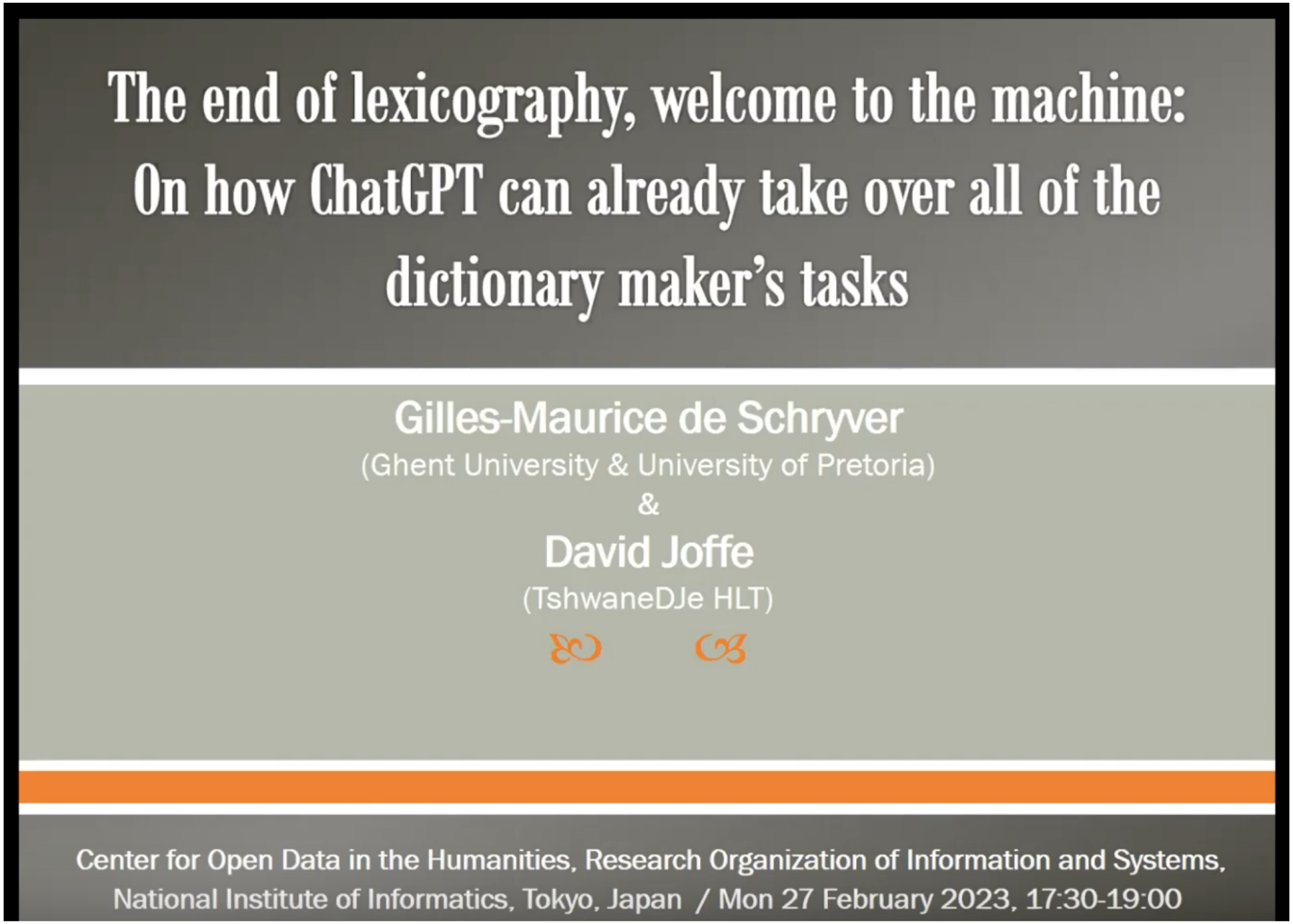
The added value of LLMs for lexicography and for lexical semantics

Erhard Hinrichs

Department of Linguistics
Eberhard-Karls Universität Tübingen



Source: `https://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/`

Three Use Cases

- LLM generated example sentences for word senses in GermaNet (joint work with Reinhild Barkey, Marie Hinrichs, and Claus Zinn)
- Harvesting of neologisms by LLMs (on-going work at Tübingen)
- Semantic Clustering of word sense by collocations (on-going work by Mark Davies at english-corpora and at Tübingen)

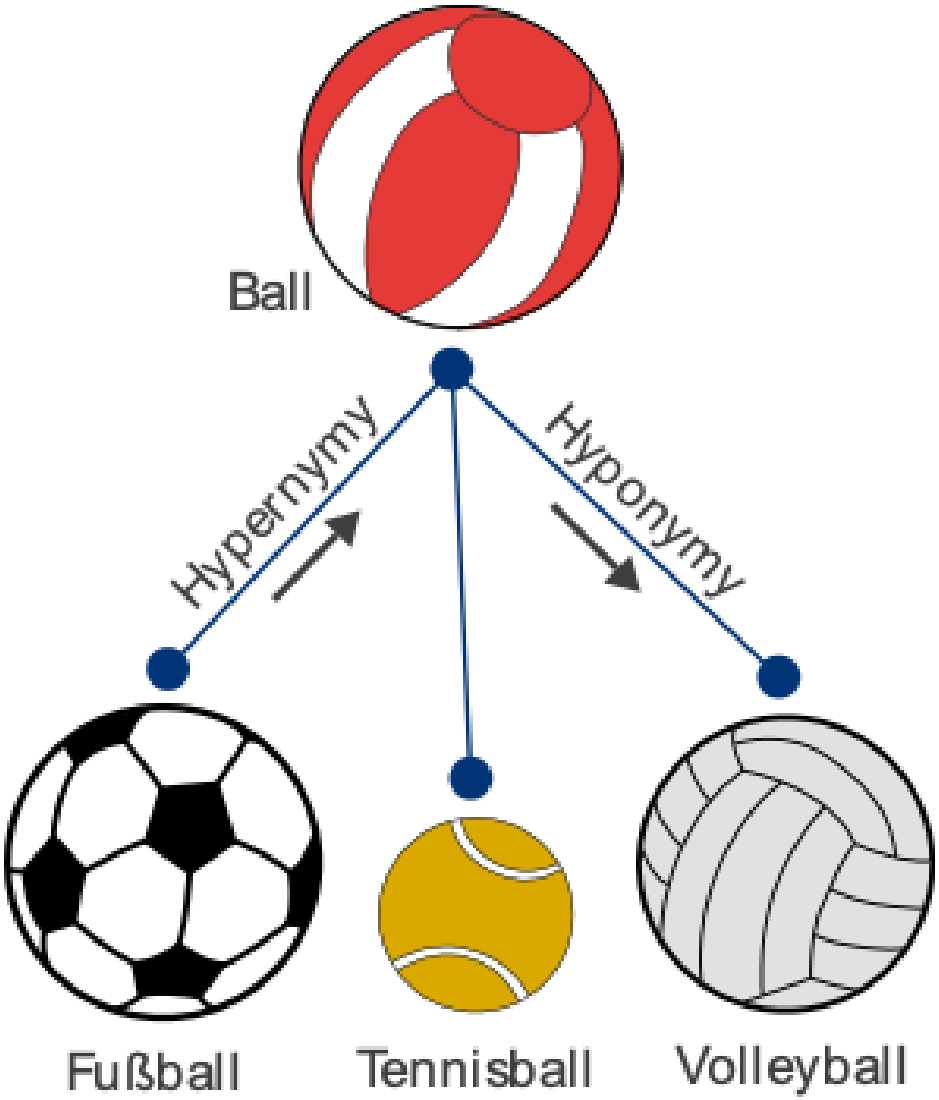
Use Case 1: Can generative AI models provide characteristic examples for word senses without human intervention?

- Using GermaNet, the German counterpart of the Princeton WordNet for English, as our data source
- Using ChatGPT model 4.0, Deep Seek, and Claude as deep in-context platforms
- Task for our pilot study: adding characteristic examples to nouns from different semantic fields represented in GermaNet

GermaNet Release 19.0

- 225,000 lexical units (181,047 nouns, 22,003 verbs, and 21,950 adjectives)
- 174,579 synsets
- 189,330 conceptual relations, and 13,346 lexical relations (synonymy excluded)
- 126,733 compounds (with pointers to their constituent parts)

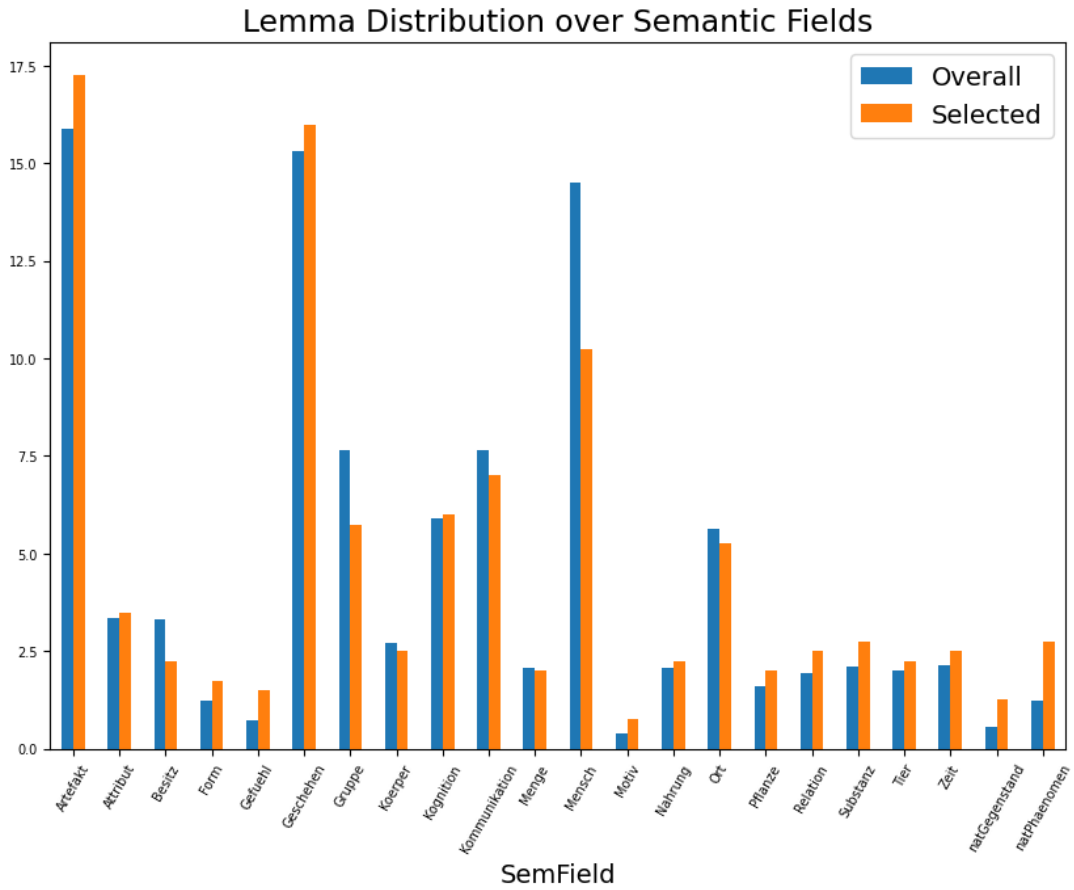
Conceptual relations



Construction of the data sample for the LLM experiments

- 100 word senses for monosemous nouns and 300 word senses for polysemous nouns
- Monosemous lemmas were chosen randomly according to their distribution across the 22 semantic fields in GermaNet.
- Polysemous lemmas were selected in such a way as to reflect as closely as possible the distribution of word senses across the semantic fields and of the number of word senses per lemma in GermaNet

Distribution of monosemous word senses across semantic fields



Task Specification for the LLM

> *It is your task to assist language experts in the creation of dictionary entries. In particular, the goal is to produce example sentences that document the typical usage of a given word as well as possible. These example sentences should be typical, natural, informative, and understandable. The property of an example sentence being typical means that typical contexts, syntactic patterns, and collocations in which a word most frequently occurs are included in the sentence ...*

using the criteria for good examples put forth by Atkins and Rundell (2008)

Criteria for good examples (due to Atkins and Rundell 2008)

typical	sentence contains typical syntactic patterns and collocations of the word in question.
natural	the sentence captures the literal meaning of the word in question.
informative	The sentence illustrates this literal meaning.
understandable	The sentence avoids specialized vocabulary.

Prompt schema for all three LLMs

Example prompt for monosemous nouns:

Engl.: 'Künstler is a hypernym of 'Musiker'. Give three example sentences for the word: 'Musiker'.

Example prompt for a polysemous nouns*:*

Engl.: The word 'Existentialistin' has two interpretations, which can be described by the following two hypernyms: 'philosopher' and 'spiritually oriented person'. Give me three example sentences for each interpretation.

Annotation scheme and results for monosemous nouns

	claude	deepseek	gpt-4o	llm combined
Σ acceptable	439 (441)	518 (518)	491 (493)	1448 (1452)
FA: wrong word sense	4 (8)	3 (7)	4 (9)	11 (24)
FG: incorrect grammar	8 (10)	2 (1)	4 (4)	14 (15)
FL: wrong lemma	46 (42)	20 (18)	36 (29)	102 (89)
FO: incorrect orthography	10 (4)	0 (2)	0 (2)	10 (8)
HA: data hallucination	14 (14)	8 (8)	9 (10)	21 (32)
NI: not informative	2 (2)	3 (3)	3 (3)	8 (8)
NN: not natural	57 (60)	30 (30)	45 (45)	132 (135)
NT: not typical	17 (19)	13 (13)	5 (5)	35 (37)
Σ not-acceptable	158 (159)	79 (82)	106 (107)	343 (348)
Σ total	597 (600)	597 (600)	597 (600)	1791 (1800)

Post-adjudication Inter-annotator agreement for monosemes and polysemes

Error Type	Metric	Value	Lower CI	Upper CI
conflated	Cohen's Kappa	0.72 (0.71)	0.69 (0.64)	0.75 (0.7)
conflated	PABAK	0.75 (0.7)	0.72 (0.67)	0.77 (0.73)
non-conflated	Cohen's Kappa	0.71 (0.64)	0.69 (0.61)	0.74 (0.66)
non-conflated	PABAK	0.69 (0.61)	0.66 (0.57)	0.71 (0.64)

Table: Post-adjudication Inter-annotator agreement for the judgement of example sentences for **polysemes**, raw annotator values in parenthesis.

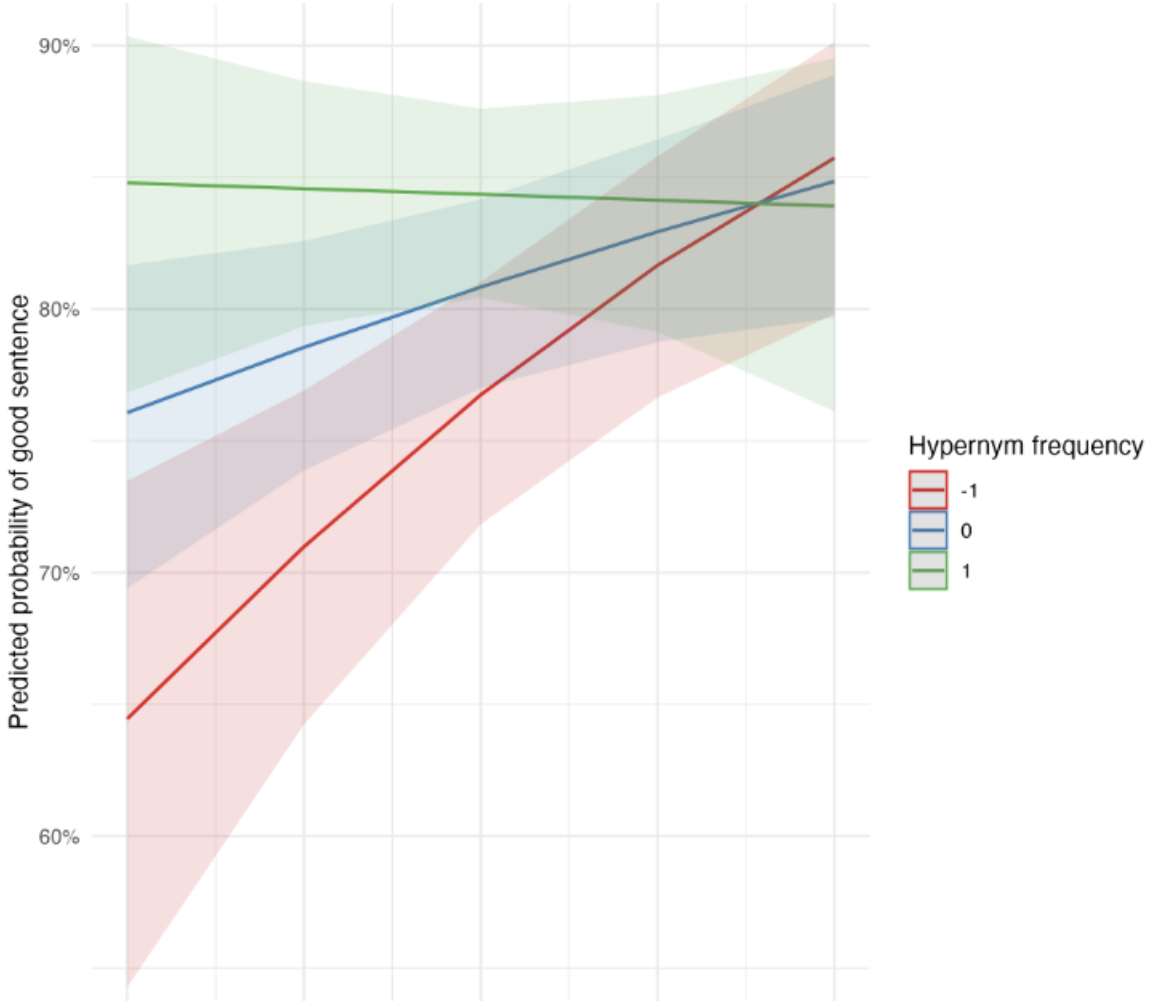
Evaluation of LLM Example Sentences Output

- distinguishes between monosemes and polysemes
- is based on linear (lmer) and generalized (glmer) mixed effects models, using the lme4 package in R
- A mixed model combines mixed effects and random effects in a single statistical model.
- Our fixed effects determine, for instance, how the sentence quality of the LLM-generated models is affected by lemma frequency, lemma's polysemy, or by the LLM which generates the sentence.
- Our random effects include the annotator rating the sentence or the length of the example sentence.

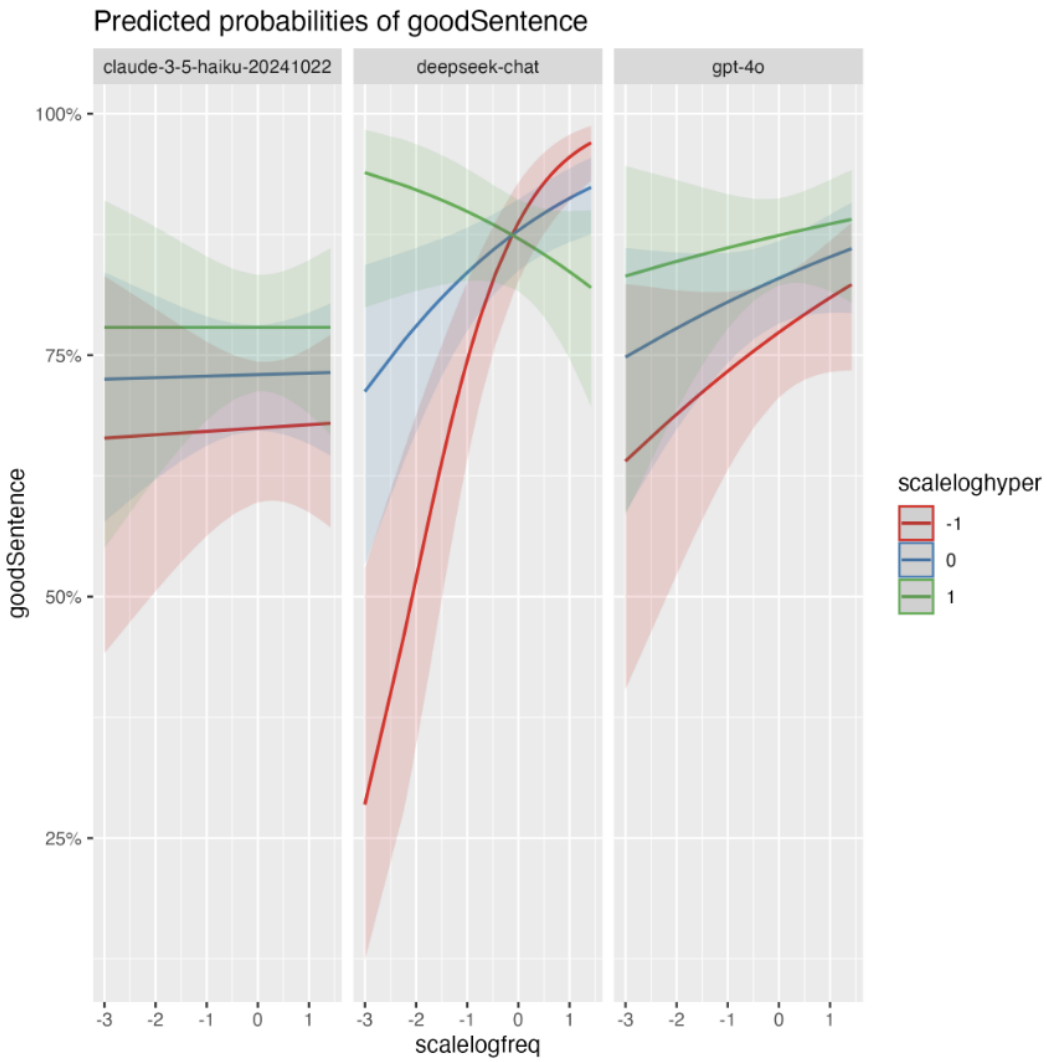
Evaluation for Monosemes

	`claude`	`deepseek`	`gpt-4o`	llm combined
Σ acceptable	439 (441)	518 (518)	491 (493)	1448 (1452)
FA: wrong word sense	4 (8)	3 (7)	4 (9)	11 (24)
FG: incorrect grammar	8 (10)	2 (1)	4 (4)	14 (15)
FL: wrong lemma	46 (42)	20 (18)	36 (29)	102 (89)
FO: incorrect orthography	10 (4)	0 (2)	0 (2)	10 (8)
HA: data hallucination	14 (14)	8 (8)	9 (10)	21 (32)
NI: not informative	2 (2)	3 (3)	3 (3)	8 (8)
NN: not natural	57 (60)	30 (30)	45 (45)	132 (135)
NT: not typical	17 (19)	13 (13)	5 (5)	35 (37)
Σ not-acceptable	158 (159)	79 (82)	106 (107)	343 (348)
Σ total	597 (600)	597 (600)	597 (600)	1791 (1800)

Lemma and hypernym frequencies interaction predicts sentence quality



Evaluation for Monosemes



Estimated Marginal Means by Models – Monosemes

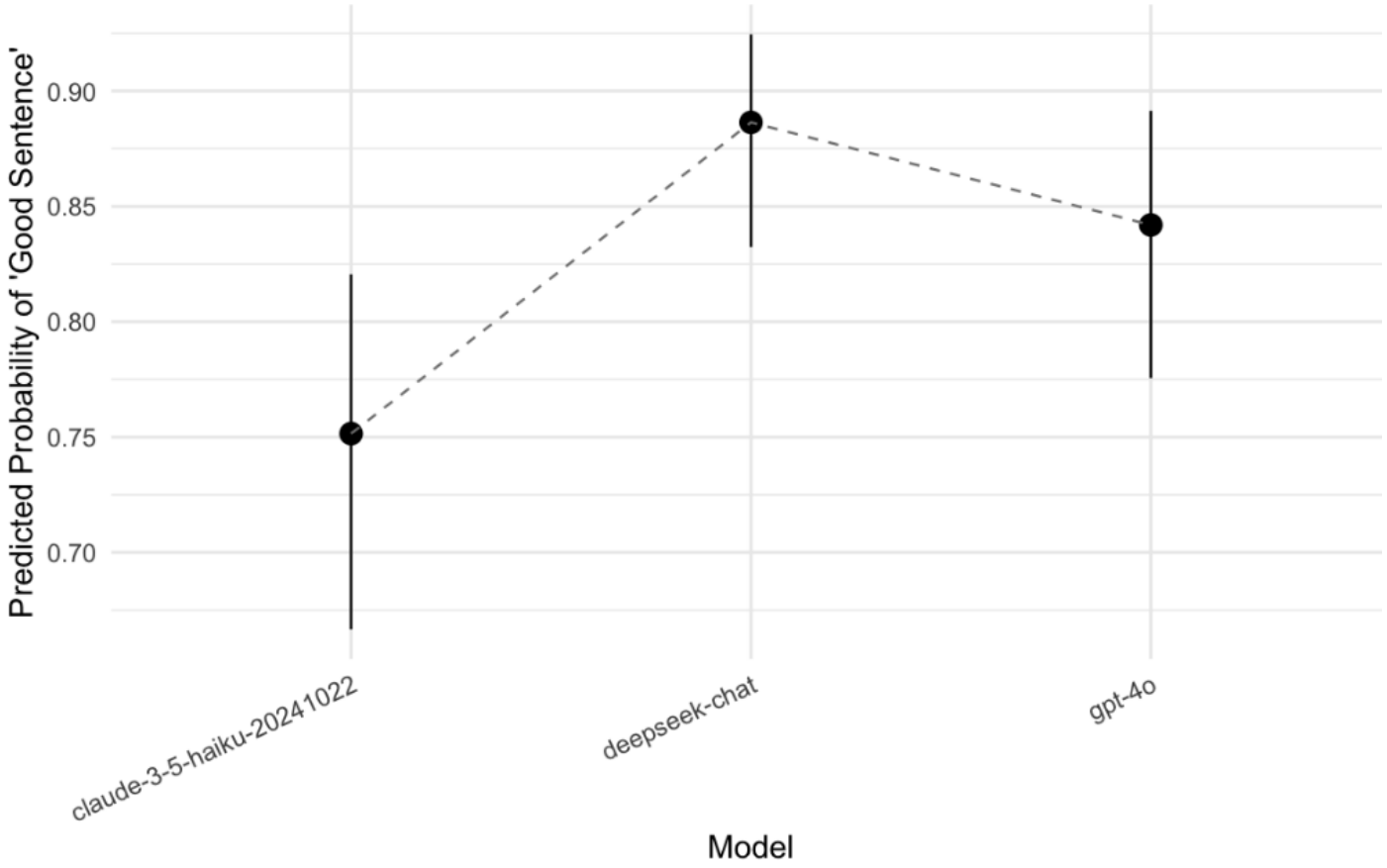


Figure 6: Predicted probability for sentence quality (main effect of model).

Estimated Marginal Means by Models – Polysemes

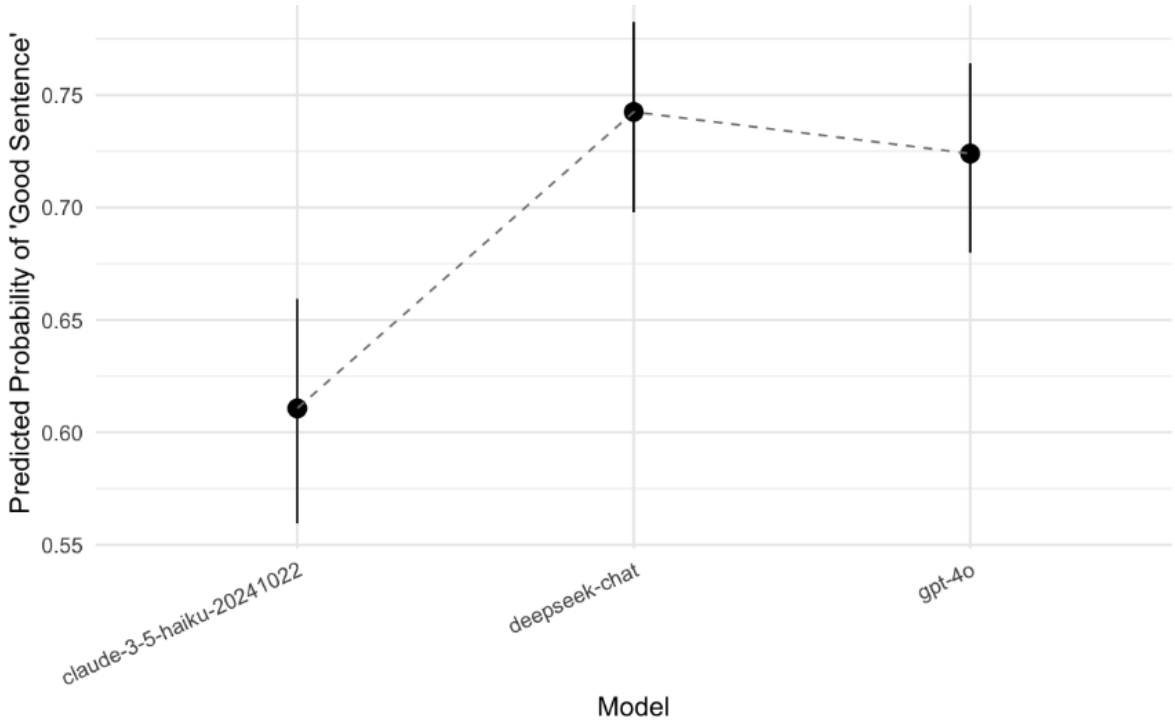


Figure 9: Predicted probability of sentence quality (main effect by model) - polysemes.

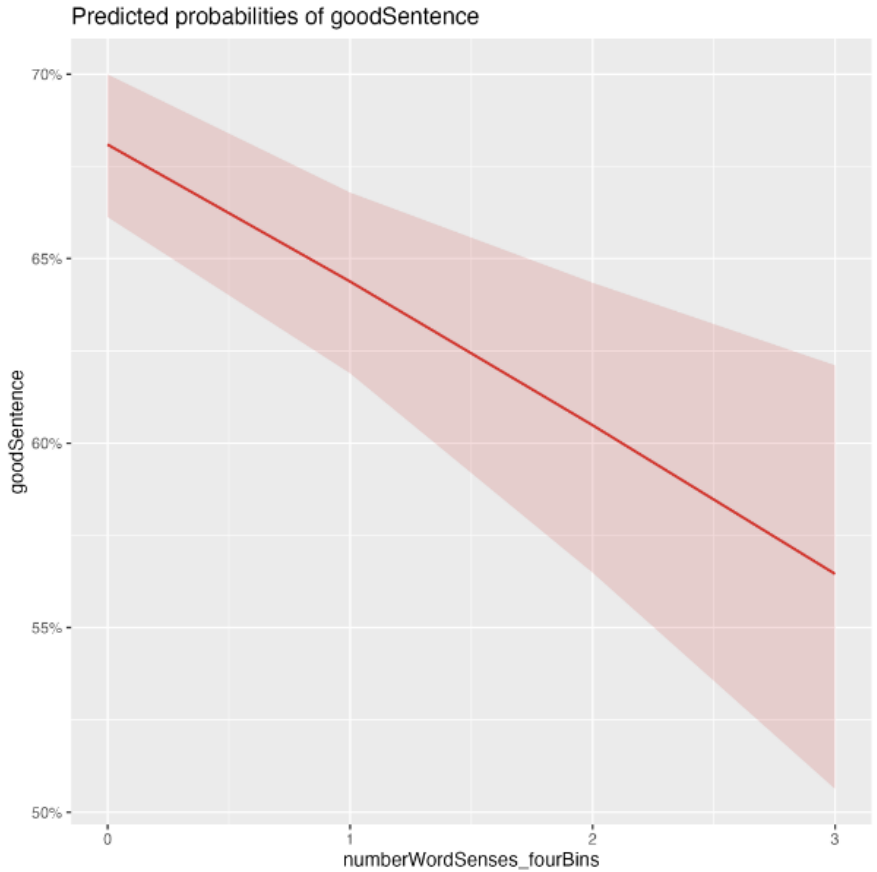
Evaluation for Polysemes

	`claude`	`deepseek`	`gpt-4o`	llm combined
Σ acceptable	1060 (1049)	1252 (1256)	1227 (1220)	3539 (3525)
FA: wrong word sense	256 (258)	205 (194)	221 (220)	682 (672)
FG: incorrect grammar	20 (20)	15 (14)	10 (12)	45 (46)
FL: wrong lemma	247 (252)	142 (144)	130 (132)	519 (528)
FO: incorrect orthography	3 (8)	3 (2)	4 (4)	10 (14)
HA: data hallucination	11 (14)	17 (25)	16 (19)	44 (58)
NI: not informative	24 (24)	20 (23)	22 (23)	66 (70)
NN: not natural	130 (128)	103 (101)	134 (134)	367 (365)
NT: not typical	49 (47)	43 (41)	36 (36)	128 (124)
Σ not-acceptable	740 (751)	548 (544)	573 (580)	1861 (1875)
Σ total	1800 (1800)	1800 (1800)	1800 (1790)	5400 (5400)

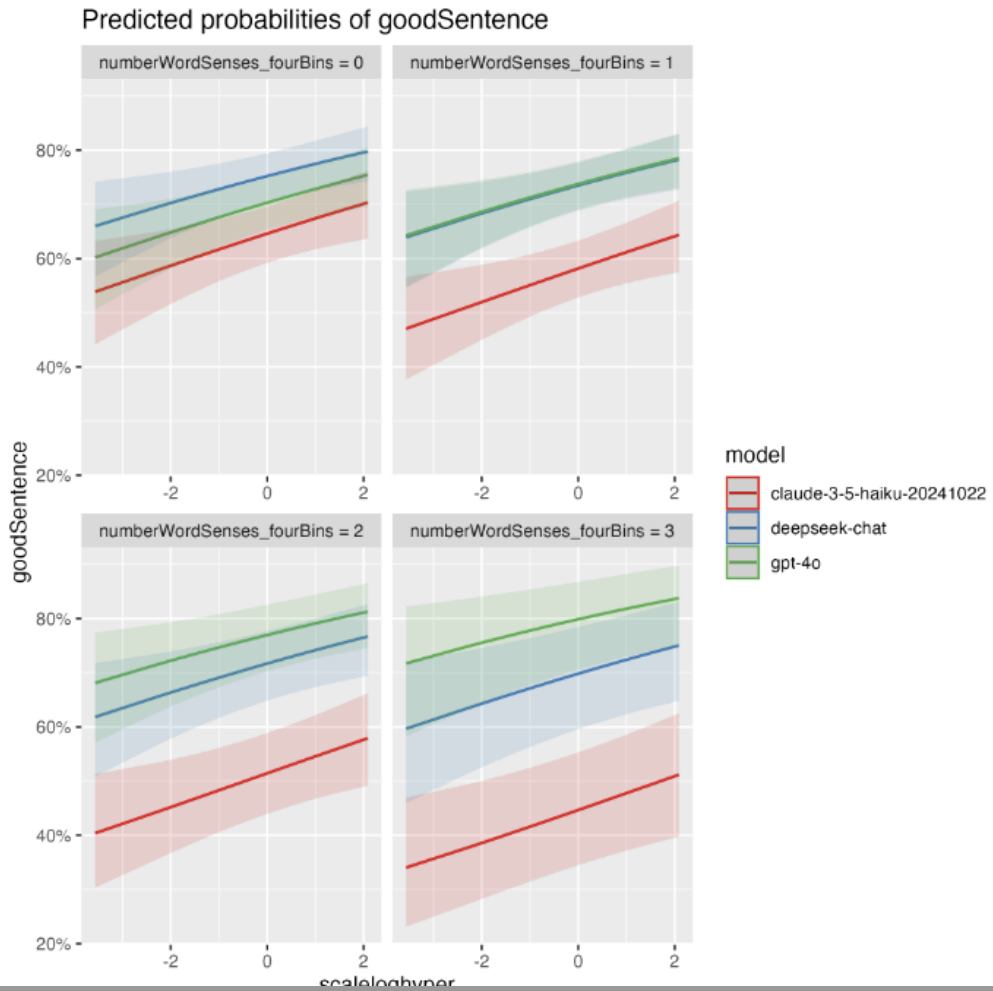
Distribution of Word Senses for Polysemes

lemmas	93	14	9	3	1	1	1	$\sum 6$
word senses	2	3	4	5	6	7	8	$ws \geq 5$
example sentences	3348	756	648	270	108	126	144	$\sum 648$

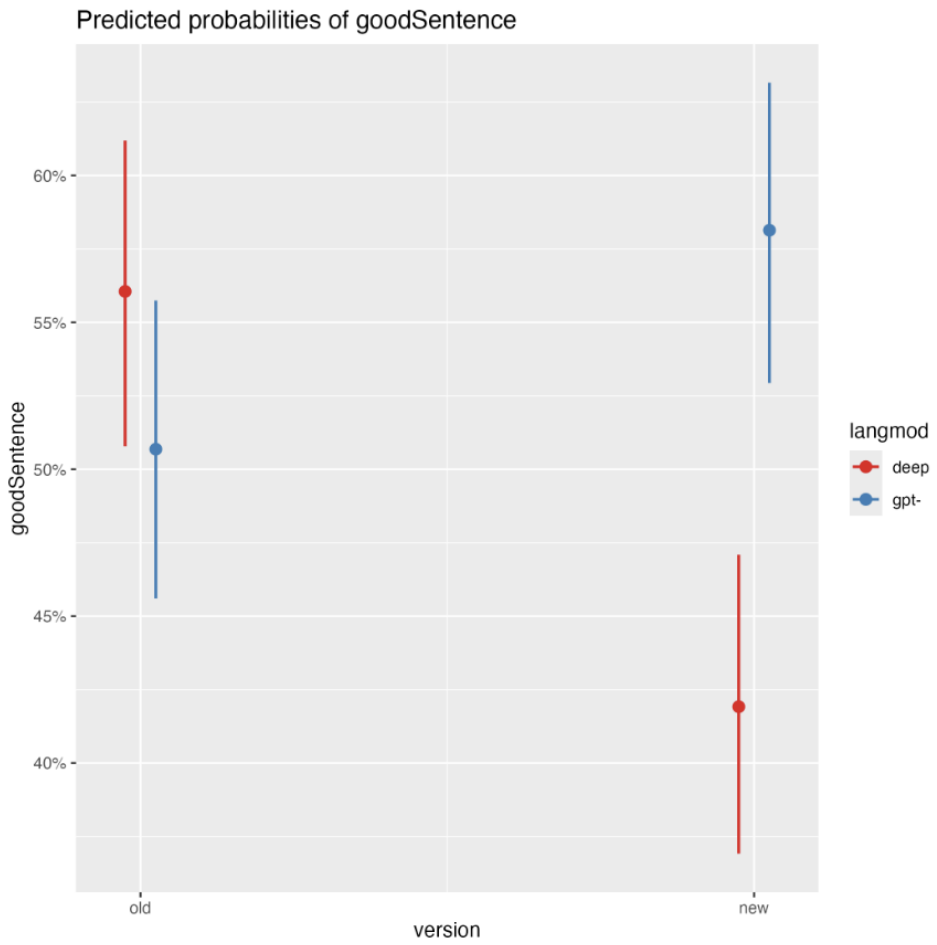
Number of Word Senses as Predictor of Polysemes



Probability of Good Sentences by Model, Polysemy and Hypernym Frequency



Follow-up Study: Improvement of LLM models (Polysemes)



Use Case 2: Can generative AI models generate a list of German neologisms with frequency information?

Prompt:

> *Give me a list of neologisms for German for the period of 2018 - 2025 with frequency information. Please name the sources from which you compiled this list.*

ChatGPT reply:

> *Unfortunately I cannot deliver the desired list of German neologisms for the period of 2018 - 2025 with concrete frequency information. Here is a brief explanation:*

Explanation given by ChatGPT

Even though there are several projects for the documentation of neologisms in German, e.g., the project Neuer Wortschatz at the Institute for the German Language in Mannheim, they are missing a publically accessible table with lemma and frequency information for the period of 2018-2025. (ids-mannheim.de)

Response given by DeepSeek

- Of course. Here is a list of German neologisms from the period of approximately 2018 to 2025, complete with frequency information and the sources used for compilation.
- Frequency Information: Providing exact, scientific frequency data (like a corpus count) is beyond the scope of this format.
- Instead, frequency is indicated qualitatively as:
  - High: Very common in mainstream media, everyday conversation, and social media.
  - Medium: Regularly used in specific contexts (e.g., politics, tech) and understood by the general public.
  - Niche/Establishing: Primarily used within specific communities (e.g., climate activism, digital culture) but gaining recognition.

Response given by DeepSeek

- Sources: The list is compiled from the following authoritative sources, which are the primary references for new words in the German language:
  1. Gesellschaft für deutsche Sprache (GfdS): Publishes the "Wort des Jahres" (Word of the Year) list, which captures significant words reflecting the public discourse of a given year.
  2. Langenscheidt: Publishes the "Jugendwort des Jahres" (Youth Word of the Year), highlighting trends in the language of younger generations.
  3. Duden: Germany's pre-eminent dictionary publisher. Duden's online edition regularly adds new words, marking their official entry into the standard language.
  4. Digital Discourse: Observation of usage in German online media (e.g., Spiegel Online, ZEIT Online), social media platforms (Twitter/X, Instagram), and political commentary.

Use Case 3: Comparing collocates harvested from large corpora and from LLMs (Mark Davies)

- LLM output from seven different LLMs: Chat-GPT, Gemini, Claude (Anthropic), Perplexity, DeepSeek, Llama (Meta), and Mistral
- Corpora available at English-Corpora.org such as the 14 billion-word Intelligent Web-based Corpus (iWeb) corpus or the 1 billion-word Corpus of Contemporary American English (COCA)
- Corpus tools such as Sketch Engine and word sketches offered at english-corpora.org can only produce lists of collocates, but cannot categorize and cluster them.

GPT Summary of Collocates for the English noun dagger (Example due to Mark Davies)

1. Noun Collocates: Words like "blade," "weapon," and "hilt" emphasize the physical attributes and functionality of a dagger, while "scabbard" and "knife" indicate its close relation to other sharp tools or weapons.
2. Adjective Collocates: Descriptors such as "sharp," "deadly," and "poisoned" highlight its lethality, while "ancient," "ceremonial," and "ornate" suggest its use in historical or ritualistic contexts.
3. Verb Collocates (Subject): Actions like "glints," "pierces," and "slashes" suggest dynamic scenes where the dagger is depicted as an active entity, reflecting its role in vivid, often violent imagery.
4. Verb Collocates (Object): Words like "wield," "brandish," and "plunge" show that the dagger is primarily handled by a person, emphasizing its practical use in combat or symbolic gestures.

GPT can infer word meanings from its collocates

Overall Insights:

The collocates paint a vivid picture of the dagger as a dual-purpose item: a functional weapon used for combat and a ceremonial object imbued with historical and cultural significance. It is depicted as both a practical tool and a symbol of danger, power, or tradition.

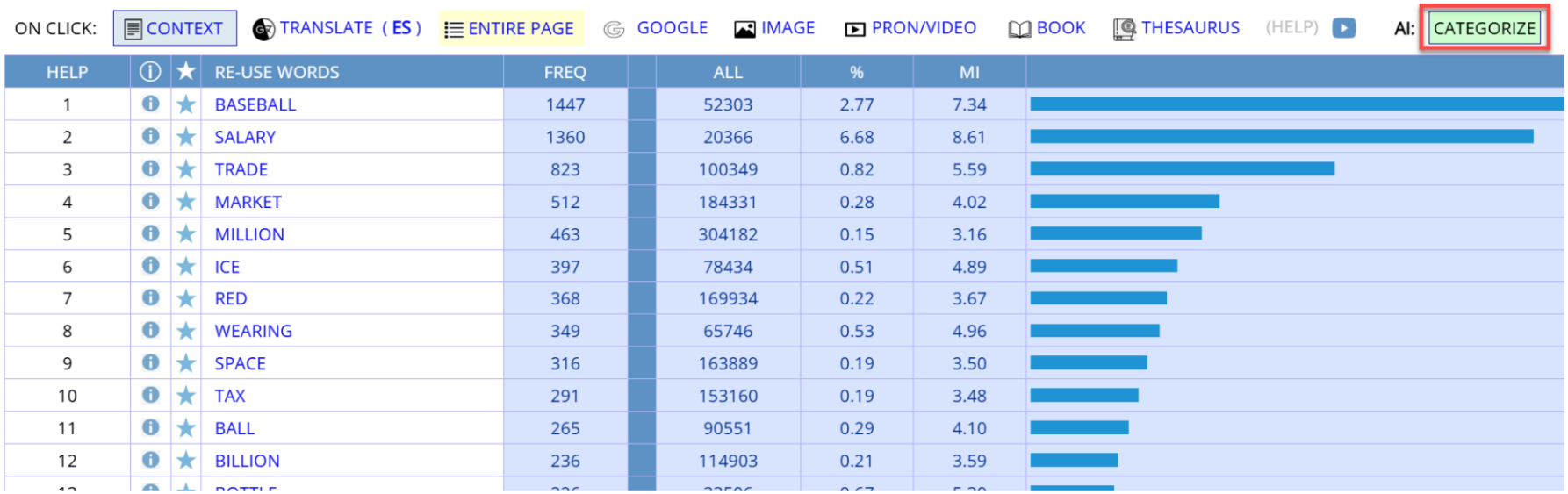
Source: Mark Davies `https://www.english-corpora.org/ai-llms/collocates.pdf`

Categorizing and Classifying Collocates by LLMs

- The richness and quality of the collocates from LLMs suggest that their underlying model generates better connections between words than the standard "collocates" model of corpora, which simply look at nearby words.
- The LLMs usually provide lists of collocates that are more insightful than the corpora, especially for low-frequency words.

Source: Mark Davies `https://www.english-corpora.org/ai-llms/collocates.pdf`

The Corpus View: Collocates for the English noun cap

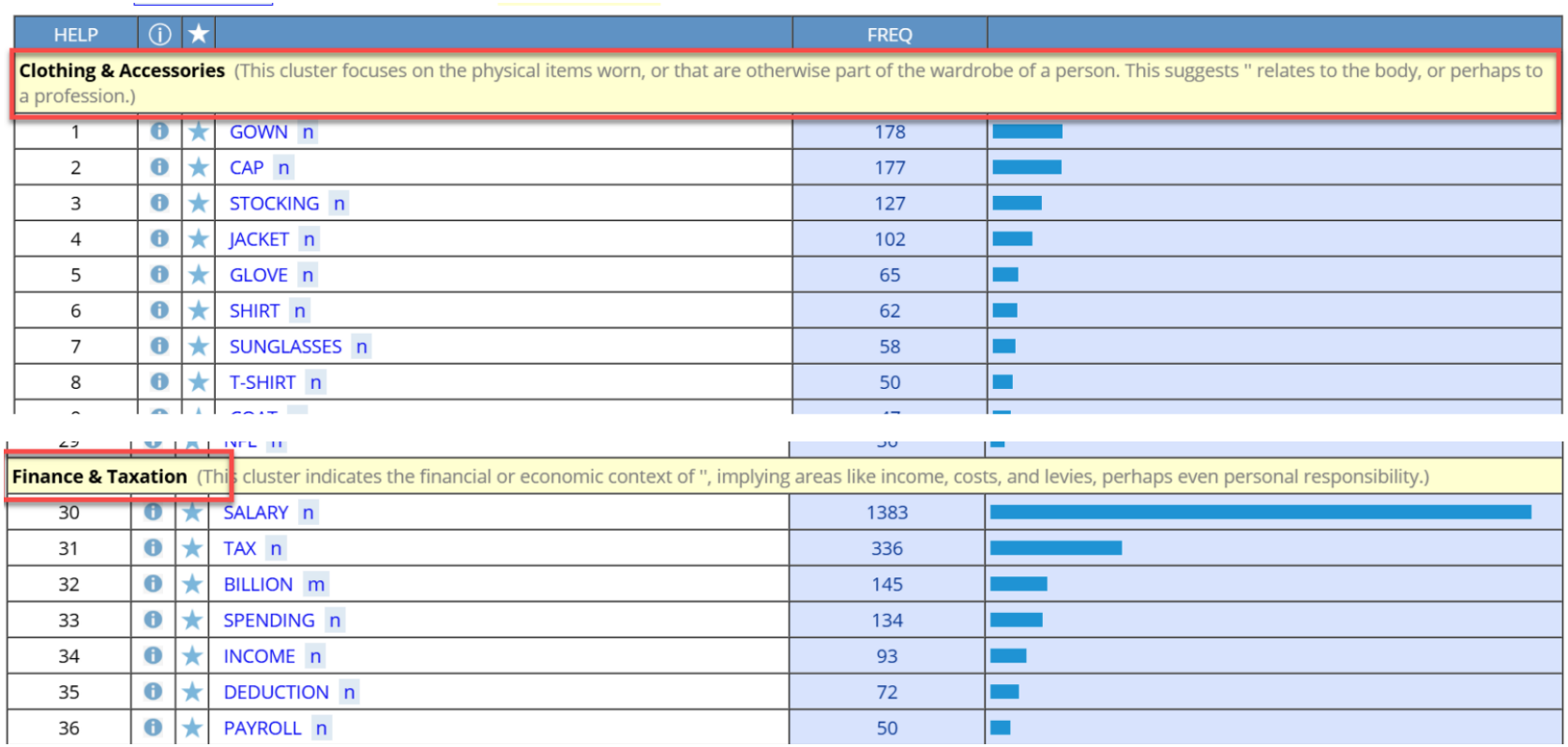


ON CLICK: CONTEXT | TRANSLATE (ES) | ENTIRE PAGE | GOOGLE | IMAGE | PRON/VIDEO | BOOK | THESAURUS | (HELP) | AI: CATEGORIZE

HELP	RE-USE WORDS	FREQ	ALL	%	MI
1	BASEBALL	1447	52303	2.77	7.34
2	SALARY	1360	20366	6.68	8.61
3	TRADE	823	100349	0.82	5.59
4	MARKET	512	184331	0.28	4.02
5	MILLION	463	304182	0.15	3.16
6	ICE	397	78434	0.51	4.89
7	RED	368	169934	0.22	3.67
8	WEARING	349	65746	0.53	4.96
9	SPACE	316	163889	0.19	3.50
10	TAX	291	153160	0.19	3.48
11	BALL	265	90551	0.29	4.10
12	BILLION	236	114903	0.21	3.59

Source: `https://www.english-corpora.org/ai-llms/collocates.pdf`

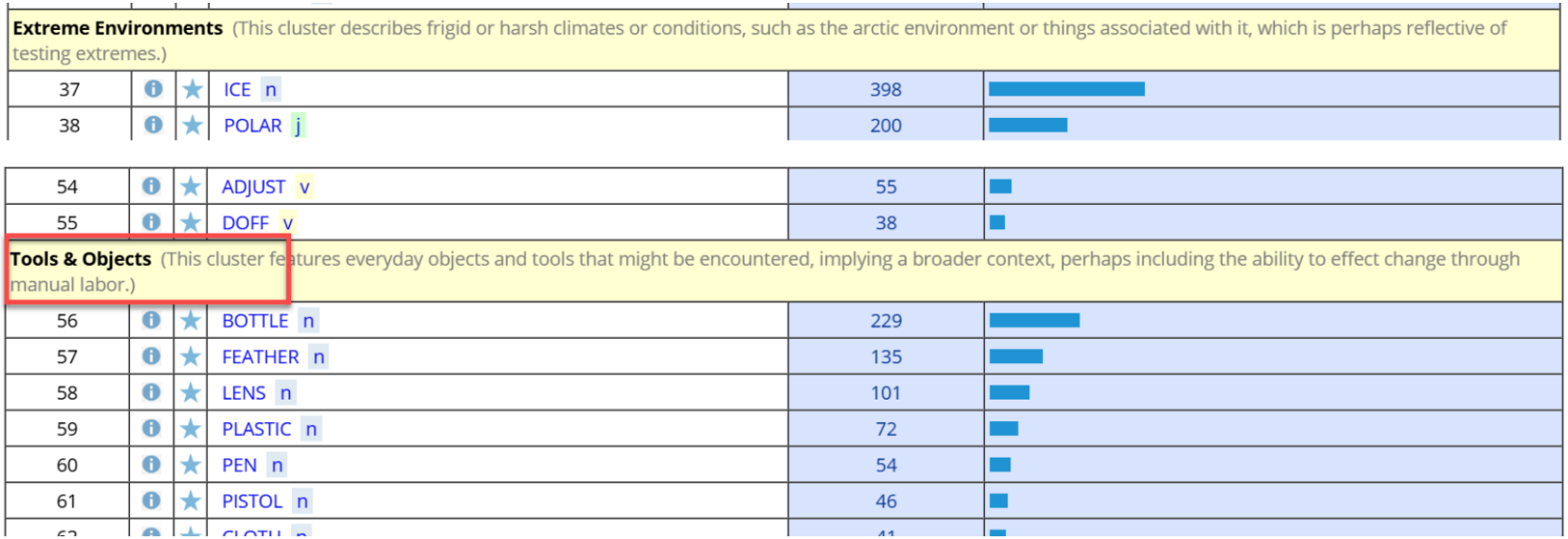
Example: Classifying the Collocates for the English noun cap, clustered by word sense



HELP			FREQ	
Clothing & Accessories (This cluster focuses on the physical items worn, or that are otherwise part of the wardrobe of a person. This suggests " relates to the body, or perhaps to a profession.)				
1		GOWN n	178	
2		CAP n	177	
3		STOCKING n	127	
4		JACKET n	102	
5		GLOVE n	65	
6		SHIRT n	62	
7		SUNGLASSES n	58	
8		T-SHIRT n	50	
Finance & Taxation (This cluster indicates the financial or economic context of ", implying areas like income, costs, and levies, perhaps even personal responsibility.)				
30		SALARY n	1383	
31		TAX n	336	
32		BILLION m	145	
33		SPENDING n	134	
34		INCOME n	93	
35		DEDUCTION n	72	
36		PAYROLL n	50	

Source: Mark Davies `https://www.english-corpora.org/ai-llms/collocates.pdf`

Example: Classifying the Collocates for the English noun cap, clustered by word sense



Extreme Environments (This cluster describes frigid or harsh climates or conditions, such as the arctic environment or things associated with it, which is perhaps reflective of testing extremes.)			
37	ICE n	398	
38	POLAR j	200	
54	ADJUST v	55	
55	DOFF v	38	
Tools & Objects (This cluster features everyday objects and tools that might be encountered, implying a broader context, perhaps including the ability to effect change through manual labor.)			
56	BOTTLE n	229	
57	FEATHER n	135	
58	LENS n	101	
59	PLASTIC n	72	
60	PEN n	54	
61	PISTOL n	46	

Source: Source: Mark Davies `https://www.english-corpora.org/ai-llms/collocates.pdf`

Main Take-home Messages

- LLM-generated content or linguistic analysis by humans is not an "exclusive or".
- Rather: progress in linguistics and NLP can best be facilitated by a synergy between the two approaches.
- Well-curated (and linguistically annotated) corpora have by no means become obsolete (see neologism use case). Their usefulness is by no means restricted to lesser-studied languages. This highlights the importance of ESFRI research infrastructures such as CLARIN and DARIAH.
- Well-prepared and well executed use cases which follow state-of-the-art practises are needed to evaluate task-specific uses of LLMs, rather than sweeping statements about for instance the supposed end of lexicography.