

The Game of Prompts – Identifying Auxiliary Constructions with LLMs and Human Participants in a Language Game

Tímea Borbála Bajzát¹²³, Balázs Indig³⁴

¹Eötvös Loránd University, Department of Modern Hungarian Linguistics

²National Laboratory for Digital Heritage

³Eötvös Loránd University, Doctoral School of Linguistics

bajzat.timea@btk.elte.hu

⁴Eötvös Loránd University, Department of Artificial Intelligence

indig.balazs@inf.elte.hu

Abstract

Recent work in theoretical linguistics has begun to integrate large language models (LLMs) with traditional corpus analysis to address limitations such as annotator fatigue and ethical constraints in human-subject research. While both methods rely on the statistical distribution of linguistic elements, LLMs offer a unique, scalable means of simulating human-like linguistic behaviour. In this paper, we examine Hungarian auxiliary constructions to investigate how closely LLMs align with human linguistic intuition. We compare responses from a gamified cloze test administered to human participants with outputs from multiple versions of GPT. Our results show that, although GPT-4o can achieve a human-level accuracy, it does so by using a high-recall, low-diversity strategy. In contrast, human participants demonstrate higher precision and more varied lexical choices. Although LLMs can mimic human-like semantic and syntactic intuitions, the qualitative differences in their frequency-driven strategies highlight their complementary role in linguistic research, rather than their ability to replace human judgement.

Keywords: linguistic constructions, cloze-test, LLMs, modality, gamification

1. Introduction

Cognitive linguists are increasingly integrating statistical methods from corpus analysis into their research. More recently, they have also begun incorporating LLMs too (Fitzmaurice and Kortmann eds., 2023; Fried and Nikiforidou eds., 2025). Within this framework, language is understood as consisting of symbolic units – form-meaning pairings that reflect patterns of usage and conceptual structure. The observable forms of language, as captured in a corpus, provide empirical data that can be both analysed directly and used to train LLMs. In turn, these models simulate aspects of human cognition by identifying and reproducing statistical patterns in language use. To capture a more comprehensive picture of linguistic behaviour, we adopt a three-pronged approach: triangulating data from human participants, LLM-generated outputs, and corpus-based patterns. This methodology enables us to examine the relationship between form and meaning from various angles. Our pilot study focuses on modality expression in Hungarian, specifically in constructions involving auxiliary verbs and auxiliary adjectives followed by infinitival clauses. Since Hungarian modality and auxiliary verbs are not considered closed classes, but rather gradient categories organised around prototypes, our analysis is based on the functional theoretical frameworks of Mortelmans (2010), Declerck (2011) and Van Linden (2010).

We use the construction candidates presented by Indig and Bajzát (2024) as input data, because their method identifies corpus-based patterns related to the selected linguistic phenomena. Their work is particularly relevant because it not only addresses the precise identifiability of verbal lemmas but also supports the analysis of interactions between construction frames and auxiliary verb slots. For the human-involved part of our experiment we adapted the language game introduced by Indig and Lévai (2023): a

gamified test framework based on the cloze test. The original game was conducted online, where participants were asked to guess the word that best fits in a given context. In each round the player is presented with a sentence fragment containing a missing word (a blank) and a fixed-length context of surrounding words. Participants could make multiple guesses, skip rounds or request additional examples of the same type, though, as Indig and Lévai note, this rarely improves performance. This observation aligns with our own findings. In the original study, both human participants and the BERT model (Devlin et al., 2019) achieved approximately 30% accuracy when predicting randomly masked target words, despite this being the model's primary training objective. They also found that human guesses were more semantically similar to the target word than those produced by the model. This insight prompted us to specifically examine the quasi-fixed elements of easily identifiable constructions, with the aim of empirically verifying their presence and their small number of possible substitutes. The third wheel of our experiments was multiple versions of GPT models (GPT-4o-mini, GPT-4o and GPT-4.1) available through API with different temperature values and prompt variations were used as a contemporary counterpart to BERT, which was used in the original experiment. In the following sections, we outline how we integrated these three data sources to gain deeper insights, and we discuss how these could be generalised.

2. Method

First, we examined the available examples from the Hungarian Webcorpus 2.0 (Nemeskey 2020) corpus based on the provided construction candidates (Indig and Bajzát, 2024). These were selected using the type-token ratio: we then selected 10 of these candidates (or all available items

where there were fewer than 10) and sampled the resulting examples to avoid duplication and erroneous matches. We randomly selected 20 examples for each construction candidate and manually verified them, resulting in a test set of approximately 4 000 instances. In total, we examined the constructions of 13 auxiliary verbs and adjectives (see Table 1).

Auxiliary verbs and adjectives	Sample Size from Hungarian Webcorpus 2.0	Filtered Sample Size	Sample size in the language game
<i>Akar</i> ('want to')	650 000	518 123	400
<i>bír</i> ('can')	179 846	112 112	280
<i>hajlamos</i> ('prone to')	272 806	29 142	280
<i>hajlandó</i> ('willing')	63 343	79 330	280
<i>imád</i> ('love to')	79 430	32 997	360
<i>képes</i> ('able to')	650 000	462 001	380
<i>képtelen</i> ('unable to')	164 909	104 274	300
<i>kíván</i> ('want to')	650 000	391 413	380
<i>mer</i> ('dare to')	473 966	278 887	400
<i>óhajt</i> ('wish to')	21 225	15 161	60
<i>szeret(ne)</i> ('like to / would like to')	650 000	448 324	400
<i>tud</i> ('can')	650 000	540 175	400
<i>utál</i> ('hate to')	16 652	9 464	60

Table 1: Sample size of the whole dataset (original sample from the Hungarian Webcorpus, filtered sample and the sample used in the language game)

We modified the original game to accept all paradigmatic forms of the lemmas based on regular expressions, allowing for all sensible forms of the examples. We also removed the hint about the number characters to be guessed to avoid bias. To prevent annotator fatigue and the priming of their guesses, we used filler examples which consisting of easily guessable postpositions and preverbs. The fillers were drawn from the Hungarian Gigaword Corpus UD version (HGC-UD) (Kalivoda et al., 2024). We applied a random sampling method but reviewed the samples manually. The guesses were analysed and the examples which were presented to the participants were used in the experiments involving the selected GPT models. To compare models of different sizes, and capabilities, and to examine how these differences influence linguistic predictions, we tested GPT-4.0-mini, GPT-4o, and GPT-4.1. In terms of prompt design, we experimented with (a) a prompt written by us, based on the description of the word-guessing game presented to the players, (b) a prompt generated by GPT-4.1, and (c) three

different temperature settings (0.0, 0.5 and 1.0). We also evaluated zero-shot and few-shot prompting using 1, 3 or 6 examples.

We wrote the prompt in Hungarian, since the language game and its environment were also in Hungarian. We defined the role for the model ("you are a native Hungarian speaker who has to participate in a language game..."). As mentioned above, we tested two different prompts in the case of zero-shot prompts. Since the GPT-4.1-generated prompt produced better results based on the percentages (see Table 2), we focused our experiments on that prompt. This prompt was different in that it specified what not to do when solving the task (e.g. "do not reason your answer", "do not use additional formatting" etc.).

	Temp / Prompt	0.0	0.5	1.0
1 st version	0-shot gpt-4o-mini	5.01%	8.28%	8.92%
	0-shot gpt-4o	31.21%	28.66%	31.21%
	0-shot gpt-4.1	24.84%	25.48%	25.48%
2 nd version	0-shot gpt-4o-mini	7.32%	6.37%	8.92%
	0-shot gpt-4o	31.85%	32.17%	35.03%
	0-shot gpt-4.1	26.75%	26.75%	23.57%
	1-shot gpt-4o-mini	20.38%	14.97%	17.52%
	1-shot gpt-4o	38.22%	40.45%	37.90%
	1-shot gpt-4.1	35.67%	32.48%	32.80%
	3-shot gpt-4o-mini	20.38%	19.43%	21.02%
	3-shot gpt-4o	46.50%	46.50%	45.22%
	3-shot gpt-4.1	40.45%	38.55%	39.49%
	6-shot gpt-4o-mini	24.20%	27.39%	26.75%
	6-shot gpt-4o	54.78%	54.14%	48.73%
	6-shot gpt-4.1	47.13%	48.09%	48.73%

Table 2: The performance of the models.

The example clauses were infinitival constructions and taken from the HGC-UD corpus (Kalivoda et al. 2024). To avoid influencing the content of the generated response, we provided several verb examples as output without specifying the correct answer. The few-shot procedure was necessary for the model to recognise the construction type (verb and infinitive) and learn the expected output form (one word per tip, multiple tips are permitted, and the expected outputs are verbs or adjectives). The number of output tokens was limited to 50.

Auxiliary verbs and adjectives	Number of the games	Clauses in the games	Median of the clauses for one game	Performance (humans) %	Performance (LLM) %	Lemma types (humans)	Lemma types (LLMs)
<i>akar</i> (‘want to’)	42	93	1.00	76.19	57.14	26	29
<i>bír</i> (‘can’)	20	72	3.50	40.00	55.00	24	11
<i>hajlamos</i> (‘prone to’)	19	47	2.00	47.37	63.16	27	12
<i>hajlandó</i> (‘willing’)	16	32	1.00	12.50	31.25	23	13
<i>imád</i> (‘love to’)	34	132	3.00	26.47	23.53	49	13
<i>képes</i> (‘able to’)	23	77	3.00	47.83	82.61	34	18
<i>képtelen</i> (‘unable to’)	24	78	3.00	45.83	34.78	35	20
<i>kíván</i> (‘want to’)	23	63	2.00	26.09	47.62	31	18
<i>mer</i> (‘dare to’)	35	109	3.00	42.86	40.00	41	14
<i>óhajt</i> (‘wish to’)	4	12	2.50	0.00	0.00	11	4
<i>szeret(ne)</i> (‘like to / would like to’)	32	84	2.00	71.88	93.32	41	15
<i>tud</i> (‘can’)	39	96	1.00	76.92	74.36	29	18
<i>utál</i> (‘hate to’)	3	9	4.00	33.33	0.00	13	6
All	314	904	2.50	50.00	54.78	155	59

Table 3: Basic descriptive statistics related to the test material and the performance (human and GPT)

3. Results and discussion

3.1. Results of the prompt experiments

As shown in Table 2, the best performance, at 54.78%, was achieved using 6-shot prompting with a temperature of 0.0 on GPT-4o. Interestingly, the results at a temperature of 0.5 were very similar, suggesting that an increase in randomness has little effect on the output in this setting. Overall, GPT-4o performed best at low temperatures (0.0 and 0.5) across the different prompt designs. This suggests that more deterministic generation tends to produce higher accuracy. The human participants performed at an overall level of 50.00%, which was not higher than the model’s results (see Table 3 for more details). Because this study has a limited scope, we are focusing on comparing the performance of the top-performing model with that of the human participants.

3.2. Comparison of human and LLM performance

First, we present a detailed comparison of the results obtained from human participants and those from the prompt design that performed best. Interestingly, we found that success rates varied significantly across the different types of auxiliary verbs. This consistent trend was observed in both human and LLM outputs (see Table 3).

As can be seen in Table 3, the LLM performs slightly better overall, though this difference is not significant. Both the LLM and the human participants performed poorly regarding non-prototypical or less common auxiliary verbs (e.g. *imád*) and adjectives. Overall, humans use many more

distinct types of lemmas. It is reasonable to assume that the human participants were more creative and had a more diverse word choice, while the model was more uniform. This is not surprising, since we set the temperature value to zero. However, it can be seen in Table 2 that higher temperature values lead to less accurate results. It is also important to note that the human data contributors also showed greater diversity in terms of the parts of speech they selected. For example, the human side used in some cases conjunctions, whereas the model guessed almost exclusively verbs and adjectives. The model performs much better with positive forms (82.61%) than negative ones (34.78%), whereas humans perform similarly in both cases (47.83% and 45.83% respectively). This suggests that the model is sensitive to polarity in a way that humans are not (e.g. due to its positiveness).

	Lemma	%	Prec.	Recall	F-score
1.	tud	13.23	0.2273	0.7692	0.3509
2.	akar	10.62	0.3019	0.7619	0.4324
3.	szeret	10.22	0.2255	0.7188	0.3433
5.	képes	7.31	0.1644	0.4783	0.2447
9.	mer	3.01	0.5000	0.4286	0.4616

Table 4: The top 10 most frequent guesses for lemmas from human participants (only the tested lemmas are shown).

	Lemma	%	Prec.	Recall	F-score
1.	tud	19.32	0.1706	0.7434	0.2776
2.	szeret	16.02	0.2128	0.9332	0.3476
3.	akar	13.18	0.2069	0.5714	0.3038
4.	képes	7.73	0.2794	0.8261	0.4176
6.	kíván	4.43	0.2308	0.4762	0.3109
10.	hajlandó	2.05	0.7778	0.3125	0.4459

Table 5: the top 10 most frequent guesses for lemmas from GPT-4o (only the tested lemmas are shown).

A comparison of human participants and GPT-4o showed that both groups relied heavily on a small set of frequently used words, such as *tud*, *akar* and *szeret*. However, their guessing strategies differed. Human participants achieved higher precision with these frequent lemmas, especially with *akar* ('want to') (F-score: 0.4324 vs. 0.3038). However, the model made more frequent guesses overall, resulting in high recall (e.g. 0.9332 vs. 0.7188 for *szeret*), but lower precision. GPT-4o demonstrated clear strengths when handling certain items, such as *képes* (F-score: 0.4176 vs. 0.2447 for humans) and *hajlandó* (precision: 0.7778 vs. no guesses observed).

Despite using far fewer different types (59 vs. 155), GPT-4o slightly exceeds the human average for the examined lemmas in Tables 4 and 5, which suggests a lack of lexical diversity. These findings suggest that large language models may use a high-recall, high-risk strategy involving the excessive use of common word forms. While this improves coverage, it lowers precision, which is favourable for generation. However, the experiment also shows that GPT-4o cannot replace human participants, as its strategy leads to qualitatively different lexical choices.

Both humans and the model often guessed different lemmas to the ones being examined (e.g. *fog* 'will', *lehet* 'may', *szabad* 'allowed to'). This is not surprising, as these typical guesses are very abstract and fit functionally. For instance, *fog* is an auxiliary verb in Hungarian that is used to express the future tense, so it can appear in a variety of contexts. Both humans and LLM also frequently associated verbs expressing aspect (*próbál* 'try') and habituality (*szokott* 'used to') with the masked lemmas.

Although many F-scores fall within the range of 0.2–0.4, this does not necessarily indicate poor performance. The task involves selecting the exact target lemma from a large set of alternatives (result in grammatically correct output), which is a challenging task even for human participants, which helped to reveal associates.

Lemma	Modal/Premodal					Non-modal	
	Ab	V	D	Ha	At	TA	O
akar	10.0	42.0	25.0	3.0	1.0	11.0	8.0
bír	31.7	9.8	28.0	2.4	6.1	6.1	15.9
hajl.	27.6	1.7	10.3	12.1	27.6	5.2	15.5
hajlan	22.2	12.7	15.9	3.2	27.0	11.1	7.9
imád	12.5	8.1	19.4	1.9	25.6	10.0	22.5
képes	28.6	2.9	28.6	2.9	11.4	11.4	14.3
képt.	39.0	5.2	11.7	0.0	5.2	16.9	22.1
kíván	16.2	35.9	21.4	1.7	0.9	16.2	7.7
mer	20.0	16.4	27.1	2.9	16.4	10.7	6.4
óhajt	17.6	29.4	23.5	5.9	0.0	23.5	0.0
szeret	11.0	26.4	20.9	15.4	2.2	9.9	14.3
tud	37.2	6.4	28.7	3.2	2.1	8.5	13.8
utál	15.8	0.0	21.1	5.3	21.1	36.8	0.0

Table 6: The semantic source domains of the guesses from human participants (hajl. = hajlamos, hajlan. = hajlandó, képt. = képtelen, Ab = Ability, V = Volition, D = Deontic, Ha = Habitual, At = Attitude, TA = Tense and aspect O = Other)

Lemma	Modal/Premodal					Non-modal	
	Ab	V	D	Ha	At	TA	O
akar	15.2	32.6	12.1	3.0	9.8	16.7	10.6
bír	51.7	16.7	20.0	0.0	5.0	5.0	1.7
hajl.	36.7	1.7	1.7	43.3	10.0	6.7	0.0
hajlan	37.5	15.0	0.0	7.5	27.5	10.0	2.5
imád	17.1	15.9	2.4	7.3	51.2	3.7	2.4
képes	53.3	5.0	8.3	3.3	10.0	16.7	3.3
képt.	47.7	15.4	6.2	0.0	12.3	13.8	4.6
kíván	16.2	35.9	21.4	1.7	0.9	16.4	13.1
mer	29.6	14.8	16.7	0.9	29.6	10.2	0.9
óhajt	50.0	50.0	0.0	0.0	0.0	0.0	0.0
szeret	12.0	38.0	2.2	0.0	34.8	8.7	4.3
tud	37.5	18.3	16.3	5.8	0.0	8.7	13.5
utál	28.6	0.0	0.0	14.3	28.6	0.0	28.6

Table 7: The semantic source domains of the guesses from GPT-4o (hajl. = hajlamos, hajlan. = hajlandó, képt. = képtelen, Ab = Ability, V = Volition, D = Deontic, Ha = Habitual, At = Attitude, TA = Tense and aspect O = Other)

Tables 6 and 7 show that both humans and GPT-4o demonstrated strong semantic consistency. Even when they produced incorrect guesses, these were usually within the same conceptual domain as the target lemma. For instance, *akar* ('want to') predominantly attracted volition-related guesses (42% for humans and 32.6% for GPT-4o), while *bír* ('can') and *képes* ('able to') were mostly classified as ability (31.7% and 28.6% for humans, and 51.7% and 53.3% for GPT-4o). This clearly distinguished between ability-related verbs (*tud*, *képes*, *bír*, *képtelen*) from volition/desire verbs (*akar*, *szeret(ne)*, *kíván*, *óhajt*), with the additional distinct groups of habitual verbs (*hajlamos*) and attitudinal verbs (*imád*, *utál*, *szeret*).

The verb *szeret* is a special case because, depending on the mood marking (conditional), it can express either volition (*szeretne*, *szeret-Cond.3Sg*, 'would like to') or attitude (*szeret*, *szeret.Prs.1Sg*, 'like to'), which explains its distribution across both categories (Kálmán et al., 1989). GPT-4o is even more consistent than human participants in remaining within a single semantic category, which often improves its performance (e.g. *képes*, *imád*), but sometimes results in miscategorisation (e.g. *mer* is often mapped to the Ability or Deontic categories, resulting in lower precision).

Humans made more guesses in the Other category, likely reflecting their greater lexical diversity. In contrast, GPT-4o showed a stronger tendency towards categorical clustering. The empirical study showed, that both humans and the model rely on conceptual domains. This is a topic worthy of detailed examination in the future.

4. Conclusion and future work

We conducted an experiment involving corpus patterns, GPT models, and human participants using a cloze-test-like task to explore their underlying semantic strategies. Our findings show that model performance is highly sensitive to prompt design and significantly improves with the inclusion

of multiple examples. By contrast, humans approached the task with ease and consistency.

Although the best-performing model setting achieved accuracy comparable to humans, their strategies were notably different. GPT-4o used a high-recall, low-diversity approach, whereas human responses were more varied and exhibited higher precision in targeting specific lemmas. Both humans and the model maintained strong semantic consistency, typically generating guesses within the same conceptual domain, verifying the theory empirically.

Our future work will focus on developing prompts that more closely align with human strategies. Our aim is to mitigate annotator fatigue and improve coverage of the blind spots in the corpus cheaply and reliably. This will enable us to collect data on underrepresented cases and conduct further investigations on the divergence between human and model behaviours. On the other hand, we plan to apply collostructional analysis (Gries and Durrant, 2020) to the examined patterns to enhance our comprehension of how human participants and LLMs approach them, providing more insight into their distinct strategies. Finally, any candidate words suggested by the guessers that are not present in the original list of auxiliaries should be thoroughly checked in the corpus. This will improve the recall of possible patterns and close the feedback loop.

Acknowledgement

This paper was supported by the Bolyai János Research Scholarship no. BO/00260/25/1 of the Hungarian Academy of Sciences.

References

- Decklerck R. (2011). The Definition of Modality. In: Patard A. and Brisard F. (Eds.) *Cognitive Approaches to Tense, Aspect, and Epistemic Modality*. Amsterdam, Philadelphia: John Benjamins, pp. 21–44.
- Devlin J., Chai M., Lee K. and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J., Doran C. and Solorio T. (Eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*
- Fried M. and Nikiforidou K. (Eds.) (2025). *The Cambridge Handbook of Construction Grammar*. Cambridge: Cambridge University Press.
- Fitzmaurice S. M. and Kortmann B. (Eds.) (2023). *Models of Modals*. Berlin, Boston: De Gruyter Mouton.
- Gries S. T. and Durrant P. (2020). Analyzing co-occurrence data. In: Paquot M. and Gries, S. T. (Eds.) *A practical handbook of corpus linguistics*. Champ: Springer, pp. 141–159.
- Indig B. and Lévai D. (2023). I’m Smarter than the Average BERT! – Testing Language Models Against Humans in a Word Guessing Game. In: Vetulani Z. and Paroubek P. (Eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań: Adam Mickiewicz University, pp. 106–110.
- Indig B. and Bajzát T. (2024). Compressing Noun Phrases to Discover Mental Construction in Corpora. In: Hämäläinen M., Pirinen F., Macias M. and Avila C. M. (Eds.) *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*. Helsinki: Association for Computational Linguistics, pp. 96–103.
- Kalivoda Á., Malouf R., Ackerman, F., Sass B. (2022). Building a dependency treebank from the Hungarian Gigaword Corpus. *15th International American Association for Corpus Linguistics Conference (AACL 2022)*. Flagstaff (poster).
- Kálmán C. Gy., Kálmán L., Nádasdy Á., and Prószéky G. (1989). A magyar segédigék rendszere. In: Telegdi Zs. and Kiefer F. (Eds.) *Általános Nyelvészeti Tanulmányok XVII. Tanulmányok a magyar mondatok köréből*. Budapest: Akadémiai Kiadó, pp. 49–103.
- Mortelmans T. (2010). Modality in Cognitive Linguistics. In: Geeraerts D. and Cuyckens H. (Eds.) *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, pp. 869–889.
- Nemeskey D. M. (2020). *Natural Language Processing Methods for Language Modeling*. PhD thesis. Eötvös Loránd University.
- Van Linden A. (2010). From premodal to modal meaning: Adjectival pathways in english. In: *Cognitive Linguistics*, 21, pp. 537–571

