

From Solitary Corpus Analysis to Collective Insight: a Glimpse into Translation and Lexicography

Łukasz Grabowski (University of Opole)

Slovko 2023, Bratislava (18 October 2023)

Plan of my talk

Corpus linguists' work: towards expanding and scaling up corpus analyses

- Interdisciplinarity & collaboration

Case Study 1: Exploring repetition in translation

- Translation of reporting verbs from English into Polish and Italian

Case Study 2: Exploring English and Polish lexicographers' habits

- Which multi-word expressions get recorded in dictionaries?

Corpus linguistics and corpus linguists in the 2020s and beyond

- Reflecting upon and reimagining the corpus linguist's role in a collaborative landscape

Typical corpus linguists' work

- **Corpus linguists are really just linguists who have decided to research language using corpora** -> their research interests often align with those of other linguists (Ihrmark & Tyrkkö 2023)
 - Some of them **use corpora to address theoretical questions**,
 - Some **turn to corpora to collect the data they need to produce an accurate description of a linguistic phenomenon**
 - Others are preoccupied with **how corpora should be compiled and structured to be representative of a population of language speakers** and useful to a group of corpus users (Divjak et al. 2017)
- For most corpus linguists, corpus data analysis (text analysis) is part and parcel of their work
 - **a corpus linguist seems to be a bit of a data scientist**

Is corpus linguists a data scientist? (Ihrmark & Tyrkkö 2023)

Differences in how corpus linguists and data scientists view language

- Data scientists are often doing something via text analysis (e.g. identify trends, sentiment analysis, fraud detection), while **linguists are interested in learning more about language itself**

Differences in the type of training they typically receive

- Corpus linguists rarely receive explicit training in programming or statistics
 - although basic statistical concepts are often included to some degree
- Corpus linguists are often trained in the use of specific corpus tools to carry out their queries
 - only some in the field engage with the things that happen “under the hood” (Mason 2018)
- Data scientists are often trained in programming and advanced statistics
 - rarely introduced to theoretical details about language

Towards expanding corpus analyses

- **Simple descriptive statistics or concordance analysis is often not enough** to explain more fine-grained patterns of language use in massive collections of linguistic data
 - if we put forward explanatory hypotheses rather than descriptive ones
 - if we deal with large swathes of linguistic data
- **The range of statistical methods of corpus analysis as well as data visualization techniques have expanded considerably in recent years**
- **Corpus analysis can be also scaled up by incorporating computational methods from natural language processing (NLP) (Dunne 2022) and statistics**

Two examples as a case in point

Research studies on:

- **repetition in translation of reporting verbs** (Mastropierro & Grabowski 2023, in progress)
- **lexicographer's habits with respect to recording multi-word expressions in dictionaries** (Maziarz, Grabowski, Piotrowski, Rudnicka & Piasecki 2023)

What skills should corpus linguists possess in the 2020s and beyond?

- **How to succeed (or survive) in the coming era of AI-assisted (corpus) research, datasphere growth and attention economy?**

Case study #1

Repetition in translation

- Mastropiero & Grabowski (2023-2024, in progress)

Repetition in translation

A **corpus-based study** - both descriptive and explanatory - dealing with **literary translation, language use, and statistics**

How are reporting verbs translated in literary novels and why they are translated the way they are

- RQ1: What are the **linguistic factors that can predict the replacement of reporting verbs** in the source-text with different target translations?
- RQ2: **Which of the factors have a significant effect on either the avoidance or reproduction of reporting verbs' repetition** in the English-to-Polish and English-to-Italian translation?

Repetition in translation

- ▶ **“Avoiding repetition** of words or phrases is part of a **set of translation norms found to operate consciously or subconsciously on the translator**” (Ben-Ari 1998: 2).
- ▶ Corpus-based studies of the translation of repeated features in literary texts (Čermáková & Fárová 2010; Čermáková 2015; Mastropierro & Mahlberg 2017, Čermáková & Mahlberg 2018, Mastropierro 2020) show that **translators tend to avoid repetition in favour of variation.**
 - ▶ **Replacing repetition with lexical variety** can have repercussion on the contribution repetition makes in the creation of stylistic effect

Repetition in translation

- ▶ **Translators should restrain themselves from “synonimization”** (i.e. using synonyms in an attempt to avoid repetition), **they should preserve all repetitions found in the original**, which is sometimes difficult given the fact that “the translators are nuts about repetitions” (Kundera 2015: 164)
- ▶ **“The original’s repetitions deserve respect”** (Bray 2017)
- ▶ English speakers tend to choose lexemes with broad meaning rather than specific while speakers of Russian, Polish and other Slavic languages tend to choose specific lexemes, which is one of the reasons why **they regard texts with multiple repetitions as ones with plain, simple, or even bad style** (Piotrowski 1994: 95-96)
- ▶ Anecdotal evidence: avoiding repetition recommended when teaching writing at Polish primary/secondary schools
 - ▶ Use thesauri, don’t repeat the same words

A study of repetition in translation

- ▶ **Are there linguistic factors that can predict the replacement of a repeated lexical item with different target translations?** (e.g. *said* -> *powiedział, stwierdził, zaznaczył, odnotował, podkreślił ...*)
 - ▶ Item frequency;
 - ▶ Number of different dictionary translations;
 - ▶ Number of different dictionary meanings;
 - ▶ Semantic category.

- ▶ **What type of repeated items are more likely to be translated in different ways?**

Earlier studies of reporting verbs in translation

▶ Why reporting verbs? Stylistic and functional significance!

- ▶ English-Italian (Mastropiero 2020; 2022), English-Spanish (Bourne 2002), English-Persian (Ardekani 2002), Czech-English and Czech-French (Nádvorníková 2020)
- ▶ **Patterns of reporting verbs contribute importantly to characterisation** (Ruano San Segundo 2016, Ruano San Segundo 2017, Eberhardt 2017, Mastropiero 2020).
- ▶ In literary texts, **English reporting verbs are regularly substituted by words with emotive and evaluative meanings** when translated from English into Russian (Urza (2019)
- ▶ In Russian-to-Polish translation of dialogues in literary texts (novels by V. Pelevin), there is **a wide variety of Polish equivalents of the source-language reporting verb *сказать* 'say'** Lubocha-Kruglik & Malysa (2019)

▶ Translation criticism

- ▶ Defending her choices, APT explains that **the Polish language, unlike English, is not able to bear the constant repetition of the form "he said" when recording the dialogue.** (Umiński 2020, 2022: *Three Translators/Trzy Tłumaczki*. Warsaw: Marginesy)
 - ▶ It's about the translation of W. Faulkner's "The Sound and the Fury" into Polish by **A. Przedpełska-Trzeciakowska**

A study of repetition in translation

- ▶ Translation from English into Polish and Italian;
- ▶ 11 + 11 target texts;
 - ▶ **Reporting verbs in 11 English novels and their translations into Polish and Italian**, retrieved via InterCorp (Čermák & Rosen 2012).
- ▶ 5 different authors (Rowling, Brown, Steinbeck, Adams, Tolkien);
- ▶ Multiple translators

Data collection

- ▶ English novels with both Italian and Polish translations;
- ▶ ≥ 500 reporting verbs (tokens) per book (both ST and TT);
- ▶ Parallel English-Italian and English-Polish queries in Intercorp
 - ▶ CQL query for English:

```
[word="\"] [tag="NP"|word="he|she"] [tag="VBD"] | [word="'"] [tag="NP"|word="he|she"] [tag="VBD"] | [word="''"]  
[tag="NP"|word="he|she"] [tag="VBD"] | [word="'''"] [tag="NP"|word="he|she"] [tag="VBD"] | [word="\\"]  
[tag="VBD"] [tag="NP"|word="he|she"] | [word="''"] [tag="VBD"] [tag="NP"|word="he|she"] | [word="''"] [tag="VBD"]  
[tag="NP"|word="he|she"] | [word="''"] [tag="VBD"] [tag="NP"|word="he|she"]
```

- ▶ Corresponding queries for Italian and Polish

Data collection

- ▶ English query paired with equivalent Italian/Polish query to obtain parallel concordance lines.

InterCorp v15 - English <input checked="" type="checkbox"/>		InterCorp v15 - Italian <input checked="" type="checkbox"/>	
Tolkien-Pan_prstenu_1	And I might say it to others, ' he added with a look at the stranger and the miller .	Tolkien-Pan_prstenu_1	E lo dico anche a voi » , aggiunse lanciando uno sguardo al mugnaio ed al forestiero .
Tolkien-Pan_prstenu_1	Well-preserved indeed ! ' he snorted .	Tolkien-Pan_prstenu_1	E poi dicono che mi mantengo bene ! ? » , sbuffò . « Io che mi sento tutto magro , come dire , teso ; rendo l' idea ? Come del burro spalmato su di una fetta di pane troppo grande .
Tolkien-Pan_prstenu_1	' No , it does not seem right , ' he said thoughtfully .	Tolkien-Pan_prstenu_1	« Hai ragione , non può essere normale » , disse pensoso .
Tolkien-Pan_prstenu_1	' Is n't that odd now ? ' he said softly to himself .	Tolkien-Pan_prstenu_1	» , mormorò incantato .
Tolkien-Pan_prstenu_1	' I think , Bilbo , ' he said quietly , ' I should leave it behind .	Tolkien-Pan_prstenu_1	« Credo , Bilbo » , disse pacatamente , « che sarebbe meglio lasciarlo , quest' anello .
Tolkien-Pan_prstenu_1	' You will be a fool if you do , Bilbo , ' he said .	Tolkien-Pan_prstenu_1	« Sei un pazzo se lo fai , Bilbo » , disse ;
Tolkien-Pan_prstenu_1	' It will be my turn to get angry soon , ' he said .	Tolkien-Pan_prstenu_1	« Fra poco sarò io ad arrabbiarmi » , disse . « Guai a te se ripeti una sola volta quel che hai detto !
Tolkien-Pan_prstenu_1	' I do n't know what has come over you , Gandalf , ' he said .	Tolkien-Pan_prstenu_1	« Non capisco che cosa ti succeda , Gandalf » , disse ; « non ti ho mai visto così prima d' oggi .
Tolkien-Pan_prstenu_1	' I have never called you one , ' Gandalf answered .	Tolkien-Pan_prstenu_1	« Non ti ho mai accusato di esserlo » , rispose Gandalf , « e nemmeno io lo sono .
Tolkien-Pan_prstenu_1	' I am sorry , ' he said .	Tolkien-Pan_prstenu_1	« Mi dispiace » , disse , « ma mi sentivo così strano .

Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápvěda

Korpus: InterCorp v15 - English | Dotaz: ,] [tag=,] [word=,] [tag=,] within <te... (3 129 výskytů) ~ Podrobnost

Výskytů: 3 129 | i.p.m.: Spočítat | ARF: 92,07 | Výsledek je setříděn

Výběr řádků: základní

1 / 79

InterCorp v15 - English <input checked="" type="checkbox"/>		InterCorp v15 - Polish <input checked="" type="checkbox"/>	
<input type="checkbox"/>	adams-sbohem_a_dik " Thank you , " he called out , " thank you very ... "	<input type="checkbox"/>	adams-sbohem_a_dik -- Dziękuję ! – krzyknęła postać . – Bardzo wam ...
<input type="checkbox"/>	adams-sbohem_a_dik " It ' s my big number , " she said .	<input type="checkbox"/>	adams-sbohem_a_dik -- To mój hit – powiedziała , zauważając zachwyty Forda .
<input type="checkbox"/>	adams-sbohem_a_dik " Then the planet got demolished , " he shouted .	<input type="checkbox"/>	adams-sbohem_a_dik -- Potem planeta została zniszczona .
<input type="checkbox"/>	adams-sbohem_a_dik " What did she say ? " he asked anxiously .	<input type="checkbox"/>	adams-sbohem_a_dik -- Co ona powiedziała ? – z lękiem spytał Artur .
<input type="checkbox"/>	adams-sbohem_a_dik " Do n't worry about the umbrella , " she said to him as she climbed in .	<input type="checkbox"/>	adams-sbohem_a_dik -- Proszę się nie martwić parasolką – powiedziała wsiadając .
<input type="checkbox"/>	adams-sbohem_a_dik " No , London , " she said .	<input type="checkbox"/>	adams-sbohem_a_dik -- Nie , w Londynie – odparła .
<input type="checkbox"/>	adams-sbohem_a_dik " I think , " she said , with a momentary sigh of relief , " you should buy me a drink before my train goes . "	<input type="checkbox"/>	adams-sbohem_a_dik -- Sądzę – powiedziała Fenny z westchnieniem ulgi – że powinien mi pan przed odjazdem pociągu postawić drinka .
<input type="checkbox"/>	adams-sbohem_a_dik " I do hope you win , " she said with a smile that suddenly snapped together like a piece of advanced origami , " the prizes are so nice . "	<input type="checkbox"/>	adams-sbohem_a_dik -- Mam ogromną nadzieję , że pan wygra – powiedziała z uśmiechem , który rozkwitł jak origami dla zaawansowanych . – Nagrody są takie ładne ...
<input type="checkbox"/>	adams-sbohem_a_dik " No , that ' s all right , really , " he said with a wave of his hand , and she started with an awful deliberation to peel off five tickets , one by one .	<input type="checkbox"/>	adams-sbohem_a_dik -- Nie , proszę dać sobie spokój – powiedział machając ręką , kobieta zaczęła jednak z przerażającą ślamazarnością odrywać pięć losów . Jeden po drugim .
<input type="checkbox"/>	adams-sbohem_a_dik " All right , " she said with a laugh , " you can ask me .	<input type="checkbox"/>	adams-sbohem_a_dik -- No dobrze – powiedziała ze śmiechem .
<input type="checkbox"/>	adams-sbohem_a_dik " I ' ll let you see the prizes in a minute , " she said , and left .	<input type="checkbox"/>	adams-sbohem_a_dik -- Zaraz panu pokazę nagrody – dodała na odchodnym .
<input type="checkbox"/>	adams-sbohem_a_dik " Now we can relax , " she said with a slow smile which filled Arthur till he thought he would burst .	<input type="checkbox"/>	adams-sbohem_a_dik -- No to możemy odetchnąć – rzekła z uśmiechem , który tak wypełnił duszę Artura , iż zaczął się bać , że zaraz pęknie jak balonik .

Raw data

	A	B	C	D	E	F	G	H	I	J	K
1	rowling	" Third time this week I	" he roared	across the table .	rowlingova	- To już trzeci raz w tym tygodniu !	- ryknął	na niego poprzez stół .			
2	rowling	" Do I look stupid ?	" snarled Uncle	Vernon , a bit of fried egg dang	rowlingova	- Czy ja wyglądam na głupca ?	- warknął	wuj Vernon . Z krzaczastego wąsa zwiśa			
3	rowling	" There 's more in the frying	" said Aunt	Petunia , turning misty eyes on	rowlingova	- Jest jeszcze trochę na patelni , syneczku	- odpowiedziała	ciotka Petunia , spoglądając kłkliwe na s			
4	rowling	" Nonsense , Petunia , I nev	" said Uncle	Vernon heartily .	rowlingova	- Ależ to nonsens , Petunio ! Kiedy ja był em w	- oświadczył	stanowczo wuj Vernon .			
5	rowling	" You 've forgotten the mag	" said Harry	irritably .	rowlingova	-hpot tajemn	- Zapomniał	eś magicznego słowa - odpowiedział ze			
6	rowling	" I meant ` please ' !	" said Harry	quickly .	rowlingova	- Chodziło mi o " proszę " !	- powiedział	prędko Harry .			
7	rowling	" HOW DARE YOU THREATE	" roared Uncle	Vernon , pounding the table w	rowlingova	- Jak śmiesz grozić Dudleyowi !	- ryknął	wuj Vernon , waląc pięścią w stół . - ostr			
8	rowling	" All right ,	" said Harry	," all right ... "	rowlingova	- Dobrze , wuju , dobrze	- powiedział	.			
9	rowling	" This could well be the day	" said Uncle	Vernon .	rowlingova	- To może być dzień , w którym dokonam najv	- rzekł	wuj Vernon .			
10	rowling	" I think we should run thro	" said Uncle	Vernon .	rowlingova	- Myślę , że dobrze by było jeszcze raz przejrze	- powiedział	wuj Vernon .			
11	rowling	" In the lounge ,	" said Aunt	Petunia promptly , " waiting to	rowlingova	- W salonie	- odpowiedziała	natychmiast ciotka Petunia - gotowa po			
12	rowling	" They 'll love him !	" cried Aunt	Petunia rapturously .	rowlingova	- Będą nim zachwyceni !	- zawołała	entuzjastycznie ciotka Petunia .			
13	rowling	" Excellent , Dudley ,	" said Uncle	Vernon .	rowlingova	- Znakomicie , Dudley	- pochwalił	go wuj Vernon , po czym zwrócił się do l			
14	rowling	" I 'll be in my bedroom , ma	" said Harry	tonelessly .	rowlingova	- Ja będę siedział cicho w swojej sypialni , udaj	- odrzekł	Harry bezbarwnym tonem .			
15	rowling	" Exactly ,	" said Uncle	Vernon nastily .	rowlingova	- Dokładnie	- powiedział	dobitnie wuj Vernon .			
16	rowling	" I 'll announce dinner ,	" said Aunt	Petunia .	rowlingova	- Oznajmię , że kolacja gotowa	- powiedziała	ciotka Petunia .			
17	rowling	" May I take you through to	" said Dudley	," offering his fat arm to an invi	rowlingova	- Czy mogę panią zaprowadzić do jadalni , pan	- powiedział	Dudley , oferując ramię niewidzialnej kc			
18	rowling	" My perfect little gentleman	" sniffed Aunt	Petunia .	rowlingova	- Mój doskonały mały dżentelmen !	- zagdakala	ciotka Petunia .			
19	rowling	" And you ?	" said Uncle	Vernon viciously to Harry .	rowlingova	- A ty ?	- warknął	wuj Vernon , patrząc na Harry ego .			
20	rowling	" I 'll be in my room , makin	" said Harry	dully .	rowlingova	- Ja będę siedział cicho w swoim pokoju , udaj	- powiedział	tępo Harry .			
21	rowling	" I 'll be in my room , makin	" he said	.	rowlingova	- Ja będę siedział cicho w swoim pokoju i udav	- wyrecytował	.			
22	rowling	" Too right , you will ,	" said Uncle	Vernon forcefully .	rowlingova	- Tak jest i są ku temu powody	- rzekł	dobitnie wuj Vernon .			
23	rowling	" And you ,	" he snarled	at Harry .	rowlingova	- A ty	- warknął	w kierunku Harry'ego - nie pałętaj się po			
24	rowling	" I know what day it is ,	" sang Dudley	," wadding toward him .	rowlingova	- A ja wiem , co dzisiaj jest , aha !	- zaśpiewał	Dudley , zmierzając w jego stronę .			
25	rowling	" What ?	" said Harry	," not taking his eyes off the spc	rowlingova	- Co ?	- zapytał	Harry , nie spuszczać wzroku z miejsca			
26	rowling	" I know what day it is ,	" Dudley repeat,	coming right up to him .	rowlingova	- Wiem , co dzisiaj jest	- powtórzył	Dudley , podchodząc do niego .			
27	rowling	" Well done ,	" said Harry	.	rowlingova	- Brawo !	- powiedział	Harry .			

Data collection

- ▶ Reporting verbs with minimum frequency ≥ 2 ;
- ▶ 23-79 ST reporting verbs, types (average 47.45 per text);
- ▶ 16,742 ST reporting verbs (tokens).
- ▶ Data preparation using custom-designed Python scripts

Methods

- ▶ **What linguistic factors have an effect on the number of different translations a ST item is translated into?**
- ▶ Outcome variable:
 - ▶ **Types** – the number of different TT translations (types) for each ST item.
- ▶ Predictor variables:
 - ▶ **Freq** – the frequency of occurrence of the ST item;
 - ▶ **Trans** – the number of different possible TL translations of the ST item;
 - ▶ **Senses** – the number of different possible meanings of the ST item;
 - ▶ **Verbtype** - the category of reporting verb in Caldas-Coulthard's (1987) taxonomy.

Methods

- ▶ **Trans** – the number of different possible TL translations of the ST item
- ▶ *Treq* (Translation Equivalents Database) (Skrabal & Vavrin 2017)
 - ▶ Lemmas;
 - ▶ Only verbs translations;
 - ▶ Proportion $\geq 4\%$

Treq | TRANSLATION
EQUIVALENTS
DATABASE

Source language: English | Target language: Italian | Restrict to: Collection(s): 7

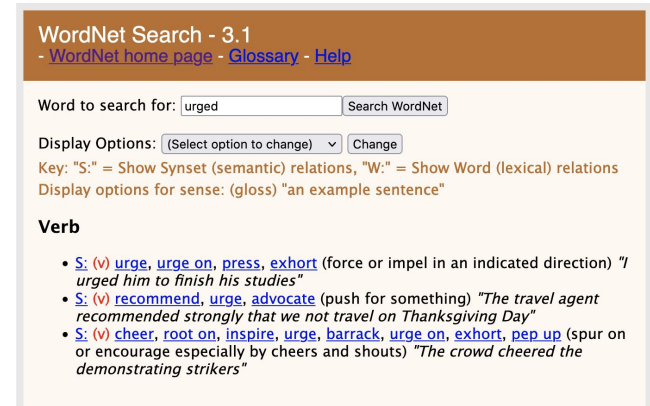
interject Search

Lemma Multiword RegEx A = a

▲ Frequency ▼	▲ Proportion ▼	▲ English ▼	▲ Italian ▼
10	35.7	interject	intervenire
8	26.6	interject	interrompere
2	7.1	interject	ribatté
2	7.1	interject	interloquire
1	3.6	interject	introdurre
1	3.6	interject	intromettere
1	3.6	interject	precisare
1	3.6	interject	domandare
1	3.6	interject	spiegare
1	3.6	interject	intercedere
28			

Methods

- ▶ **Senses** – the number of different possible meanings of the ST item.
- ▶ *Wordnet* (Lexical Database of English):
 - ▶ Number of different senses reported.



WordNet Search - 3.1
- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Verb

- **S:** (v) [urge](#), [urge on](#), [press](#), [exhort](#) (force or impel in an indicated direction) *"I urged him to finish his studies"*
- **S:** (v) [recommend](#), [urge](#), [advocate](#) (push for something) *"The travel agent recommended strongly that we not travel on Thanksgiving Day"*
- **S:** (v) [cheer](#), [root on](#), [inspire](#), [urge](#), [barrack](#), [urge on](#), [exhort](#), [pep up](#) (spur on or encourage especially by cheers and shouts) *"The crowd cheered the demonstrating strikers"*

Methods

- ▶ **Verbtype** - the category of reporting verb in Caldas-Coulthard's (1987) taxonomy.

- ▶ Caldas-Coulthard's (1987) taxonomy:
 - ▶ Seven main categories;
 - ▶ Both linguistic & paralinguistic reporting verbs;
 - ▶ Functional as well as semantic distinction.

Category	Sub-category	Examples
Neutral		<i>say, tell</i>
Structuring		<i>ask, inquire, reply, answer</i>
Metapositional	Assertive	<i>exclaim, proclaim, agree</i>
	Directive	<i>urge, instruct, order</i>
	Expressive	<i>accuse, lament, swear</i>
Metalinguistic		<i>narrate, quote, recount</i>
Prosodic		<i>cry, shout, scream</i>
Paralinguistic	Voice qualifier	<i>whisper, murmur, mutter</i>
	Voice qualification	<i>laugh, sigh, groan</i>
Signalling discourse		<i>repeat, add, go on, hesitate</i>

Methods

- ▶ **Generalised linear model with negative binomial regression:** overdispersion in our data (variance > mean); outcome variable is a count variable
- ▶ Two analyses: **combined Italian data** and **combined Polish data** (some novels with less than 15/20 observations per predictor, cf. Hair et al. 2009: 176);
- ▶ Preliminary analyses to find **the best-fitting model** (lowest AIC).
 - ▶ Tailor-made R scripts

Results: Italian

AIC: 1469.2

- ▶ Overall (left): **Freq**, **Verbtype**, and **Trans** (borderline: $p=0.072$) are significant predictors of **Type**; **Senses** doesn't have a significant effect.

	Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
NULL			462		1659.25		
verbtype	7	1092.27	455		566.98	< 2e-16	***
logfreq	1	320.12	454		246.86	< 2e-16	***
trans	1	3.22	453		243.64	0.07262	.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.02452	0.10289	0.238	0.8117
verbtypeN	0.30647	0.16639	1.842	0.0655 .
verbtypePros	-0.17199	0.11635	-1.478	0.1393
verbtypeSdis	-0.22725	0.11531	-1.971	0.0488 *
verbtypeStr	-0.17462	0.15643	-1.116	0.2643
verbtypeVier	-0.15484	0.11431	-1.355	0.1755
verbtypeVion	-0.06859	0.08700	-0.788	0.4305
verbtypeMlin	-0.77609	1.03921	-0.747	0.4552
logfreq	<u>0.48366</u>	0.02844	17.005	<2e-16 ***
trans	0.05505	0.03059	1.800	0.0719 .

- ▶ How often a reporting verb occurs, what type of verb it is, and the number of available translation equivalents influence the chances of seeing that verb translated into multiple TL items (positive association).
 - ▶ e.g. (right) for a one-unit change (increase) in **logfreq**, *types* increase $e^{0.48}$ -fold
 - ▶ a verb (e.g. *screamed*) has its verb frequency logfreq value of 2. If we increase the frequency by one unit (to a logfreq value of 3), we would expect that the number of translation equivalents (*types*) increases by approximately 48% to its previous value

Results: Polish

AIC: 1575.5

- ▶ Overall (left): **Freq**, **Verbtype**, **Senses**, and interaction **Trans:Senses** are significant predictors of **Type**; **Trans** doesn't have a significant effect.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			451	4697.9	
verbtype	7	3651.0	444	1046.9	< 2e-16 ***
logfreq	1	832.5	443	214.5	< 2e-16 ***
trans	1	2.4	442	212.1	0.12230
senses	1	5.0	441	207.1	0.02538 *
trans:senses	1	6.5	440	200.5	0.01054 *

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.255314	0.149713	-1.705	0.088129 .
verbtypeN	0.451657	0.134111	3.368	0.000758 ***
verbtypePros	-0.158915	0.100394	-1.583	0.113440
verbtypeSdis	-0.250851	0.104430	-2.402	0.016301 *
verbtypeStr	-0.412646	0.129258	-3.192	0.001411 **
verbtypeVier	-0.195429	0.105511	-1.852	0.063996 .
verbtypeVion	-0.119487	0.084124	-1.420	0.155502
verbtypeMlin	0.181544	0.597759	0.304	0.761350
logfreq	0.592083	0.023207	25.513	< 2e-16 ***
trans	0.098669	0.031383	3.144	0.001666 **
senses	0.029276	0.017031	1.719	0.085623 .
trans:senses	-0.008774	0.003374	-2.600	0.009310 **

- ▶ How often a reporting verb occurs, what type of a verb it is, and the interaction between the number of senses and translation equivalents influence the chances of seeing that verb translated into multiple TL items
 - ▶ e.g. (right) for a one-unit change (increase) in **logfreq**, *types* increase $e^{0.59}$ -fold
 - ▶ a verb (e.g. *screamed*) has its verb frequency logfreq value of 2. If we increase the frequency by one unit (to a logfreq value of 3), we would expect that the number of translation equivalents (*types*) increases $e^{0.59}$ -fold

Results: Italian (without *said*)

AIC: 1629.2

- ▶ Overall (left): **Freq**, **Verbtype**, and **Trans** are significant predictors of **Type**; **Senses** doesn't have a significant effect.

	Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
NULL			451		480.44		
verbtype	7	48.069	444	432.37	3.452e-08	***	
logfreq	1	193.561	443	238.81	< 2.2e-16	***	
trans	1	4.544	442	234.26	0.03303	*	

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.08956	0.09806	0.913	0.3610
verbtypeN	0.28627	0.18131	1.579	0.1144
verbtypePros	-0.14079	0.10604	-1.328	0.1843
verbtypeSdis	-0.20807	0.10576	-1.967	0.0491 *
verbtypeStr	-0.12397	0.14089	-0.880	0.3789
verbtypeVier	-0.12695	0.10379	-1.223	0.2213
verbtypeVion	-0.07096	0.07966	-0.891	0.3731
verbtypeMlin	-0.80951	1.00278	-0.807	0.4195
logfreq	0.43751	0.03137	13.949	<2e-16 ***
trans	0.05982	0.02782	2.150	0.0315 *

- ▶ The presence or absence of *said* doesn't change the results.

Results: Polish (without *said*)

AIC: 1451.2

- ▶ Overall (left): **Freq**, **Verbtype**, **Senses**, and interaction **Trans:Senses** are significant predictors of **Type**; **Trans** doesn't have a significant effect.

	Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
NULL			440		694.45		
verbtype	7	163.29	433		531.16	< 2e-16	***
logfreq	1	337.79	432		193.37	< 2e-16	***
trans	1	2.41	431		190.96	0.12064	
senses	1	4.89	430		186.07	0.02706	*
trans:senses	1	5.57	429		180.51	0.01830	*

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.167841	0.154303	-1.088	0.27671
verbtypeN	0.349472	0.159390	2.193	0.02834 *
verbtypePros	-0.121462	0.099935	-1.215	0.22421
verbtypeSdis	-0.234794	0.101897	-2.304	0.02121 *
verbtypeStr	-0.310430	0.133016	-2.334	0.01961 *
verbtypeVier	-0.161899	0.103769	-1.560	0.11872
verbtypeVion	-0.120793	0.081902	-1.475	0.14026
verbtypeMlin	0.169165	0.582811	0.290	0.77162
logfreq	0.556015	0.029994	18.537	< 2e-16 ***
trans	0.092077	0.030819	2.988	0.00281 **
senses	0.025857	0.016774	1.541	0.12321
trans:senses	-0.007977	0.003329	-2.397	0.01655 *

- ▶ The presence or absence of *said* doesn't change the results

Results: verb types and number of translations

Italian		Polish	
Verbtype	Estimate	Verbtype	Estimate
Neutral	0.30647	Neutral	0.451657
Metapropositional	0.02452	Metalinguistic	0.181544
Voice qualification	-0.06859	Voice qualification	-0.119487
Voice qualifier	-0.15484	Prosodic	-0.158915
Prosodic	-0.17199	Voice qualifier	-0.195429
Structuring	-0.17462	Signaling discourse	-0.250851
Signaling discourse	-0.22725	Metapropositional	-0.255314
Metalinguistic	-0.77609	Structuring	-0.412646

- ▶ **Neutral** verbs (e.g. *said, told*) more likely to be translated into different TT verbs
- ▶ **Structuring** (e.g. *asked, replied*) and **signalling discourse** (e.g. *added, repeated*) verbs less likely to be translated into different TT verbs

Conclusions - case study # 1

- ▶ For both Italian and Polish data, the positive correlation between **Freq** and **Types** has been confirmed
- ▶ For Polish data, the interaction between **Senses** and **Trans** has a significant effect on **Types**, but **not Trans on its own**
 - ▶ those verbs which are polysemous have many translation equivalents
 - ▶ polysemy (along with homonymy) may be realized differently across languages (Srinivasan & Rabagliati 2014; Zuercher 2019)
- ▶ By learning what can prompt translation of repeated reporting verbs (e.g. *said*) into a wider lexical variety, we can improve translation strategies to deal with repetition, making them more sensitive to the **stylistic effects of the original**

Case study #2

(English and Polish) lexicographers' habits with respect to recording multi-word expressions in dictionaries

- Maziarz, M., Grabowski, Ł., Piotrowski, T., Rudnicka, E., Piasecki, M. (2023). "Lexicalisation of Polish and English word combinations: an empirical study". *Poznan Studies in Contemporary Linguistics*. <https://doi.org/10.1515/psicl-2023-2002>

MWEs (multi-word expressions) vs MWLUs (multi-word lexical units)

In theory, and in practice (lexicography), **it is difficult to determine when MWEs are treated as word combinations and when as vocabulary units**

- ▶ *blood bank, blood pressure, black lung, black coal, early night, cold night, early grave, mass grave*
- ▶ *lew morski* 'sea lion', *ssak morski* 'sea mammal', *dziecko szczęścia* 'whiz-kid/child of fortune', *przyczyna szczęścia* 'the cause of happiness/luck'

Various linguistic criteria are taken into consideration

- ▶ formal-linguistic, frequency-driven/distributional (statistical), and psycholinguistic ones
 - ▶ frequency of distribution in text and speech, (non-)standard morphosyntactic patterns, constituent (non-)substitutability, semantic (non-)compositionality, among others (Zgusta 1971, Bright 1992, Biber et al. 1999, Hunston and Francis 2000, Svensén 2009)

MWEs vs MWLUs

Any idiosyncrasy in those criteria may be a signal that a process of **lexicalisation** has started (Baldwin and Bond 2002, Baldwin and Kim 2010, Ježek 2016: 25-28, Constant et al. 2017: 838).

- ▶ the more idiosyncrasies, the closer to a lexical unit (LU) status a given word combination (MWE) gets

Lexicalisation is a syntax-to-lexicon process of transgression of **a purely compositional word combination (a syntactic unit) into a single semantic or pragmatic unit, exhibiting word-like behaviour** (Ježek 2016: 25-28, Constant et al. 2017: 838)

- ▶ a gradable process best described on a scale
 - ▶ purely compositional word combinations created ad hoc on one end, through collocations, to fixed phrases and idioms on the other end

Problems with lexicalisation

Lexicalisation of MWEs is a scalar and gradable process, yet in lexicography and lexical databases construction (e.g. wordnets) it is a dichotomous (Yes or No) process

- ▶ **clear, operational procedures for including MWEs in dictionaries or lexical databases** are scarce (there are exceptions though)
- ▶ lexicographers rely on their **linguistic intuition, linguistic experience and competence, contemporary and previous sources of information** (dictionaries, books etc.) to decide which MWE should be granted an MWLU status

More problems

Lexicographers are a very special type of native speakers!

- ▶ **They work with tens of thousands of lexical items, and their experience is far wider** than that of a typical native speaker
- ▶ A structure that may appear to be semantically opaque to a “lay” native speaker (e.g. rare or non-modern) may be transparent to a lexicographer,
 - ▶ semantically transparent word combinations are often excluded from the dictionary
- ▶ A lexicographer may include transparent word-combinations because of **publishing demands or pressure of competition**
 - ▶ unlike a linguist, a lexicographer works under enormous pressure of time, and may include some items because there is simply no time
- ▶ **Dictionaries are usually produced by teams**, which often results in uneven treatment of individual items.
 - ▶ one has to treat dictionary data with caution!

Our study (Maziarz et al. 2023)

We explore **specific factors contributing to the classification/treatment of MWEs as MWUs**

We **review a number of lexicality criteria** and select the ones with the highest potential to be operationalised

semantic compositionality (substitutability/paraphrasing, linguistic intuition), measures of collocational strength, word length (attribute and head)

We explore the criteria in a series of experiments conducted using a multivariate statistical method called **Latent Class Analysis (LCA)**

Our study (2)

We study **two samples of Polish and English MWEs**

- ▶ **bigrams**: adjectives and nouns (A+N) or nouns and nouns (N+N),
- ▶ we contrast the MWEs extracted from selected monolingual dictionaries with those formed by ourselves, and verify which of the MWEs should be accorded MWU status
- ▶ DATASET: <https://clarin-pl.eu/dspace/handle/11321/853>

We deconstruct the English and Polish lexicographers' habits in order to better understand how they deal (what criteria guide their decisions?) with registering MWEs in dictionaries

Research material

Two samples of MWEs - lexicalised and non-lexicalised ones (350 MWEs)

- ▶ **The first sample extracted - based on experts' opinion / purposive sampling - from monolingual Polish** (e.g. *butka paryska*, *lew morski*, *dziecko szczęścia*) **and English dictionaries** (e.g. *floor plan*, *flight bag*, *belly laugh*)
 - ▶ (Dunaj 1996: Słownik współczesnego języka polskiego, SWJP; Pearsall 2001: The New Oxford Dictionary of English)
 - ▶ 102 English and 91 Polish MWEs recorded in dictionaries (lexicalised, MWLUs)
- ▶ **The second (control) sample created by us and modelled on those MWEs in the first sample** (np. *twarda butka*, *ssak morski*, *przyczyna szczęścia*, *clear plan*, *garbage bag*, *loud laugh*) with the intention to obtain non-lexicalised items
 - ▶ 76 English and 91 Polish MWEs not recorded in dictionaries (not lexicalised ones)
 - ▶ We verified if they were recorded there (11 English MWEs and 1 Polish MWE were found to be listed in the dictionaries)

Methodology. Procedures and study stages

We operationalised **semantic compositionality** by testing two custom-designed criteria, i.e., **intuition and paraphrase**, and by using statistical methods (selected **measures of collocational strength: LL, PMI and Jaccard**) for checking word order fixedness and word combination specificity

We also checked **how long (in letters) the syntactic head / its complement is** (the measure highly correlated with word frequency (Piantadosi et al. 2011, Strauss et al. 2007), word familiarity and with the degree of polysemy (Grzybek 2014)

- ▶ combination of semantic compositionality with syntactic fixedness

Lexicality was identified with the presence of an MWE in a dictionary

Methodology. Procedures and study stages

Two teams of annotators (one team for Polish MWEs, and one for English ones), with three annotators in each

Intuition was verified by a test question: “Can you derive the meaning of an MWE from the meanings of its component parts and their syntactic relations?”

- ▶ If ‘yes’, the MWE is compositional;
- ▶ if ‘no’, the MWE is non-compositional

Paraphrase was verified by a test question: “Can you paraphrase an MWE using a relative clause built of MWE components and of function words (prepositions and conjunctions) combined using regular syntactic rules?”

Lexicality: for manually created MWEs, **we checked whether they are recorded in the English or Polish dictionaries**

- ▶ English: Lexico.com (OED), Merriam-Webster
- ▶ Polish: WSJP, SJP (PWN), SJP Dor, Dunaj 1996 (SWJP)

Study goal

We want to identify hidden dependencies between the studied criteria and indicators

- ▶ we run Latent Class Analysis

Why?

We want to **predict the presence of MWEs in two classes: recorded and not recorded in dictionaries**

Results (1). Inter-annotator agreement

From all pairs of annotations, we randomly selected the annotations of particular pairs and recorded them in two separate subsets

- ▶ Subsets with paired annotations and used Cohen's kappa κ (Cohen 1960)
 - ▶ Range is the confidence interval at 95% probability

Criterion	Agreement	CI	Annotation quality
Intuition	$\kappa = 0.35$	[0.351, 0.354]	“Fair”
Paraphrase	$\kappa = 0.64$	[0.638, 0.639]	“Substantial”

Table 1. Inter-annotator agreement (Cohen's kappa) for Polish and English MWEs

Given the minimal (intuition) and moderate (paraphrase) inter-annotator agreement, we averaged their decisions: if they replied “YES” and “NO”, then we proposed a new category “I DON'T KNOW”.

Results (2). Collocation measures and word length

For each MWE, we calculated its collocation strength measures: Log Likelihood (LL), Pointwise Mutual Information (PMI), and Jaccard.

- ▶ High values of PMI and Jaccard -> a strong collocation;
- ▶ High values of LL -> a weak collocation with free word order

We used different corpora for Polish and English and analyzed both data sets separately

- ▶ **For English**, we merged some corpora from the NLTK package : The Reuters-21578 benchmark corpus, Project Gutenberg Selections, Brown Corpus, Australian Broadcasting Commission 2006, Twitter Samples (50 million tokens)
- ▶ **For Polish**, we used the KGR10 Corpus with its 4 billion words (Kocoń and Gawor 2019)

All variables

Semantic and syntactic features included **three nominal variables** (intuition, paraphrase, and MWE status), **two ordinal variables** (HWL and AWL) and **three continuous variables** (collocation measures):

Semantic' features:

- Intuition (“YES” - “DON’T KNOW” - “NO”),
- Paraphrase (“YES” - “DON’T KNOW” - “NO”),
- MWE Status (“Dict” / “non-Dict”),

Syntactic' features:

- HWL and AWL (values: “short” vs. “long”, i.e. “1-4 letters” and “5+ letters”, respectively, for English and “1-5”, “6+” for Polish),
- Collocation measures: LL, PMI and Jaccard.

Criteria			Polish	English
Lexicality	MWE status	“ <u>Dict</u> ”	91 [50%]	102 [58%]
		“ <u>non-Dict</u> ”	91 [50%]	76 [42%]
Semantic	Intuition	“YES”	45 [25%]	115 [65%]
		“DON’T KNOW”	86 [47%]	27 [15%]
		“NO”	51 [28%]	36 [20%]
	Paraphrase	“YES”	101 [56%]	111 [63%]
		“DON’T KNOW”	22 [12%]	36 [20%]
		“NO”	59 [32%]	31 [17%]
Syntactic	mean <u>Jacc</u>		0.0240	0.0072
	mean LL		54.7	62.0
	mean PMI		5.22	7.44
	AWL	“short”	73 [40%]	46 [26%]
		“long”	109 [60%]	132 [74%]
	HWL	“short”	63 [35%]	62 [35%]
		“long”	119 [65%]	116 [65%]

Table 2. Statistics for variables related to syntactic and semantic features of MWEs

Statistics for all variables

Word Length (HWL and AWL): For Polish, the median of the phrase head and the attribute is 6; for English it is 5.

Word length was transposed into the ordinal variable:

- ▶ “short” -> MWEs whose head word length (HWL) / attribute word length (AWL) was shorter than the median
- ▶ “long” -> MWEs whose HWL and AWL scores were equal or longer than the median

Latent class analysis (LCA)

LCA is a statistical technique (Lazarsfeld 1950) used to **identify hidden (unseen) groups or classes** in a population **based on observed nominal variables**

- ▶ **nominal variables: lexicality** (Status = Dict or Non-Dict), **intuition** (Yes, No, I don't know) and **paraphrase** (Yes, No, I don't know)

The probability of belonging to a particular class is assigned to each word combination

In practice, we want to **select the model** (from among the ones of various number of hidden classes) **with the lowest AIC and BIC scores** (quality=data matching and model simplicity respectively)

- ▶ it means selecting a model with an optimum number of classes
- ▶ we used R library PoLCA (Linzer and Lewis 2011)

Results (3a). LCA for Polish

3 class model as an optimum one (lowest AIC and BIC values), high p-value of bootstrapped model fits statistics

Class #1 - “compositional MWEs not found in dictionaries”

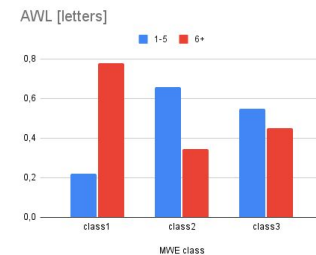
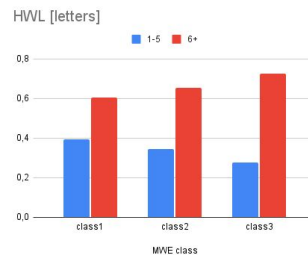
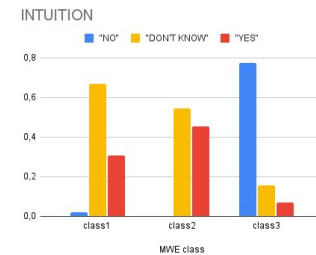
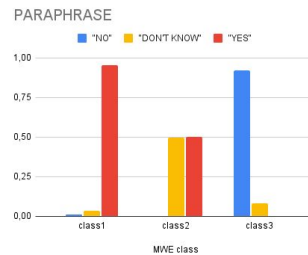
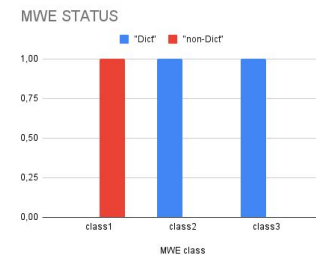
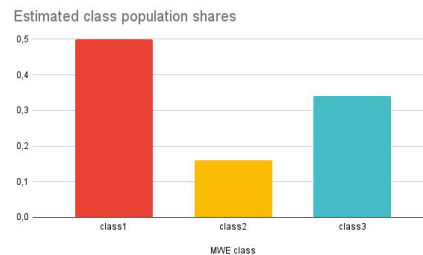
- word combinations out of a dictionary (Status = “non-Dict”),
- paraphrasable (“YES”),
- vague intuition (“DON’T KNOW”, in rare cases “YES”)
- e.g. *czerwony dywan, kozie mleko, twarda bułka*

Class #2 - “(partly) compositional MWEs found in dictionaries”

- word combinations from dictionaries
- either vague intuition of compositionality (“DON’T KNOW”) or strong intuition (“YES”),
- partly/fully paraphrasable (“DON’T KNOW”, “YES”)
- e.g. *gorąca krew, gruba forsa, krótka pamięć*

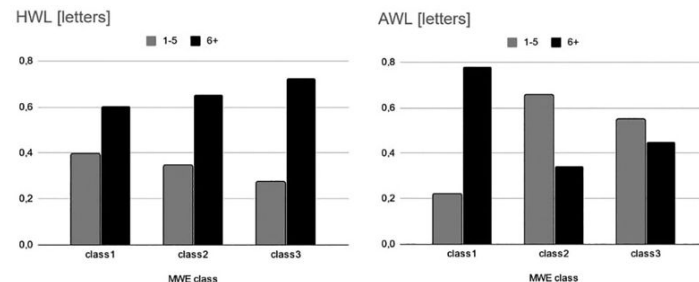
Class #3 - “non-compositional MWEs found in dictionaries”

- word combinations come from dictionaries (Status = “Dict”),
- non-paraphrasable (“NO”),
- clearly non-compositional (Intuition = “NO”)
- e.g. *gabinet cieni, klucz francuski, beczka śmiechu*



Results (3b). Polish - word length

- **Long attributive words** were characteristic for Class #1 (t -test p -values < 0.005), a likely indicator of compositional MWEs
- **Shorter attributive words** were characteristic for Classes #2 and #3: partly and non-compositional MWEs
- Such a relationship did not occur with **syntactic heads** (not statistically significant -> most probably due to our procedure of constructing non-lexicalised MWEs on the basis of syntactic heads of MWEs found in dictionaries)



Results (3c). Polish - collocational measures

LCA with co-variant variables (measures of collocational strength)

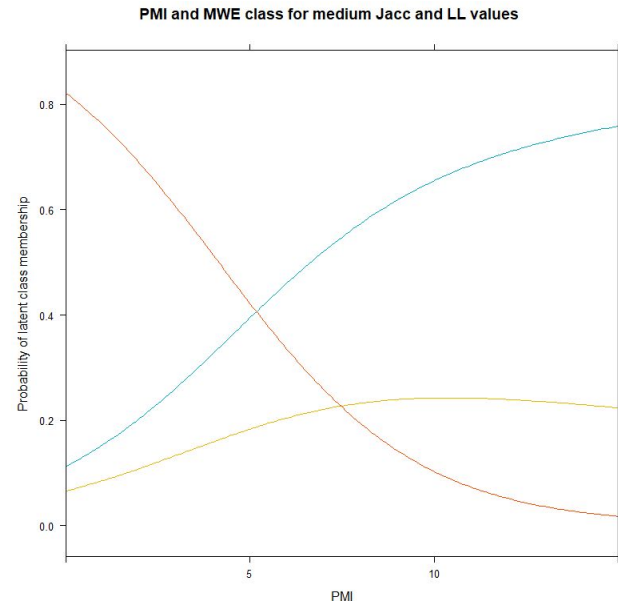
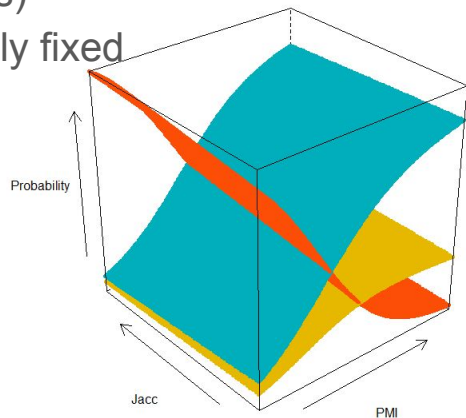
Two classes were differentiated by PMI and Jaccard measures

Class #1 (compositional)

- lower PMI scores = syntactically free word combinations ($p < 0.001$).

Class #3 (non-compositional MWEs)

- higher PMI scores = syntactically fixed



- ▶ non-compositional MWEs are syntactically fixed to a higher degree than compositional ones

— class 1 ("compositional")
— class 2 ("partly compositional")
— class 3 ("non-compositional")

Results (4). LCA for English

3 class model as an optimum one (the lowest AIC and BIC values), high p-value of bootstrapped model fits statistics

Class #1 - “compositional MWEs”

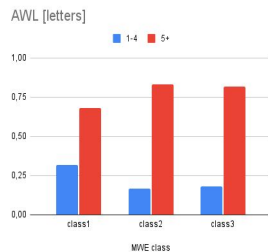
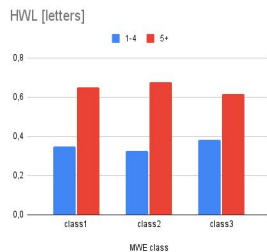
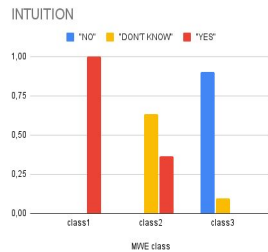
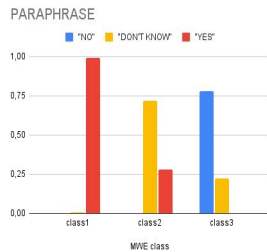
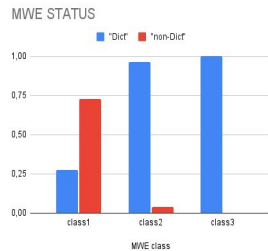
- mostly not recorded in dictionaries (“non-Dict”),
- paraphrasable (“YES”),
- intuitively compositional (“YES”)
- e.g. *electric fire, huge glasses, noisy party, short break*

Class #2 - “partly compositional MWEs found in dictionaries”

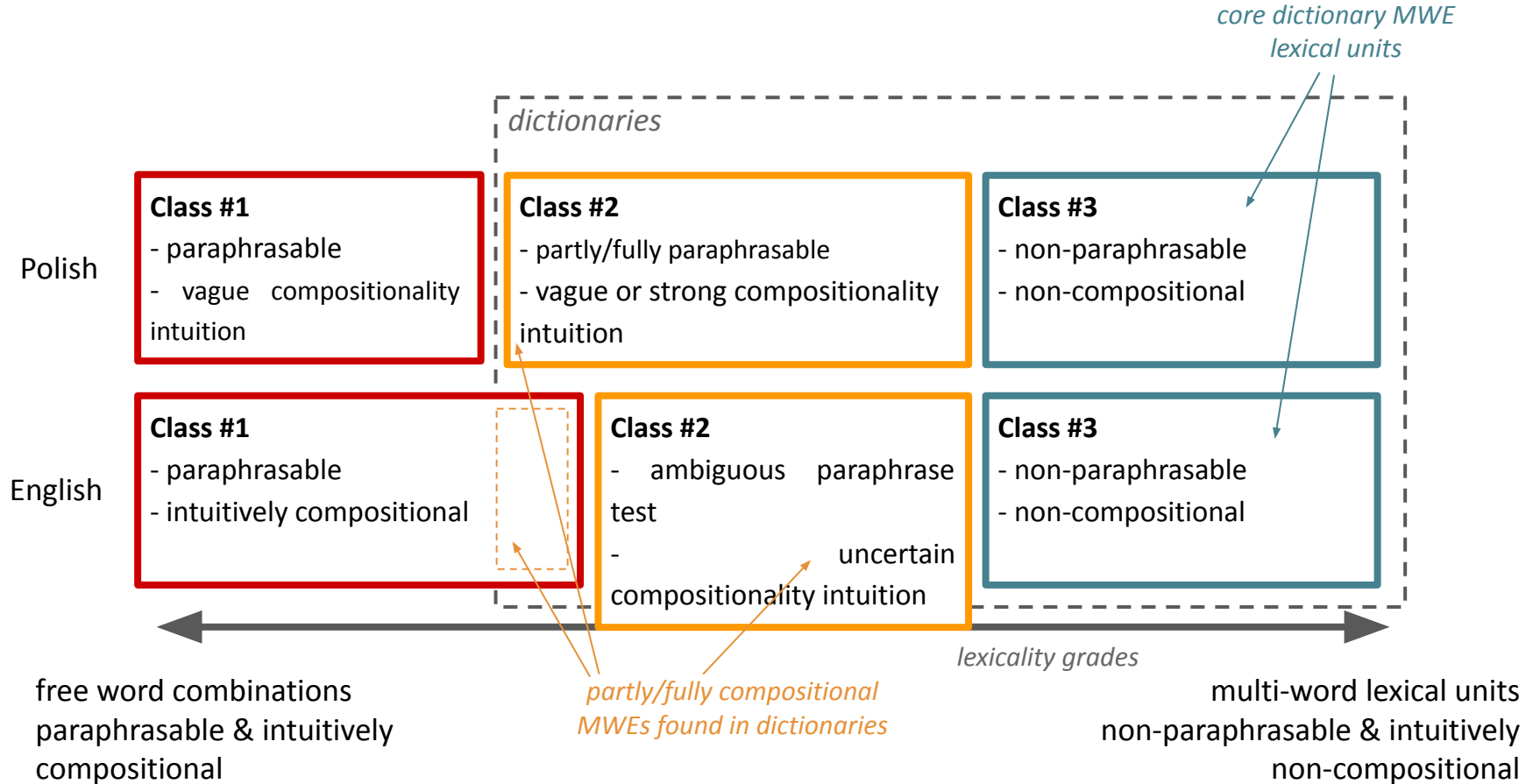
- mostly found in dictionaries (“Dict”)
- ambiguous status in the paraphrase test (“DON’T KNOW” and rarely “YES”)
- uncertain intuition regarding compositionality (“DON’T KNOW” and rarely “YES”)
- e.g. *dirty work, dry rot, first finger, floor exercise, flower bed*

Class #3 - “non-compositional MWEs found in dictionaries”

- found in dictionaries (“Dict”),
- non-paraphrasable (“NO”),
- intuitively non-compositional (“NO”)
- e.g. *blood money, dragon tree, dry run, garden cress, rotten apple*



LCA: comparison of Polish and English MWEs



Discussion

Polish lexicographers don't tend to include compositional MWEs as lexical entries in their dictionaries; the criteria of **paraphrase and intuition** are important for them

- ▶ if MWEs are not clearly and unambiguously paraphrasable and compositional, then they are recorded in dictionaries
 - ▶ approximately $\frac{1}{3}$ of compositional MWEs in Class #1 are recorded in the dictionaries of English (Merriam-Webster and Pearsall 2001); in Class #1 of Polish compositional MWEs not even a single one was recorded in the consulted Polish dictionaries (SJP, SJPDor, WSJP and Wilga)

English lexicographers tend to record also compositional and partly compositional MWEs

- ▶ 94% were recorded in the dictionaries of English

English MWLUs are more heterogeneous than the Polish ones

- ▶ English compositional MWEs covers word combinations not recorded in dictionaries but, unlike the class of Polish compositional MWEs, it also includes ($\frac{1}{3}$) compositional MWEs recorded in dictionaries

Conclusions - case study # 2

We obtained a **collective insight into the degree of semantic compositionality of MWEs and into certain criteria guiding lexicographers' decisions** as to which MWEs should be recorded, as MWUs, in dictionaries of Polish and English

We obtained **further evidence that lexicalization of MWEs is gradable in nature**

The findings are potentially useful for extending existing dictionaries or lexical databases, notably wordnets, with new MWLUs

Corpus linguistics in the 2020s (and beyond)

Corpus linguistics has become increasingly embedded in statistics, data science and NLP

- ▶ research has become increasingly collaborative and interdisciplinary

Corpus linguists benefit a lot from research and development projects

- ▶ e.g. CLARIN-ERIC or national CLARIN infrastructures offering access to computational tools for linguistic data analysis and large textual datasets

Corpus linguistics in the 2020s (and beyond)

Advances in research on application of large language models (LLMs) may bring further changes to the corpus linguists' work

- ▶ potential (not fully known yet) of using LLMs for linguistic data analysis
 - ▶ **integration of LLMs into corpus analysis tools** (Anthony 2023, in progress) as a way to overcome some limitations of corpus analysis tools
 - ▶ **Lew (2023): the quality of COBUILD-style AI-generated (by ChatGPT Plus) lexicographic definitions is comparable to the ones written by human lexicographers**
- ▶ practical difficulties in using LLMs for corpus-based research (Crosthwaite & Baisa 2023)
 - ▶ knowing the data, authenticity, **replicability**, multimodality, safety (data privacy), **hallucinations**

AntConc 4.2.4 (Anthony 2023)

AntConc

KWIC Plot File View Cluster N-Gram Collocate Word Keyword Wordcloud **ChatAI**

Tokens prompt: 655 completion: 38 total: 693 session total: 2158

Target Corpus
Name: AmE06
Files: 500
Tokens: 1017879

- AmE06_A01.txt
- AmE06_A02.txt
- AmE06_A03.txt
- AmE06_A04.txt
- AmE06_A05.txt
- AmE06_A06.txt
- AmE06_A07.txt
- AmE06_A08.txt
- AmE06_A09.txt
- AmE06_A10.txt
- AmE06_A11.txt
- AmE06_A12.txt
- AmE06_A13.txt
- AmE06_A14.txt
- AmE06_A15.txt
- AmE06_A16.txt
- AmE06_A17.txt
- AmE06_A18.txt
- AmE06_A19.txt
- AmE06_A20.txt
- AmE06_A21.txt
- AmE06_A22.txt
- AmE06_A23.txt
- AmE06_A24.txt
- AmE06_A25.txt
- AmE06_A26.txt

ChatAI

How can I help you?

You How is the word "gauge" used in the KWIC results?

AI In the KWIC results, the word "gauge" is used in various contexts such as "gauge the fiscal impact," "gauge the public's will," "gauge the results," "gauge the vitality," "gauge of the country's economic health," "gauge his health," "gauge quantitatively," "gauge whether or not," "gauge transformations," "gauge whether they will be successful," and "gauge with respect to."

You What types of nouns does the word "gauge" collocate with?

AI The word "gauge" collocates with nouns such as "impact," "public's will," "results," "vitality," "character," "economic health," "action," "group," "health," "glass transition temperature," "shotgun," "transformation," "knowledge," and "issue."

Inc. Legacy Max Tokens 50 Temp 1.0

Model Source

Based on the results, is the word "gauge" generally used with positive or negative sentiment?

Progress

Time taken (creating chatai results): 1.6353 sec

Corpus linguists in the 2020s (and beyond?)

A linguist in the first place!

- ▶ The goal of any corpus analysis is to obtain new knowledge of language, be it a language-system or language-in-use

(Some) background in statistics

- ▶ to understand, at least conceptually: descriptive and inferential statistics and methods popular in computational linguistics/NLP; specificity of linguistic data (types of variables etc.), estimation of the quality of a statistical model of linguistic data
- ▶ to understand how to interpret different visualizations of linguistic data

(Some) programming skills sufficient to

- ▶ **understand programming scripts** (in R, Python or what comes next)
- ▶ use **collaborative programming platforms** (Jupyter Notebook, Google Collab etc.)
- ▶ use **visual or low-code programming platforms for linguists**, e.g. KNIME (Ihrmark & Tyrkkö 2023)
 - ▶ <https://pladan.clarin-pl.eu/> (under construction: Pezik, Kaczyński, Grabowski, Jałowiecki, 2023)
- ▶ know how to extract linguistic data or use language tools via API

Collaboration skills

- ▶ working with specialists from other disciplines and/or stakeholders in interdisciplinary research projects (e.g. Baker et al. 2019)
 - ▶ it also facilitates social impact and public engagement



Platforma Danologiczna CLARIN-PL

Biblioteka zastosowań narzędzi, korpusów i modeli CLARIN-PL w warsztacie badawczym

Usecase query...

Search



Wyznaczanie słów kluczowych w tekście

2023-10-08 Łukasz Grabowski

Scenariusz: Wyznaczanie słów kluczowych w tekście

Wersja: 16-08-2023

Zasoby Clarin: SpokesBiz, Morphodita lub WCRFT, Respa

Poziom trudności: średni

Liczba odsłon: 7



Modelowanie tematyczne krótkich tekstów

2023-10-04 Łukasz Grabowski

Scenariusz: Modelowanie tematyczne krótkich tekstów

Wersja: 05-08-2023

Zasoby Clarin: sentencembeddings, Shortextopic

Zasoby spoza Clarin: możliwe do wykorzystania - Spacy, BERTopic

Poziom trudności: średni

Liczba odsłon: 3



Wyszukiwanie nazw własnych w tekstach - Liner

2023-10-03 Konrad Kaczyński, Łukasz Grabowski

Scenariusz: Wyszukiwanie nazw własnych w tekstach - Liner

Wersja: 01-10-2023

Zasoby Clarin: Korpus Dyskursu Parlamentarnego, NER/Liner

Zasoby spoza Clarin:

Poziom trudności: podstawowy

Liczba odsłon: 3



+ Kod + Tekst Skopiuj na Dysk

Połącz T4



Opis:



1. Z korpusu Paralela pobieramy wybrany podkorpus
2. Z pobranych danych generujemy próbkę 1000 par segmentów PL i ENG
3. Liczymy podobieństwo cosinusowe dla par zdań z podkorpusu
4. Generujemy alignment słów/fraz PL i ENG za pomocą wektorów słów/zdań
5. Generujemy słownik zawierający najczęściej występujące tłumaczenia
6. Zapisujemy dane do pliku xlsx



▼ Kroki scenariusza:

▼ Pobierz z korpusu Paralela wybrany zbiór tekstów

```
[ ] # Instalujemy biblioteki, których google collab nie ma preinstalowanych
```

```
!pip install jsonlines  
!pip install progress
```

```
Collecting jsonlines  
  Downloading jsonlines-4.0.0-py3-none-any.whl (8.7 kB)  
Requirement already satisfied: attrs>=19.2.0 in /usr/local/lib/python3.10/dist-packages (from jsonlines) (23.1.0)  
Installing collected packages: jsonlines  
Successfully installed jsonlines-4.0.0  
Collecting progress  
  Downloading progress-1.6.tar.gz (7.8 kB)  
  Preparing metadata (setup.py) ... done  
Building wheels for collected packages: progress  
  Building wheel for progress (setup.py) ... done  
  Created wheel for progress: filename=progress-1.6-py3-none-any.whl size=9610 sha256=f8b19cd1470656ebcaaa59fdc3b974046b749a0aafca9d7432fc5da4dd6b179  
  Stored in directory: /root/.cache/pip/wheels/a2/68/5f/c339b20a41659d856c93ccdce6a33095493eb82c3964aac5a1  
Successfully built progress
```



Conclusions

On the individual level, **corpus analysis is often a solitary work**

- ▶ **working in a multi-disciplinary research team** helps obtain **the collective insight into complex linguistic or social problems** involving language use
 - ▶ cyberhate (Kopytowska et al. 2017), perception of science by Poles (Dzimińska et al. 2023)

Corpus linguists' toolkit and skills have become **a mix of those of data scientists, computer programmers, statisticians and skilled collaborators**

Technological developments bring about social changes that impact the ways we do corpus linguistic research

- ▶ **attention economy** (Hyland 2023): better and better research methods may tip the scales in the race for getting published and becoming visible
 - ▶ in 2022, Elsevier alone received 2.7 million paper submissions (2.5 million in 2021); over 600,000 papers published in 2021
- ▶ **datasphere growth** (statista.com, 2023): data volume will grow from circa 125 ZB in 2023 to more than 180 ZB in 2025 (where 1 ZB equals a billion TB)

Final note

Corpus linguistics is not the sole approach to the study of language ->
rationalistic „armchair“ linguistics continues and is also likely to evolve in the future

- ▶ Its cumulative knowledge has led to contemporary developments
- ▶ It may provide a critical outlook on current developments in corpus linguistics

The period of AI-assisted (corpus) research has already started

- ▶ Corpus linguists should navigate through it with an open-mind and caution!

Thank you!

Q & A

e-mail: [lukasz\(at\)uni.opole.pl](mailto:lukasz(at)uni.opole.pl)

References (selected)

- Crosthwaite, P. & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3). <https://doi.org/10.1016/j.acorp.2023.100066>
- Daniel, I. & Tyrkkö, J. (2023). Learning text analytics without coding? An introduction to KNIME. *Education for Information* 39 (2023) 121–137
- Dunne, J. (2022). *Natural Language Processing for Corpus Linguistics (Elements in Corpus Linguistics)*. Cambridge: Cambridge University Press, 96 p.
- Hirschberg, J. & Manning, Ch. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.
- Hyland, K. (2023). Academic publishing and the attention economy. *Journal of English for Academic Purposes*, 64. <https://doi.org/10.1016/j.jeap.2023.101253>
- Jurafsky, D. & Martin, J. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (Third edition e-book: draft of January 7, 2023). Accessible at <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf> (accessed on 19 July 2023).
- Lew, R. (2023, June 12). ChatGPT as a COBUILD lexicographer. Accessible at <https://doi.org/10.31219/osf.io/t9mbu>
- NCES. (n.d.). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes). In *Statista - The Statistics Portal*. Accessible at <https://www.statista.com/statistics/871513/worldwide-data-created/>
- Winter B. (2019). *Statistics for Linguists: An Introduction Using R*. London: Routledge, 310 p.

Selected references - Case study # 1

- Ben-Ari, N. (1998). The ambivalent case of repetitions in literary translation. *Avoiding repetitions: A 'universal' of translation*. *Meta*, 43(1), 68–78.
- Caldas-Coulthard, C. R. (1987). Reported speech in Written narrative texts. In *Discussing Discourse*, edited by M. Coulthard, 149–67. Birmingham: University of Birmingham.
- Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–27.
- Čermáková, A. (2015). Repetition in John Irving's novel *A Widow for One Year*. A Corpus stylistic approach to literary translation. *International Journal of Corpus Linguistics*, 20(3), 355–77.
- De Sutter, G., & Lefer, M.-A. (2019). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, 28(1), 1–23.
- Mastropierro, L. (2020). The translation of reporting verbs in Italian: The case of the Harry Potter series. *International Journal of Corpus Linguistics*, 25(3): 241–69.
- Nádvorníková, O. (2020). Differences in the lexical variation of reporting verbs in French, English and Czech fiction and their impact on translation. *Languages in Contrast*, 20(2), 209–34.
- Piotrowski, T. (1994). *Problems in bilingual lexicography*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Selected references - Case study # 2

Baldwin, T. and Kim S.N. 2010. "Multiword Expressions". In: Indurkha, N., Damerau, F. J. (Ed.). Handbook of Natural Language Processing (Vol. 2). CRC Press.

Bauer, L. 2019. "Compounds and multi-word expressions in English". In: B. Schlücker (Ed.), Complex Lexical Units: Compounds and Multi-Word Expressions. Berlin: de Gruyter, 45-68.

Corpas Pastor, G. and Colson, J-P. (eds.). 2020. Computational Phraseology. Amsterdam: John Benjamins.

Hanks, P. 2013. Lexical Analysis: Norms and Exploitations. Cambridge, MA: MIT Press.

Hunston, S. and Francis, G. 2000. Pattern Grammar: a corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins.

Moon, R. 1998. Fixed Expressions and Idioms in English. A Corpus-Based Approach. Oxford: Clarendon.

Pęzik, P. 2018. Facets of prefabrication. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Sinclair, J. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

Wray, A. 2002. Formulaic language and the lexicon. Cambridge: CUP.

Zgusta, L. 1971. Manual of Lexicography. Prague: Czechoslovak Academy of Sciences.