slabe plati uporabe umetne inteligence v okviru znanstvenega raziskovanja tako pri samem raziskovalnem procesu kot tudi pri pisanju prispevkov. Opisali so že preizkušene metode vključevanja sodobnih jezikovnih modelov za vsako od petih stopenj izdelave znanstvenega prispevka: pregled literature, iz-

delava hipotez, načrtovanje eksperimenta, pisanje in končni pregled. Izpostavljeni so bili številni izzivi, s katerimi se raziskovalci srečujejo ob uporabi umetne inteligence pri izvedbi raziskav. Predvsem sodobni jezikovni modeli pogosto generirajo popolnoma neresnične podatke (t. i. halucinacije), izkazujejo zelo malo domensko-specifičnega znanja in le redko proizvedejo podatke visoke kakovosti.

Med kolokviji smo najprej obiskali CogALex (*Cognitive Aspects of the Lexicon*), ki združuje raziskovalce kognitivnih vidikov jezika. Na kolokviju so bile predstavljene najnovejše raziskave o tem, kako se besede in pomeni hranijo, povezujejo in obdelujejo v našem miselnem leksikonu ter kako ta zapletena, široka in dinamična omrežja pomenov in besed predstavljati v leksikografskih virih. H kolokviju smo prispevali tudi slovenski raziskovalci, in sicer s člankom *How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene* (Žagar idr., 2024a).

Na prvi izdaji kolokvija CL4Health (*Patient-Oriented Language Processing*) smo lahko poslušali raziskave, ki se osredotočajo na obdelavo jezika za potrebe pacientov. Naraščajoče zanimanje za to temo namreč izhaja iz uporabe avtomatiziranih pomočnikov in pogostejšega iskanja zdravstvenih informacij na spletu, ki so danes veliko bolj dostopne. K področju sta prispevala tudi slovenska raziskovalca, in sicer z raziskavo *Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction* (Knez in Žitnik, 2024).

V okviru kolokvija NLPerspectives *The 3rd Workshop on Perspectivist Approaches to Natural Language Processing* je bilo govora o podatkovnem perspektivizmu. Temeljna ideja perspektivizma je ta, da nesoglasja pri označevanju podatkov niso vedno posledica napačnega odločanja označevalca ali slabih navodil za označevanje, pač pa so lahko naravno pričakovana in prisotna zaradi razlogov, kot so osebna prepričanja ali inherentna subjektivnost/dvoumnost problema, ki ga je mogoče interpretirati iz različnih perspektiv. Večina člankov je predstavljala nove vire, ki omogočajo učenje modelov na neagregiranih (torej ne-povprečenih, ne-združenih) oznakah, nekaj manj člankov pa je predstavljalo modele, naučene na podlagi tovrstnih oznak. Ideja perspektivizma je sicer prisotna tudi v mnogih člankih, predstavljenih na glavnem delu konference.

V Torinu je potekal tudi kolokvij ParlaCLARIN, med organizatorji katerega sta bila Darja Fišer, Institut za novejšo zgodovino/CLARIN ERIC, ter David Bordon, Filozofska fakulteta Univerze v Ljubljani. Na dogodku se je zvrstilo kar nekaj prispevkov, pri katerih so sodelovali slovenski raziskovalci: *Parliamentary Discourse Research in Political Science: Literature Review* (Skubic in Fišer, 2024), *Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines* (Çöltekin idr., 2024), *ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments* (de Jong idr., 2024), *Historical Parliamentary Corpora Viewer* (Kavčič idr., 2024).

Na kolokviju First Workshop on Reference, Framing, and Perspective so se predstavili še Ivačič idr. (2024) s člankom *Comparing News Framing of Migration Crises using Zero-Shot Classification*, na srečanju SIGUL za manj podprte jezike pa Ljubešić idr. (2024) s člankom *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining*.

3. Glavna konferenca

Na konferenci se je velika večina sprejetih prispevkov predstavila v obliki plakata (posterja), manjši del prispevkov pa z govornimi nastopi. Članki so bili umeščeni v 26 tematskih sklopov, od katerih so bili najbolj zastopani *Corpora and Annotation*, *Applications Involving LRs and Evaluation* ter *Evaluation and Validation Methodologies*, ki sodijo bolj v domeno konference LREC. Izbire je bilo resnično preveč, saj je običajno hkrati potekalo vsaj pet sej predstavitev in prav toliko razstav plakatov. Prvi dan je bilo mnogo teh sej na daljavo. Nekoliko neugodna je bila umestitev razstav plakatov v ločeni zgradbi, saj je to pomenilo manjšo zamudo s hojo do razstavne hale, kdaj pa nam je močan dež celo popolnoma onemogočil premik tja.

Ker je bilo poslušanja in branja vrednih raziskav ogromno, naj tukaj omenimo zgolj tiste, pri katerih so sodelovale slovenske ustanove in raziskovalci[3]. Med prispevke, ki so razgrnili nove jezikovne vire, lahko uvrstimo: *SUK 1.0: A New Training Corpus for Linguistic Annota-*

3 Število člankov s slovenskimi soavtorji, ki smo jih našli, se ne ujema z uradno statistiko, a uradnega kriterija štetja organizatorji niso navedli.

tion of Modern Standard Slovene (Arhar Holdt idr., 2024a), *Towards an Ideal Tool for Learner Error Annotation* (Arhar Holdt idr., 2024b), *SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation* (Klemen idr., 2024), *CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation* (Ljubešić in Kuzman, 2024), *The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings* (Mochtak idr., 2024), *A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection* (Dobranić idr., 2024), *MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations* (Gromann idr., 2024) in *Gos 2: A New Reference Corpus of Spoken Slovenian* (Verdonik idr., 2024). Preostala polovica prispevkov je predstavila nove metode in raziskave na področju jezika in jezikovnih tehnologij, to so bili: *A Computational Analysis of the Dehumanisation of Migrants from Syria and Ukraine in Slovene News Media* (Caporusso idr., 2024), *When Cohesion Lies in the Embedding Space: Embedding-Based Reference-Free Metrics for Topic Segmentation* (Ghinassi idr., 2024), *Denoising Labeled Data for Comment Moderation Using Active Learning* (Pelicon idr., 2024), *LLMSegm: Surface-level Morphological Segmentation Using Large Language Model* (Pranjić idr., 2024), *Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across Languages* (van Noord idr., 2024) in *SENTA: Sentence Simplification System for Slovene* (Žagar idr., 2024b).

ELRA in COLING sta najboljšim prispevkom letos podelila dve glavni nagradi. Nagrado za najboljši prispevek so prejeli Taiga Someya, Ryo Yoshida in Yohei Oseki za *Targeted Syntactic Evaluation on the Chomsky Hierarchy*, nagrado za najboljši študentski prispevek pa Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, in Rachel Bawden za *When Your Cousin Has the Right Connections: Unsupervised Bilingual Lexicon Induction for Related Data-Imbalanced Languages*. V prvem članku avtorji predstavijo izsledke o tem, kako dobro veliki jezikovni modeli prepoznajo različne lastnosti jezikov v hierarhiji Chomskega, v drugem pa avtorji izpostavijo omejitve obstoječih nenadzorovanih metod za ustvarjanje dvojezičnih slovarjev v jezikih z zelo malo jezikovnimi viri in predstavijo metodo, ki deluje bolje.

4. Vabljeni govorci

Vsak dan glavne konference je bilo moč poslušati vabljeno predavanje. Med tujima vabljenima govorcema sta bila Roger Levy z univerze MIT in Li Juanzi z univerze Tsinghua v Pekingu. Prvi je v sredo izvedel predavanje z naslovom *Large Language Models and Human Cognition*, na katerem je predstavil raziskave, ki naj bi nam prav s pomočjo jezikovnih modelov odstirale pogled na kompleksne procese človeške kognicije pri učenju in procesiranju jezika s tako imenovanim povratnim inženirstvom (ang. *reverse engineering*) pa tudi z razvijanjem modelov, ki simulirajo odraslo razumevanje otroškega govora.

Drugi dan konference je pred udeleženci nastopil Nizar Habash, prejemnik nagrade Antonio Zampolli, ki je poleg svojih znanstvenih udejstvovanj, predvsem na področju obdelave arabskega jezika, predstavil tudi svoje zasebne hobije, med drugim izmišljanje novih jezikov in umetniški projekt Palisra. V tem raziskuje možnost palestinsko-izraelske enotnosti preko ustvarjalnega združevanja obeh jezikov, pisav, kultur. Zelo zanimivo je, da na univerzi v Abu Dhabiju poučuje predmet, v katerem študenti snujejo povsem nove jezike, pri čemer se pokaže pomembnost spoznavanja prvin in struktur tujih jezikov ter kritična obravnava kdaj samoumevnih prvin in struktur lastnega maternega jezika.

Popoldne je predaval lokalni vabljeni govorec, Michele Loporcaro, ki deluje na univerzi v Zürichu. V predavanju *The Language Landscape of Italy as a Linguistic Data Mine* je predstavil delo z romanskimi narečji italijanskimi narečji, ki so večinoma nenapisani jeziki, a hranijo edinstvene jezikovne prvine in strukture, ki niso značilne niti za romansko skupino niti za indoevropsko vejo jezikov nasploh in tako raziskovalcem ponujajo neprecenljiv uvid v tipologijo jezikov.

Tretji dan konference je pred zbranimi nastopila Li Juanzi. V predavanju z naslovom *Knowledge in the LLM Era: Actuality, Challenge and Potentiality* je predstavila delo tamkajšnje raziskovalne skupnosti pri razumevanju zmogljivosti velikih jezikovnih modelov in govorila o izzivih, ki jih morajo modeli še premagati za globlje razumevanje in uporabo znanja.

5. Sklep

Ni presenetljivo, da so bili pri večini prispevkov megakonference LREC-COLING v središču pozornosti jezikovni modeli, ki vse bolj nadomeščajo tradicionalnejše pristope k obdelavi naravnega jezika, vse bolj pa vstopajo tudi na področja evalvacije in učenja jezikov. Skladno s tem se je več konferenčnih sekcij posredno ali neposredno posvečalo vprašanjem pristranskosti, etike in moralnih načel, ki naj bi jih jezikovni modeli vsebovali, ob vsem tem pa se še vedno, morda celo vse bolj, kažejo razlike med angleščino kot dominantnim jezikom in dolgim repom manjših jezikov, ki iz angleščine »podedujejo« ne le znanje, ampak tudi celo vrsto kulturnih in družbenih norm.

Spodbudno pa je, da je bilo na konferenci opaziti zelo raznolik nabor prispevkov iz najrazličnejših jezikov. Raziskovalci nekaterih manj zastopanih jezikov, kot je na primer kazaščina, so na konferenci namreč predstavili prve, pionirske korake na področju razvoja jezikovnih virov in orodij. Kljub v svetovnem merilu izjemno majhnemu številu govorcev nam to kaže, da slovenščina na področju jezikovnotehnoloških raziskav nikakor ni v slabem položaju.

Naslednjo izvedbo konference COLING bodo med 19. in 24. januarjem 2025 gostili Združeni arabski emirati, do naslednje konference LREC v letu 2026 na še nerazkriti lokaciji pa bo treba še nekoliko počakati. Do takrat vas vabimo k branju prispevkov s slovenskih ustanov, navedenih v seznamu literature.

Literatura

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon, L., & Žitnik, S. (2024a). SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 15428–15435). Pridobljeno s https://aclanthology.org/2024.lrec-main.1340/

Arhar Holdt, Š., Erjavec, T., Kosem, I., & Volodina, E. (2024b). Towards an Ideal Tool for Learner Error Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 16392–16398). Pridobljeno s https://aclanthology.org/2024.lrec-main.1424

Caporusso, J., Brglez, M., Hoogland, D., Koloski, B., Purver, M., & Pollak, S. (2024). A Computational Analysis of the Dehumanisation of Migrants from Syria and Ukraine in Slovene News Media. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 199–210). Pridobljeno s https://aclanthology.org/2024.lrec-main.18/

Çöltekin, Ç., Kopp, M., Meden, K., Morkevicius, V., Ljubešić, N., & Erjavec, T. (2024) Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 94–100). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.14/

de Jong, A., Kuzman, T., Larooij, M., & Maarten, M. (2024). ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC--COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 110–115). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.16

Dobranić, F., Evkoski, B., & Ljubešić, N. (2024). A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 695–703). Pridobljeno s https://aclanthology.org/2024.lrec-main.61/

Ghinassi, I., Wang, I., Newell, C., & Purver, M. (2024). When Cohesion Lies in the Embedding Space: Embedding-Based Reference-Free Metrics for Topic Segmentation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 17525–17536). Pridobljeno s https://aclanthology.org/2024.lrec-main.1524/

Gromann, D., Gonçalo Oliveira, H., Pitarch, L., Apostol, E.-S., Bernad, J., Bytyçi, E., ..., & Zdravkova, K. (2024). MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 11783–11793). Pridobljeno s https://aclanthology.org/2024.lrec-main.1029/

Ivačič, N., Purver, M., Lind, F., Pollak, S., Boomgaarden, H., & Bajt, V. (2024). Comparing News Framing of Migration Crises using Zero-Shot Classification. *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024* (str. 18–27). Pridobljeno s https://aclanthology.org/2024.rfp-1.3

Kavčič, A., Stojanoski, M., & Marolt, M. (2024). Historical Parliamentary Corpora Viewer. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 127–132). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.19

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2024). SI-NLI: A Slovene Natural Language Inference Dataset and its Evaluation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 14859–14870). Pridobljeno s https://aclanthology.org/2024.lrec-main.1294/

Knez, T., & Žitnik, S. (2024). Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction. *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024* (str. 131–136). Pridobljeno s https://aclanthology.org/2024.cl4health-1.16/

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 3271–3282). Pridobljeno s https://aclanthology.org/2024.lrec-main.291/

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (str. 189–203). Pridobljeno s https://aclanthology.org/2024.sigul-1.23

Mochtak, M., Rupnik, P., & Ljubešić, N. (2024). The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 16024–16036). Pridobljeno s https://aclanthology.org/2024.lrec-main.1393/

Pelicon, A., Karan, M., Shekhar, R., Purver, M., & Pollak, S. (2024). Denoising Labeled Data for Comment Moderation Using Active Learning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 4626–4633). Pridobljeno s https://aclanthology.org/2024.lrec-main.413/

Pranjić, M., Robnik-Šikonja, M., & Pollak, S. (2024). LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 10665–10674). Pridobljeno s https://aclanthology.org/2024.lrec-main.933/

Skubic, J., & Fišer, D. (2024). Parliamentary Discourse Research in Political Science: Literature Review. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora* (str. 1–11). Pridobljeno s https://aclanthology.org/2024.parlaclarin-1.1/

van Noord, R., Kuzman, T., Rupnik, P., Nikola Ljubešić, N., Esplà-Gomis, M., Ramírez-Sánchez, G., & Toral, A. (2024). Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across Languages. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 5221–5234). Pridobljeno s https://aclanthology.org/2024.lrec-main.465/

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 7825–7830). Pridobljeno s https://aclanthology.org/2024.lrec-main.691/

Žagar, A., Brglez, M., & Vintar, Š. (2024a). How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024* (str. 42–48). Pridobljeno s https://aclanthology.org/2024.cogalex-1.5/

Žagar, A., Klemen, M., Robnik-Šikonja, M., & Kosem, I. (2024b). SENTA: Sentence Simplification System for Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 14687–14692). Pridobljeno s https://aclanthology.org/2024.lrec-main.1279/