

Piotr Bański/Nils Diewald/Marc Kupietz/Beata Trawiński

APPLYING THE NEWLY EXTENDED EUROPEAN REFERENCE CORPUS EuReCo

Pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish

Keywords Comparable corpora; collocation analysis; light-verb constructions; EuReCo

It is well known that the distribution of lexical and grammatical patterns is size- and register-sensitive (Biber 1986, and later publications). This fact alone presents a challenge to many corpus-oriented linguistic studies focusing on a single language. When it comes to cross-linguistic studies using corpora, the challenge becomes even greater due to the lack of high-quality multilingual corpora (Kupietz et al. 2020; Kupietz/Trawiński 2022), which are comparable with respect to the size and the register. That was the motivation for the creation of the European Reference Corpus EuReCo, an initiative started in 2013 at the Leibniz Institute for the German Language (IDS) together with several European partners (Kupietz et al. 2020). EuReCo is an emerging federated corpus, with large virtual comparable corpora across various languages and with an infrastructure supporting contrastive research. The core of the infrastructure is KorAP (Diewald et al. 2016), a scalable open-source platform supporting the analysis and visualisation of properties of texts annotated by multiple and potentially conflicting information layers, and supporting several corpus query languages.

Until recently, EuReCo consisted of three monolingual subparts: the German Reference Corpus DEREKO (Kupietz et al. 2018), the Reference Corpus of Contemporary Romanian Language (Barbu Mititelu/Tuşiş/Irimia 2018), and the Hungarian National Corpus (Váradi 2002). The goal of the present submission is twofold. On the one hand, it reports about the new component of EuReCo: a sample of the National Corpus of Polish (Przepiórkowski et al. 2010). On the other hand, it presents the results of a new pilot study using the newly extended EuReCo. This pilot study investigates selected Polish collocations involving light verbs and their prepositional / nominal complements (Fig. 1) and extends the collocation analyses of German, Romanian and Hungarian (Fig. 2) discussed in Kupietz/Trawiński (2022).

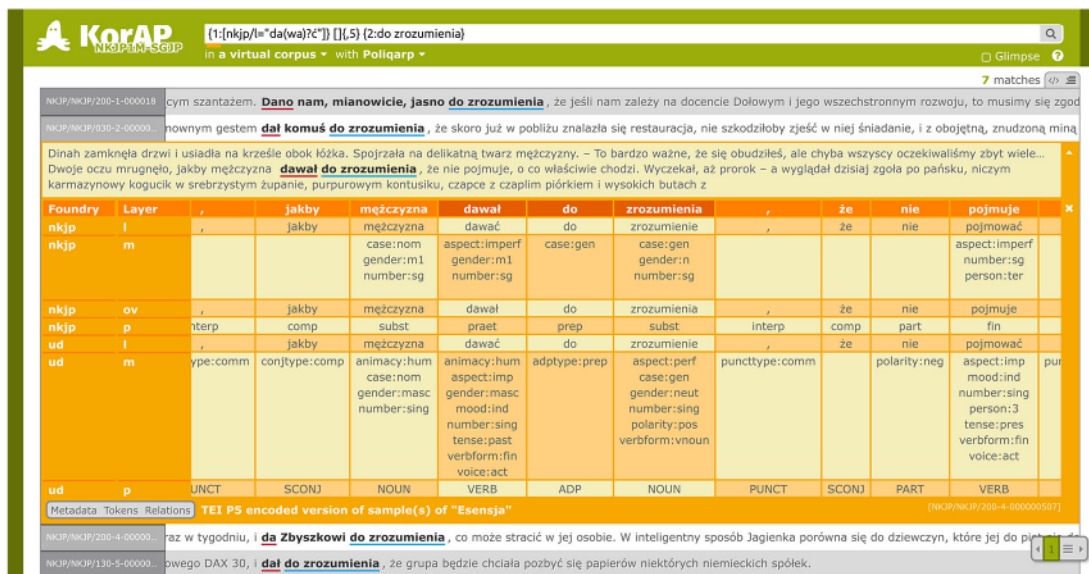


Fig. 1: Light verb constructions in Polish: concordances and PoS-annotation of *da(wa)ć do zrozumienia* (= to give sb. to understand)

<spune> în <NN> / CoRoLa			in <NN> <setzen> / vc_drukola		LVC example	FN (DeepL)	logDice
NN	logDice	EN (-DeepL)	<NN>	logDice			
pericol	11,16	Danger	Gang	10,84	nyilvánosságra hozott	disclosed	12.4
aplicare	10,74	Application	Szene	10,59	hozzuk nyilvánosságra	we publish	12.3
mișcare	10,63	Move	Brand	10,12	hoznak létre	are created	11.5
discuție	10,07	Discussion	Kenntnis	9,55	helyzetbe hozza	puts you in a position	11.3
funcțiune	9,97	Function	Bewegung	9,44	Malév-pilóta hozta Budapestre	Brought to Budapest by a Malév pilot	10.7
evidență	9,64	Highlight	Verbindung	9,16	világost hozza előnybe	worldly preference	10.5
partieă	8,05	Participo	March	8,07	pótegyezményt hozott ajándékba	a replacement convention as a gift	10.3
					szégyent hozott Magyarországra	brought shame to Hungary	10.1
					forgalomba hozott	placed on the market	10.0
					rendbe hozni	fix	10.0

Fig. 2.: Light Verb Construction comparison Romanian-German (left) and analysis Hungarian (right) using DEReKo, CoRoLa, HNC and the KorAP-APIs

References

Barbu Mititelu, Verginica/Tufiș, Dan/Irimia, Elena (2018): The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo,Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (eds.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1189>.

Biber, Douglas (1986): Spoken and written textual dimensions in English: Resolving the contradictory findings. In: Language 62 (2), pp. 384–414. <https://doi.org/10.2307/414678>.

Diewald, Nils/Hanl, Michael/Margaretha, Eliza/Bingel, Joachim/Kupietz, Marc/Bański, Piotr/Witt, Andreas (2016): KorAP architecture – diving in the deep sea of corpus data. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3586–3591. <https://www.aclweb.org/anthology/L16-1569>.

Kupietz, Marc/Trawiński, Beata (2022): Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo. In: Auteri, Laura/Barrale, Natascia/Di Bella, Arianna/Hoffmann, Sabine (eds.): Wege der Germanistik in transkultureller Perspektive. Akten des XIV. Kongresses der Internationalen Vereinigung für Germanistik (IVG). Volume 6. Bern: Lang (= Jahrbuch für Internationale Germanistik – Beihefte 6), pp. 417–439.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New developments – new opportunities. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (eds.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1689>.

Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, Sylviane/Lefer, Marie-Aude (eds.): Translating and Comparing Languages: Corpus-based Insights. Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 257–273.

Przepiórkowski, Adam/Górski, Rafał L./Łaziński, Marek/Pęzik, Piotr (2010): Recent developments in the National Corpus of Polish. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (eds.): Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/152_Paper.pdf.

Váradi, Tamás (2002): The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). Las Palmas: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf>.

Contact information

Piotr Bański

Leibniz-Institut für Deutsche Sprache
banski@ids-mannheim.de

Nils Diewald

Leibniz-Institut für Deutsche Sprache
diewald@ids-mannheim.de

Marc Kupietz

Leibniz-Institut für Deutsche Sprache
kupietz@ids-mannheim.de

Beata Trawiński

Leibniz-Institut für Deutsche Sprache
trawinski@ids-mannheim.de

Bibliographical information

This text is part of the publication: Trawiński, Beata/Kupietz, Marc/Proost, Kristel/Zinken, Jörg (eds.) (2023): 10. International Contrastive Linguistics Conference (ICLC). Book of Abstracts (pre-conference version). Mannheim: IDS-Verlag.