Piotr Bański, Nils Diewald, Marc Kupietz and Beata Trawiński

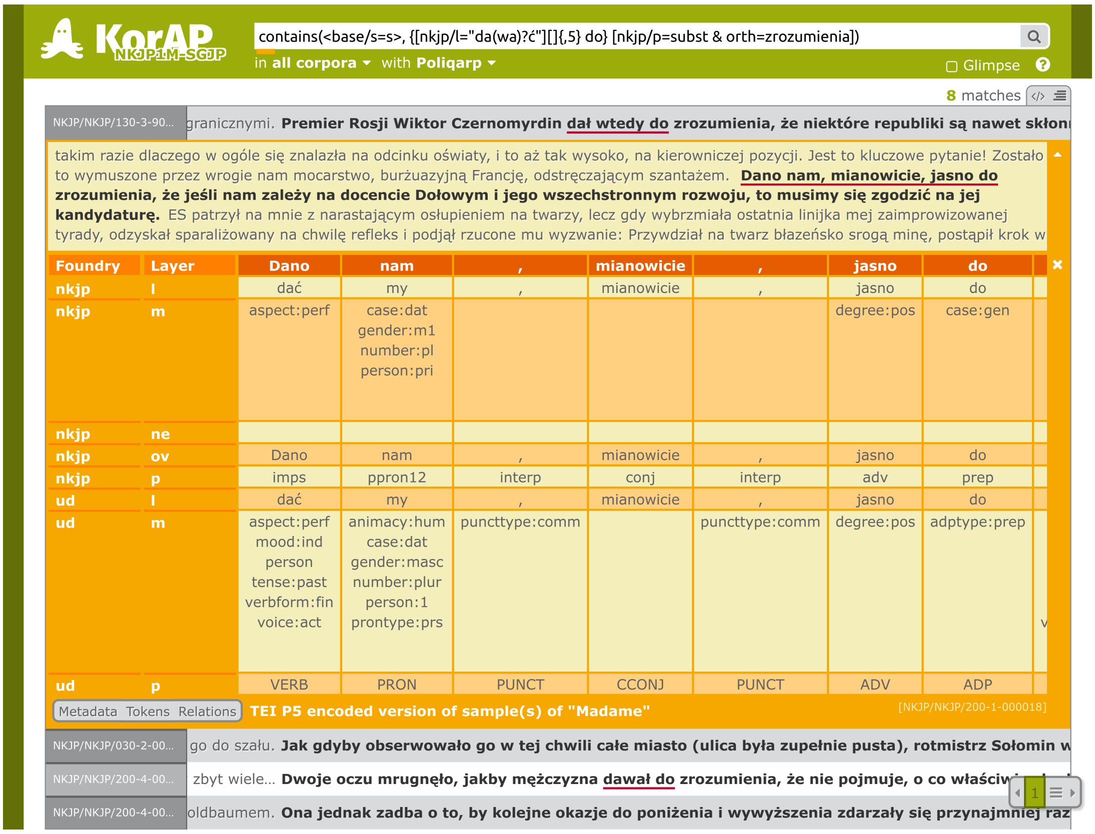
Applying the newly extended European Reference Corpus EuReCo

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 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/Tufiş/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 and Trawiński (2022).



contains(<base/s=s>, {[nkjp/l="da(wa)?ć"][]{,5} do} [nkjp/p=subst & orth=zrozumienia])

in **all corpora** with **Poliqarp**

□ Glimpse

8 matches

NKJP/NKJP/130-3-90... granicznymi. **Premier Rosji Wiktor Czernomyrdin dał wtedy do zrozumienia, że niektóre republiki są nawet skłon**

takim razie dlaczego w ogóle się znalazła na odcinku oświaty, i to aż tak wysoko, na kierowniczej pozycji. Jest to kluczowe pytanie! Zostało to wymuszone przez wrogie nam mocarstwo, burżuazyjną Francję, odstręczającym szantażem. **Dano nam, mianowicie, jasno do zrozumienia, że jeśli nam zależy na docencie Dołowym i jego wszechstronnym rozwoju, to musimy się zgodzić na jej kandydaturę.** ES patrzył na mnie z narastającym osłupieniem na twarzy, lecz gdy wybrzmiała ostatnia linijka mej zaimprowizowanej tyrady, odzyskał sparaliżowany na chwilę refleks i podjął rzucone mu wyzwanie: Przywdział na twarz błazeńsko srogą minę, postąpił krok w

Foundry	Layer	Dano	nam	,	mianowicie	,	jasno	do
nkjp	l	dać	my	,	mianowicie	,	jasno	do
nkjp	m	aspect:perf	case:dat gender:m1 number:pl person:pri				degree:pos	case:gen
nkjp	ne							
nkjp	ov	Dano	nam	,	mianowicie	,	jasno	do
nkjp	p	imps	ppron12	interp	conj	interp	adv	prep
ud	l	dać	my	,	mianowicie	,	jasno	do
ud	m	aspect:perf mood:ind person tense:past verbform:fin voice:act	animacy:hum case:dat gender:masc number:plur person:1 prontype:prs	puncttype:comm		puncttype:comm	degree:pos	adptype:prep
ud	p	VERB	PRON	PUNCT	CCONJ	PUNCT	ADV	ADP

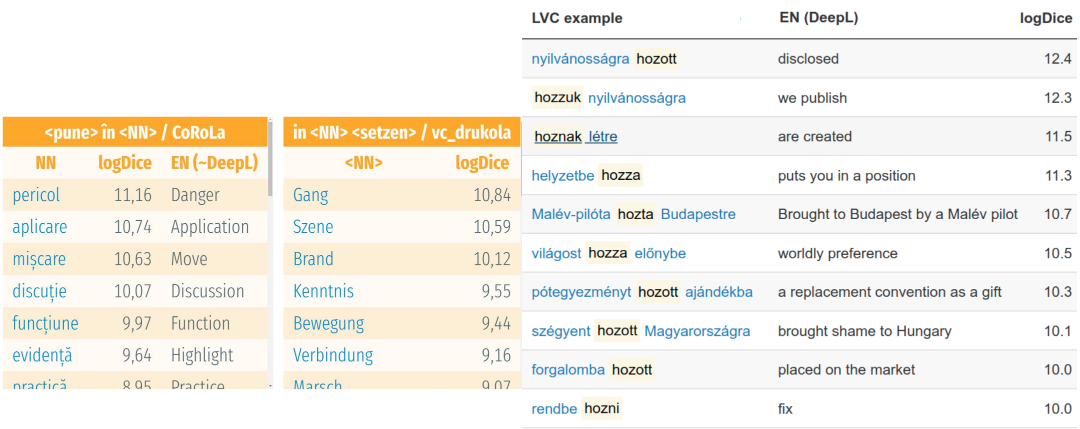
Metadata Tokens Relations **TEI P5 encoded version of sample(s) of "Madame"** [NKJP/NKJP/200-1-000018]

NKJP/NKJP/030-2-00... go do szału. **Jak gdyby obserwowało go w tej chwili całe miasto (ulica była zupełnie pusta), rotmistrz Sołomin w**

NKJP/NKJP/200-4-00... zbyt wiele... **Dwoje oczu mrugnęło, jakby mężczyzna dawał do zrozumienia, że nie pojmuje, o co właściw**

NKJP/NKJP/200-4-00... oldbaumem. **Ona jednak zadba o to, by kolejne okazje do poniżenia i wywyższenia zdarzały się przynajmniej raz**

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



<pune> în <NN> / CoRoLa		
NN	logDice	EN (~DeepL)
pericol	11,16	Danger
aplicare	10,74	Application
mișcare	10,63	Move
discuție	10,07	Discussion
funcțiune	9,97	Function
evidență	9,64	Highlight
practică	8,95	Practice

in <NN> <setzen> / vc_drukola	
<NN>	logDice
Gang	10,84
Szene	10,59
Brand	10,12
Kenntnis	9,55
Bewegung	9,44
Verbindung	9,16
Marsch	9,07

LVC example	EN (DeepL)	logDice
nyilvánosságra hozott	disclosed	12.4
hozzuk nyilvánosságra	we publish	12.3
hoznak létre	are created	11.5
helyzetbe hozza	puts you in a position	11.3
Malév-pilóta hozta Budapestre	Brought to Budapest by a Malév pilot	10.7
világost hozza előnybe	worldly preference	10.5
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/Tufiş, Dan/Irimia, Elena (2018): The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: 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: 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/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: 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.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1689.

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) (Bd. 6). Bern: Peter Lang (Jahrbuch für Internationale Germanistik - Beihefte - 6), pp. 417–439.

Przepiórkowski, Adam/Górski, Rafa\l L./\Laziński, Marek/Pęzik, Piotr (2010): Recent Developments in the National Corpus of Polish. In: 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, Canary Islands - Spain: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf.