Светла Коева, Валентина Стефанова

ЕВРОПЕЙСКО ЕЗИКОВО РАВЕНСТВО

Статията представя накратко широкомащабно проучване, имащо за цел да проследи и анализира технологичната поддръжка за европейските езици (с фокус върху българския език). Изследването се стреми не само да очертае състоянието на езиковите технологии към момента, но и да набележи пропуските, както и факторите, които възпрепятстват развитието на научните изследвания и технологиите в областта. Резултат от анализа е разработването на програма за стратегически изследвания и иновации, както и на пътна карта за постигане на равенство на европейско равнище по отношение на развитието на езиковите технологии и тяхното ефективно използване.

Проучването е проведено от повече от 40 изследователски институции за повече от 30 европейски езика в рамките на проекта „Европейско езиково равенство“[1] и включва процедура за събиране на данни, които предоставят информация за технологичната поддръжка за европейските езици в началото на 2022 г.

Изследването на съществуващите езикови технологии за български език се базира основно на колекцията от ресурси, програми за обработка на езика и услуги, събрани и разпространявани от „Европейска езикова мрежа“[2], но също и от източници като ГитХъб[3], МЕТА-ШЕЪР[4], КЛАРИН[5] и др. Като цяло езиковите технологии обхващат широка интердисциплинарна научна област, която се занимава с изучаването и разработването на системи, способни да обработват, анализират, възпроизвеждат и „разбират“ човешките езици, независимо дали са в писмена, или в устна форма. Езиковите технологии предоставят решения за следните основни области: автоматичен анализ на текста, обработка на реч, машинен превод, автоматично извличане на информация, генериране на естествен език, взаимодействие между човек и компютър. По-долу ще представим накратко развитието на езиковите технологии за български език за съответните области.

Няколко съвременни библиотеки за компютърна обработка на езика предоставят множество от инструменти, които са неизменна част от повечето приложения, базирани на езикови технологии: токънизация (разделяне текста на последователности от символи); разделяне на изречения; откриване на границите на абзаци; проверка на правописа; тагиране (разпознаване на граматичните характеристики на думите или морфологичен анализ); лематизация (извеждане на основната форма на думите); разпознаване на имена на лица, организации и географски обекти; синтактичен и семантичен анализ и др. Част от тях предоставят техники за „дълбоко обучение“ и очертаване на граф на знанието за семантично търсене, както и отчитат добро ниво на точност и бързодействие (например Спарк еНеЛПи [1]). Тези библиотеки обаче не включват предварително обучени модели за български език. Сравнително скоро станаха достъпни две системи (включващи токънизация, разделяне на изречения, морфологичен анализ, лематизация и анализ на синтактични зависимости), които са обучени за езици с ресурси за универсални синтактични зависимости, включително и за български: ЮДипайп [2] и еНеЛПи-Кюб [3]. Друга система за компютърна обработка на българския език интегрира изброените фунционалности и разпознаване на имена на лица, организации и географски обекти, „плитък“ синтактичен анализ за разпознаване на именни фрази и откриване на термини [4].

Текстовият анализ все още доминира в областта на езиковите технологии за български език [5], а мултимодални данни (текст, изображения, аудио и видео) се обработват рядко едновременно, въпреки че прогнозите сочат, че видеосъдържанието скоро ще доминира в интернет.

Наличието на езикови модели за даден език е съществена предпоставка за развитието на приложения за компютърна обработка. Обучението на подобни модели отнема много време и изисква голямо количество подходящи ресурси. Могат да бъдат посочени: моделът БЕРТ за разпознаване на имена на лица, организации и локации на няколко славянски езика, включително български [6], многоезиковите модели РоБЪРТа бейз [7] и ЕКСеЛеМ-аР [8], които се базират на последващо обучение на големи трансформационни модели за много езици, сред които и за български. Разработването на подходящи модели за български език ще

1 https://european-language-equality.eu/
2 https://live.european-language-grid.eu/
3 https://github.com/
4 http://www.meta-net.eu/meta-share
5 https://www.clarin.eu/

помогне за успешното решаване на редица задачи като автоматично отговаряне на въпроси, резюмиране на текстове, маркиране на семантични роли, определяне на кореферентни връзки, извличане на имена на лица, организации и географски обекти и анализ на настроенията на потребителите по отношение на продукти, събития и организации.

Въпреки че се разработват от няколко десетилетия, качеството на технологиите за обработка на българска реч все още не е задоволително. Основната причина е, че за български няма достатъчно езикови ресурси. Българските речеви корпуси са разработени за специфични изследователски цели и са със сравнително малък обем, например БеГеСпийч[6] – корпус от транскрибирана разговорна реч, или БулФонСи[7] – корпус от съобщителни и въпросителни изречения, прочетени от 147 различни диктори. СпийчЛаб 2.0[8] е система за синтез на реч, специално разработена за български език. Синтезаторът СкайКоде титиес[9] се предлага за онлайн употреба и като приложение за Андроид.

Въпреки че има алгоритми за разпознаване на реч, специално разработени за български език [9], все още няма достъпна система за преобразуване на реч към текст за български език, макар че би била приложима в различни социални сфери. В рамките на проекта „Синхронен преводач на европейски езици“[10] е разработена автоматична система за субтитриране на срещи на живо и конференции, осигуряваща устен езиков превод от английски на български език в реално време.

Освен автоматичен превод, който се предлага от гиганти като Гугъл и Майкрософт, съществуват и други системи за машинен превод от и на български език: СИСТРАН преводач[11] – услуга за превод на текстови фрагменти, базирана на технологията за невронен машинен превод; ДийпеЛ преводач[12] – система, базирана на невронни мрежи, която е способна да улавя нюанси и да ги възпроизвежда при превод, и т.н. Потребителски невронни мрежи за машинен превод от български на английски и обратно са разработени по проекта „Автоматичният превод на Механизма за свързване на Европа за Председателството на Съвета на Европа“[13]. Платформата иТранслейшън, разработвана от Европейската комисия, предлага невронен машинен превод на 24-те официални езика на Европейския съюз и е достъпна (за публичната администрация и малките и средните предприятия) чрез приложно-програмен интерфейс, уебинтерфейс и услугата „Преводач за социални медии“[14].

Оценката на качеството на съществуващите услуги за машинен превод, броят на езиковите двойки и обхватът на тематичните области все още определят технологиите за машинен превод за българския език като недостатъчно развити.

Напоследък се наблюдава сериозен напредък в изследванията, базирани на автоматичното извличане на информация за български: проследяване на събития [10]; анализ на настроенията на потребителите [11]; откриване на фалшиви новини [12, 13]; прогнозиране на резултати [14]; проверка на достоверността на факти [15].

Съществуват популярни системи за взаимодействие между човек и компютър като Волфрам Алфа и АйБиЕм Уотсън. Към момента личните асистенти като Алекса или услуги като ОК Гугъл или Сири не поддържат български език.

При съпоставка на езиковите технологии за европейски езици се забелязват следните тенденции. През последните години езиковите технологии отбелязват забележителен напредък. Появата на „дълбоко обучение“ и невронни мрежи, заедно със значителното увеличаване на броя и качеството на ресурсите за различни езици, водят до забележителен напредък, който обаче не е достатъчен, нито е еднакъв за всички езици. За да бъде сравнено нивото на технологична поддръжка за европейските езици, в рамките на проекта „Европейско езиково равенство“ бяха разгледани повече от 11 500 езикови приложения и ресурси, събрани в каталога на платформата „Европейска езикова мрежа“ (към 2022 г.). Сравнението е направено в редица основни области на приложение на езиковите технологии (подобно на сходно изследване през 2012 г. [16]) като:

– Обработка на текста (например маркиране на част на речта, синтактичен анализ);

– Извличане на информация (например търсене и извличане на информация);

– Системи за превод (например машинен превод, компютърно подпомогнат превод);

– Генериране на естествен език (например автоматични резюмета);

– Обработка на реч (например синтез на реч, разпознаване на реч);

– Обработка на изображения или видео (например лицево разпознаване);

– Взаимодействие между човек и компютър (например инструменти за генериране на диалог).

Направена е оценка за относителната технологична поддръжка за 87 национални и регионални европейски езика по отношение на посочените категории в следната четиристепенна скала

1. Слаба или липсваща поддръжка: езикът присъства (като съдържание, входен или изходен език) в по-малко от 3 % от ресурсите от същия тип;

[6] http://bgspeech.net
[7] http://lml.bas.bg/BulPhonC/
[8] https://play.google.com/store/apps/details?id=org.bacl.android.speechlab2g&hl=en_GB&gl=BG
[9] https://tts.skycode.com/
[10] https://elitr.eu/
[11] https://www.systran.us
[12] https://www.deepl.com/translator
[13] CEF Automated Translation for the EU Council Presidency
[14] CEF Social Media Translator

2. Частична поддръжка: езикът присъства в повече от 3 % и по-малко от 10 % от ресурсите от същия тип;

3. Умерена поддръжка: езикът присъства в повече от 10 % и по-малко от 30 % от ресурсите от същия тип;

4. Добра поддръжка: езикът присъства в повече от 30 % от ресурсите от същия тип.

Общото ниво на поддръжка за даден език е изчислено въз основа на средното покритие във всички изследвани категории.

Единственият език, който е класифициран в групата за добра поддръжка, е английският език. Френски, немски и испански формират група на езици с умерена поддръжка. Всички други официални езици на Европейския съюз, включително и българският език, са оценени като езици с частична поддръжка (с изключение на ирландски и малтийски, които имат слаба или липсваща поддръжка в отделните категории).

При езиковите технологии за български доминира анализът на текста, а обработката на многомодални данни се предлага рядко. Все още автоматичният превод от и на български език не се характеризира с отлично качество, особено когато текстовете включват преносни и идиоматични изрази. Често използвани и нужни на хората езикови технологии не са разработени за български език (например за взаимодействие между човек и компютър, за едновременна обработка на текст, реч, изображение или видео), а за други, макар че има известен напредък на технологично равнище, няма разработени приложения за широко използване (например за автоматично резюмиране на съдържанието на документи, за автоматично отговаряне на въпроси и др.). Наблюдава се съществена разлика между броя и качеството на езиковите технологии не само между английски и български език, но и между езици като немски, френски, испански и български език.

В заключение може да се обобщи, че езиковите технологии като цяло бележат значителен напредък през последните десет години. Разликата обаче между езиците с развити езикови технологии и езиците със слаборазвити езикови технологии се запазва и през 2022 г. Тази разлика трябва да бъде преодоляна (или на първо време значително намалена), за да се осигури равнопоставеност на европейските езици по отношение на възможностите за компютърна обработка, основана на езикови технологии.

ЛИТЕРАТУРА

[1] *Zaharia, M., R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, Sh. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, Sc. Shenker, and I. Stoica*. Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11):56-65, November 2016. ISSN 0001-0782. doi: 10.1145/2934664.

[2] *Straka, M., J. Straková*. Tokenizing, POS tagging, lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88-99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[3] *Boroș, T., St. Dumitrescu, and R. Burtica.* NLP-cube: End-to-end raw text processing with neural networks. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 171-179, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K18-2017.

[4] *Koeva, Sv., N. Obreshkov, and M. Yalamov.* Natural language processing pipeline to annotate Bulgarian legislative documents. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6988-6994, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.863.

[5] *Hristova, Gl.* Text analytics in Bulgarian: An overview and future directions. Cybernetics and Information Technologies, 21(3):3-23, 2021. doi: doi:10.2478/cait-2021-0027. URL https://doi.org/10.2478/cait-2021-0027.

[6] *Arkhipov, M., M. Trofimova, Y. Kuratov, and A. Sorokin.* Tuning multilingual Transformers for language-specific Named Entity Recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 89-93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3712. URL https://aclanthology.org/W19-3712.

[7] *Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov.* Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[8] *Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov*. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440-8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.aclmain.747.

[9] *Mitankin, P., St. Mihov, and T. Tinchev.* Large vocabulary continuous speech recognition for bulgarian. In Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria, pages 246-250, 2009. URL https://aclanthology.org/R09-1046/.

[10] *Tanev, Hr., and J. Steinberger*. Semi-automatic acquisition of lexical resources and grammars for event extraction in Bulgarian and Czech. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 110-

118, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2416.

[11] *Kapukaranov, B., and Pr. Nakov*. Fine-grained sentiment analysis for movie reviews in Bulgarian. In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 266-274, Hissar, Bulgaria, September 2015. INCOMA Ltd. URL https://aclanthology.org/R15-1036.

[12] *Dinkov, Yo., I, Koychev, and Pr. Nakov.* Detecting Toxicity in News Articles: Application to Bulgarian. arXiv preprint arXiv:1908.09785, 2019.

[13] *Karadzhov, G., P. Gencheva, Pr. Nakov, and I. Koychev*. We built a fake news & click-bait filter: What happened next will blow your mind! CoRR, abs/1803.03786, 2018. URL http://arxiv.org/abs/1803.03786.

[14] *Velichkov, B., I. Koychev, and Sv. Boytcheva.* Deep learning contextual models for prediction of sport event outcome from sportsman's interviews. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_142. URL https://aclanthology.org/R19-1142.

[15] *Atanasova, P., Pr. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, Ts. Mihaylova, M. Mohtarami, and J. Glass.* Automatic fact-checking using context and discourse information. Journal of Data and Information Quality, 11(3):1-27, Jul 2019. ISSN 1936-1963. doi: 10.1145/3297722. URL http://dx.doi.org/10.1145/3297722.

[16] *Blagoeva, D., Sv. Koeva, and Vl. Murdarov*. Българският език в дигиталната епоха – The Bulgarian Language in the Digital Age. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012. Georg Rehm and Hans Uszkoreit (series editors). http://www.meta-net.eu/whitepapers/e-book/bulgarian.pdf

Svetla Koeva, Valentina Stefanova

EUROPEAN LANGUAGE EQUALITY

(Abstract)

This paper reports on the current status of technology support for the Bulgarian language and highlights the identified gaps that should be overcome by further development of research and technology.

Language Technology (LT) is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, irrespective of their being written, spoken or embodied. LT is trying to provide solutions for the following main application areas: Text Analysis; Speech Processing; Machine Translation; Information Extraction and Information Retrieval; Natural Language Generation; Human-Computer Interaction.

The study of the available language technology for Bulgarian is based on the overall collection recorded on the European Language Grid (ELG) in February 2022. The observations are briefly as follows. Text analysis still dominates the field of Bulgarian language technology, and multimodal input data, such as simultaneous text, images, audio and video, are rarely processed, although forecasts indicate that video content will soon dominate on the internet. There are also applications for automatically translating language, although these still fail to produce linguistically and idiomatically correct translations, especially when Bulgarian is the target language. Many commonly used and necessary technologies are still not available for Bulgarian (human-computer interaction, multimodal processing, etc.) and for others, even if some advance in technologies is recorded, there are no certain available applications (summarisation, question answering, etc.). Many technologies are advanced abroad and Bulgarian is part of some multilingual systems for machine translation, speech analysis and recognition.

The general conclusion is that there is still a yawning technological gap between English and Bulgarian and even between German, French, Italian, Spanish and Bulgarian. A comparison of international technology and the one for Bulgarian shows that results for the automatic analysis of English and of some other European languages are far better than those for Bulgarian. The LT field as a whole has significantly progressed in the last ten years. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality.