IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Applying the newly extended European Reference Corpus EuReCo

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

Piotr Bański, Nils Diewald, Marc Kupietz and Beata Trawiński

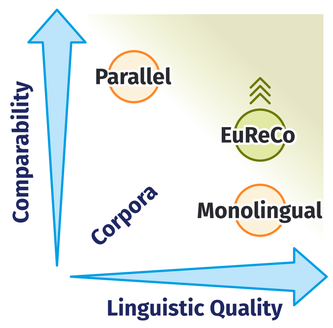
The EuReCo initiative aims to provide a large, distributed, virtual corpus that allows users to conduct contrastive studies in comparable corpora across language boundaries. For this purpose, KorAP offers a search and analysis platform that provides uniform access to all data. The Polish National Corpus is a recent addition to EuReCo, which previously comprised the German Reference Corpus, the Reference Corpus of Contemporary Romanian Language and the Hungarian National Corpus. Initial pilot studies now allow contrastive research across four languages.

EURECO

- **combine existing national and reference corpora**
  - maintained by sustainable institutions
  - instead of building new ones
- provide predefined virtual (comparable) corpora
- enable users to define **specific virtual comparable corpus pairs** dynamically

BENEFITS

- more **economical**
- **scalable**
- **sustainable**
- providing **high linguistic quality** and **high comparability**



PILOT STUDY

Polish

- Latest addition to EuReCo: **Polish** (2023)
- Pilot Study: **Identification of Light Verb Constructions using Collocation Analysis** (involving light verbs and their prepositional/nominal complements; Taborek 2020)
- Extends the analyses of **German**, **Romanian** and **Hungarian** (Kupietz and Trawiński 2022)

```
01 new("KorAPConnection",
02     KorAPUrl = "https://korap.ids-mannheim.de/instance/nkjp1m-sgjp") %>%
03 collocationAnalysis(
04   'focus({[nkjp/l="da(wa)?ć"] []{,5}} [ud/p=NOUN])',
05   leftContextSize = 0,
06   rightContextSize = 1,
07   # relative to { ... } in focus()
08   addExamples = TRUE
09 )
```

Collocation analysis of da(wa)?ć (=give) + NOUN in NKJP1M-SGJP using UDPipe2 annotations (Straka 2018) and RKorAPClient

Collocate	EN	Example	logDice	pmi
radę	-	Nie trza, damy se sami radę!	9.03	8.80
spokój	peace	- Daj spokój, ja jestem	8.71	7.59
znak	sign	rzuciła Gun, dając jednocześnie znak Przystupie, by	7.79	7.14
szansę	chance	" każdemu dawała szansę.	7.78	7.03
wygraną	win	producenci nie dają za wygraną, dotrzymując kroku	7.64	9.59
wyraz	expression	postępować aby dawać temu wyraz.	7.36	7.20

https://korap.ids-mannheim.de/instance/nkjp1m-sgjp

- Using the **RKorAPClient** (Kupietz/Diewald/Margaretha 2020), an R-library for direct access to the KorAP API, it is possible to query all federated corpora in a reproducible way

LIGHT VERB CONSTRUCTIONS IN EURECO

Hungarian

LVC Example	EN	logDice	pmi	ll
nyilvánosságra hozott	disclosed	12.4	12.5	329,887.5
hozzuk nyilvánosságra	we publish	12.3	12.5	257,710.9
hoznak létre	are created	11.5	11.4	224,613.8
helyzetbe hozza	puts you in a position	11.3	11.7	120,263.9
forgalomba hozott	placed on the market	10.0	10.7	37,160.9
rendbe hozni	fix	10.0	11.1	34,710.2
szóba hozták	have been mentioned	10.0	10.3	39,319.4
hozom összefüggésbe	I put it in context	10.0	12.2	25,788.1
összefüggésbe hozza	puts it in context	9.9	12.1	30,667.1
felszínre hoztak	have brought to the surface	9.7	11.3	25,356.7

Collocation analysis for lemma hoz (=bring) with noun in sublative or illative – focus([hnc/p="FN.(SUB|ILL)"] {[hnc/l=hoz]})

Romanian

LVC	EN	logDice	pmi	ll
pune în pericol	danger	11.16	10.97	125,874.70
pune în aplicare	application	10.74	9.24	171,155.67
pune în mișcare	movement	10.63	10.43	82,962.34
pune în discuție	discussion	10.07	10.18	50,411.90
pune în funcțiune	function	9.97	9.94	46,895.08
pune în evidență	emphasis	9.64	8.64	45,253.29
pune în practică	practice	8.95	8.04	24,389.76
pune în executare	execution	8.85	7.82	23,764.97
pune în scenă	scene	8.81	9.18	17,577.59
pune în vânzare	sale	8.51	7.88	15,324.66

Collocation analysis for »pune în NN« (= to put in NN) in CoRoLa (Kupietz/Trawiński 2022).

German

LVC	EN	logDice	pmi	ll
in Szene setzen	scene	10.75	11.93	1,191,593.45
in Gang setzen	aisle	10.68	12.03	890,108.47
in Kenntnis setzen	knowledge	9.79	11.32	363,663.16
in Brand setzen	fire	9.76	10.98	504,157.30
in Bewegung setzen	movement	9.73	10.85	597,460.51
in Verbindung setzen	connection	9.60	10.75	510,847.60
in Kraft setzen	force	8.65	9.63	363,908.00
in Marsch setzen	march	8.40	10.80	82,467.47
in Beziehung setzen	relationship	7.47	8.85	64,474.66
in Anführungszeichen setzen	quotation marks	7.38	11.89	33,600.12

Collocation analysis for »in ... setzen« (= to put in NN) in DeReKo

Contact:
Piotr Bański
Leibniz-Institut für Deutsche Sprache
Dpt. for Digital Linguistics
PO Box 10 16 21
D-68016 Mannheim
Germany

Phone: +49 621 1581-612
banski@ids-mannheim.de

Street Address:
Leibniz-Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
Germany

Phone: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Leibniz Association

REFERENCES

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2020): **RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP**. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 7015–7021.

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. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain (Corpora and language in use), pp. 257–273.

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

Przepiórkowski, Adam/Bańko, Mirosław/Górski, Rafał L./Lewandowska-Tomaszczyk, Barbara (eds) (2012): **Narodowy Korpus Języka Polskiego**. Warsaw: Wydawnictwo Naukowe PWN.

Straka, Milan (2018): **UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task**. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium: Association for Computational Linguistics, pp. 197–207.

Taborek, Janusz (2020): **Kookkurrenz und syntagmatische Muster der Funktionsverbgefüge aus kontrastiver deutsch-polnischer Sicht am Beispiel in Not geraten**. In: De Knop, Sabine/Hermann, Manon (eds.): Funktionsverbgefüge im Fokus: Theoretische, didaktische und kontrastive Perspektiven, Berlin:de Gruyter, pp. 211–233.