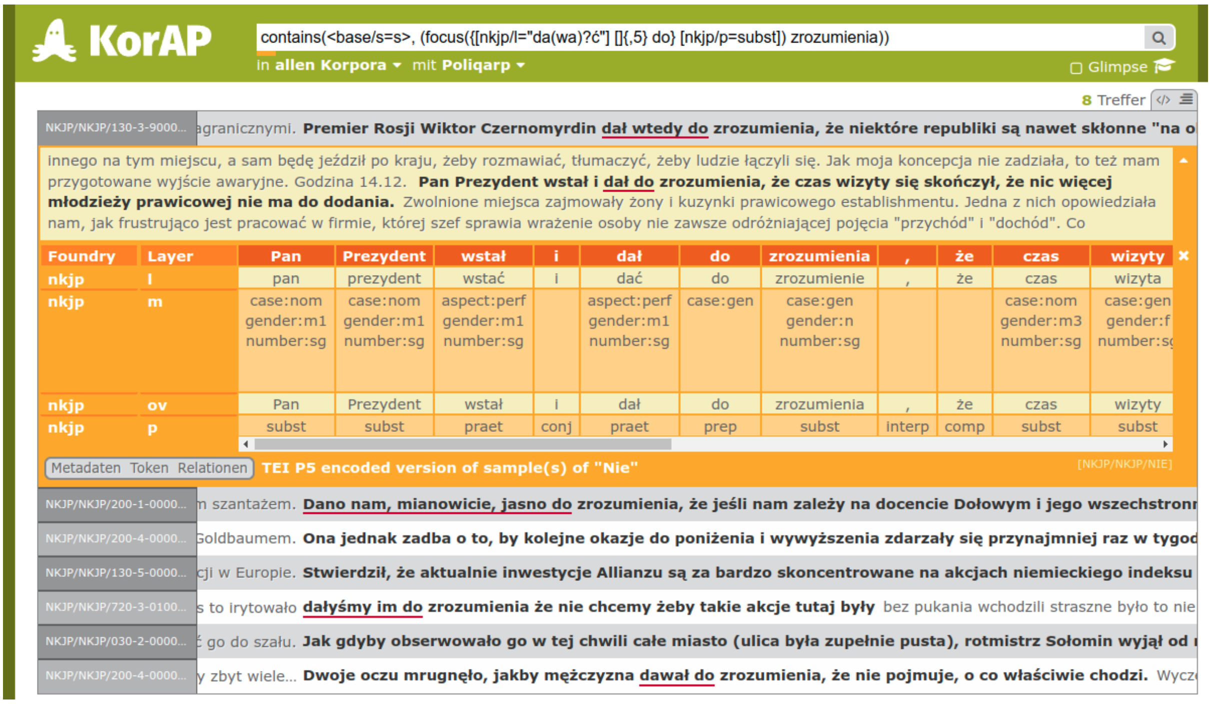
Applying the newly extended European Reference Corpus EuReCo: pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish

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 and Trawiński, to appear), 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 et al., 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 (to appear).



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

NKJP/NKJP/130-3-9000... agranicznymi. **Premier Rosji Wiktor Czernomyrdin dał wtedy do zrozumienia, że niektóre republiki są nawet skłonne "na o**

innego na tym miejscu, a sam będę jeździł po kraju, żeby rozmawiać, tłumaczyć, żeby ludzie łączyli się. Jak moja koncepcja nie zadziała, to też mam przygotowane wyjście awaryjne. Godzina 14.12. **Pan Prezydent wstał i dał do zrozumienia, że czas wizyty się skończył, że nic więcej młodzieży prawicowej nie ma do dodania.** Zwolnione miejsca zajmowały żony i kuzynki prawicowego establishmentu. Jedna z nich opowiedziała nam, jak frustrująco jest pracować w firmie, której szef sprawia wrażenie osoby nie zawsze odróżniającej pojęcia "przychód" i "dochód". Co

Foundry	Layer	Pan	Prezydent	wstał	i	dał	do	zrozumienia	,	że	czas	wizyty
nkjp	l	pan	prezydent	wstać	i	dać	do	zrozumienie	,	że	czas	wizyta
nkjp	m	case:nom gender:m1 number:sg	case:nom gender:m1 number:sg	aspect:perf gender:m1 number:sg		aspect:perf gender:m1 number:sg	case:gen	case:gen gender:n number:sg			case:nom gender:m3 number:sg	case:gen gender:f number:s
nkjp	ov	Pan	Prezydent	wstał	i	dał	do	zrozumienia	,	że	czas	wizyty
nkjp	p	subst	subst	praet	conj	praet	prep	subst	interp	comp	subst	subst

TEI P5 encoded version of sample(s) of "Nie" [NKJP/NKJP/NIE]

NKJP/NKJP/200-1-0000... n szantażem. **Dano nam, mianowicie, jasno do zrozumienia, że jeśli nam zależy na docencie Dołowym i jego wszechstronı**

NKJP/NKJP/200-4-0000... Goldbaumem. **Ona jednak zadba o to, by kolejne okazje do poniżenia i wywyższenia zdarzały się przynajmniej raz w tygod**

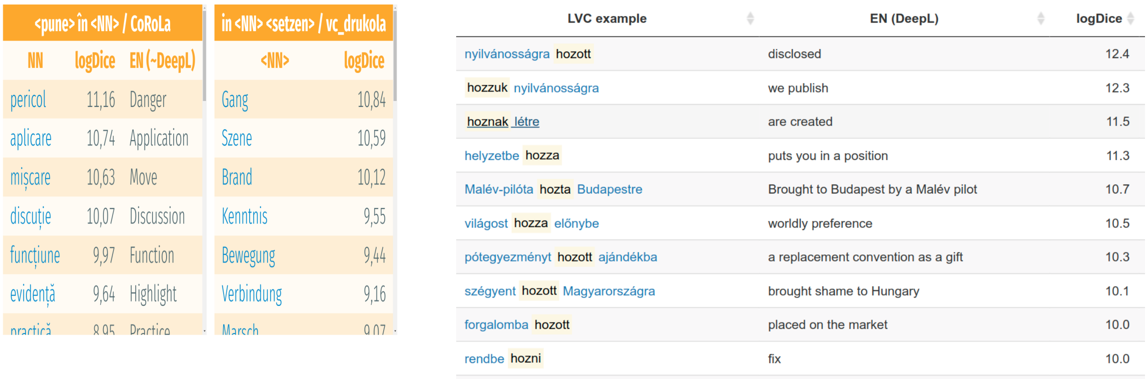
NKJP/NKJP/130-5-0000... cji w Europie. **Stwierdził, że aktualnie inwestycje Allianzu są za bardzo skoncentrowane na akcjach niemieckiego indeksu**

NKJP/NKJP/720-3-0100... s to irytowało **dałyśmy im do zrozumienia że nie chcemy żeby takie akcje tutaj były** bez pukania wchodzili straszne było to nie

NKJP/NKJP/030-2-0000... ć go do szału. **Jak gdyby obserwowało go w tej chwili całe miasto (ulica była zupełnie pusta), rotmistrz Sołomin wyjął od ı**

NKJP/NKJP/200-4-0000... y zbyt wiele... **Dwoje oczu mrugnęło, jakby mężczyzna dawał do zrozumienia, że nie pojmuje, o co właściwie chodzi.** Wyczı

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, V., Tufiş, D., Irimia, E., 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa), in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Biber, D. (1986) Spoken and written textual dimensions in English: resolving the contradictory findings. *Language*, 62, 384–414.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., Witt, A., 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). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3586–3591.

Kupietz, M., Diewald, N., Trawiński, B., Cosma, R., Cristea, D., Tufiş, D., Váradi, T., Wöllstein, A., 2020. Recent developments in the European Reference Corpus EuReCo. Transl. Comp. Lang. Corpus-Based Insights Sel. Proc. Fifth Using Corpora Contrastive Transl. Stud. Conf. Louvain--Neuve Press. Univ. Louvain, Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference 257–273.

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A., 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Kupietz, M. and Trawiński, B. (appears). Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo, in: Janusz Taborek, Henning Lobin, Fabio Mollica (Eds.): *Kontrastive Korpuslinguistik.* Akten des XIII. Internationalen Germanistenkongresses Shanghai 2015: Germanistik zwischen Tradition und Innovation. Peter Lang Verlag.

Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P., 2010. Recent Developments in the National Corpus of Polish, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta.

Váradi, T., 2002. The Hungarian National Corpus, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain.