A language model approach to near-synonymous grammatical constructions

Yusuke Minami[1], Yuzo Morishita[2]
[1]Kobe University, [2]St. Andrew's University
[1]y-minami@people.kobe-u.ac.jp
[2]ymorishi@andrew.ac.jp

Abstract

Many linguistic studies have examined subtle semantic/functional differences between near-synonymous pairs of grammatical constructions. In such studies, however, as synonymity is identified based solely upon scholars' intuitions, potential connections across constructions could be left unnoticed. To addresses this issue, this study seeks a relationship between the personal necessity construction (*You don't need to do that*.) and the impersonal necessity construction (*There's no need to do that*.), which have seldom been treated as being synonymous presumably because of the huge structural differences. We conducted an experiment with a (large) language model, GPT-2. First, 50 examples of each construction, including its preceding context, were randomly sampled from the *Corpus of Contemporary American English*. Then an alternative version for each construction was created by replacing the original with the other, and the same procedure was carried out without preceding contexts. Finally, each of the 400 stimuli was evaluated using GPT-2, and their cross-entropy values were computed. The differences between the original and alternative ones were statistically examined by paired *t*-tests. The results suggest that, while the constructions differ with respect to preferred probabilistic contexts, they are treated as being more similar in isolation, thereby supporting the synonymous relationship between the two constructions.

Keywords: synonymy, modals, language model, cross-entropy

1. Introduction

Close comparison of different lexical items that are in a (near-)synonymous relation has long been one of the most common practices in functional linguistics. Since the advent of Construction Grammar toward the end of the 20th century (Fillmore et al., 1988; Goldberg, 1995; Fillmore & Kay, 1999, etc.), this practice has been extended to the clause-level phenomena (e.g., syntactic alternations of argument structure constructions; Goldberg, 1995; Bresnan et al., 2007). The purpose of this paper is to seek the possibility of utilizing a large language model, GPT-2 (**G**enerative **P**re-trained **T**ransformer 2), to properly capture near-synonymous relations among distinct grammatical constructions, including ones which could fall outside of researchers' intuition-based observations. As a case study, we will focus on a potentially near-synonymous pair of clause-level modal expressions which has received little attention in the literature of modality (see 2.1 for details): personal necessity construction (*You don't need to do that.*) and impersonal necessity construction (*There is no need to do that.*).

This paper is organized as follows. Section 2 provides theoretical, practical and technical backgrounds for the present study. Section 3 explains the methods of the experiment conducted in the present study. Section 4 presents the results of the experiment. Section 5 discusses what the results reveal about the nature of synonymous relations across clause-level constructions, and argues for the utility of LLMs that will complement the traditional, intuition-based analysis. Section 6 concludes this paper.

2. Background

2.1 Near-synonymous relations between grammatical constructions

Traditionally, the relation of synonymy was a matter of word meaning, i.e., it was treated as a relation that holds between a pair (or set) of distinct lexical items. However, the central assumption adopted in the present study, which aligns with a relatively bnew grammatical theory called Construction Grammar, is that basic linguistic units in the knowledge of language are linguistic signs (i.e., form-meaning pairs) of any size and any schematicity that are uniformly called "constructions." Under this theoretical assumption, constructions which are larger in size than words have also come into scope, and synonymy relations among those constructions have naturally come under scrutiny. In particular, clause-level syntactic alternations such as dative alternation (*Mary gave me a book*. vs. *Mary gave a book to me*.) and locative alternation (*John loaded hay onto the truck*. vs. *John loaded the truck with hay*.) have been extensively studied (Perek, 2015, a.o.), building upon the assumption that the relation between the two constructions that constitute each alternation is characterized by near-synonymity. It should be noted that it is *near*-synonymy, not *absolute* synonymy, that is supposed to be observed between the two variants of alternations. Underlying this idea is the *principle of isomorphism*, which states that language should be organized in the way that there are exclusive mappings between form and meaning (Bolinger, 1977; Haiman, 1980, a.o.).

2.2 Why these two constructions?

Under the tenets outlined in 2.1, a number of qualitative studies on the nuanced semantic/functional differences between near-synonymous clause-level constructions have been conducted. Among the phenomena commonly treated are constructions in English that include modal verbs (henceforth, "model verb constructions"). Traditionally, a major part of research on modal verb constructions was devoted to close comparison between pairs of modal verbs (including "quasi-modals") characterized by a considerable degree of functional overlap (e.g., *must* vs. *have to*, *should* vs. *ought to*, *may* vs. *can*, etc.). With the advancement of large-scale electric corpora, recent development has seen a major shift to more quantity-oriented studies that often go along with statistical modelling and psycholinguistic experiments (Depraetere & Cappelle, 2023).

However, there remains an area that is yet to be explored in the literature: comparisons across modal expressions of different lexical categories. It is well known that along with modal verbs there are also modal adjectives and nouns, but research on the latter is scarce, let alone comparison between the former and the latter. This is somewhat surprising, considering the fact that an onomasiological

approach based upon a limited number of specific "modal" concepts (e.g., obligation, permission, ability, possibility, etc.) is as well-established as a semasiological one.

The present study addresses this gap by focusing on two grammatical constructions which denote the "absence of necessity" with different lexical categories: the personal necessity construction which involves modal verb *need* (e.g., *You don't* ***need*** *to do that*) and the impersonal necessity construction which involves modal noun *need* (e.g., *There's no* ***need*** *(for you) to do that*). The two constructions differ also in the way the logical subject of the *to*-infinitival verb is encoded: it is realized as the subject of the main clause in the former, while in the latter it is left unexpressed as an "implicit argument" or expressed as the non-finite subject with the marker *for*. These two differences combined may have contributed to diverting researchers' attention from the potential relationship between the two constructions. It is for these reasons that we chose this particular pair as the test case.

2.3 Quantitative analyses of alternation

Research on syntactic alternations and two related constructions was once centered on traditional qualitative analysis, as in Levin (1993) and Pinker (1989), but with the spread of statistical methods using corpora in linguistic research, quantitative research has also become active over the past decades. Since Bresnan et al. (2007), whose work has been one of the catalysts for the spread of quantitative analyses of grammar, logistic regression analysis has been the central quantitative method for grammatical research of English and other languages. They set up 14 variables, such as the recipient's accessibility and the theme's definiteness, and created a model that predicts under what conditions and with what probability each construction is realized.

The common goal of linguistic research on grammatical alternation is to clarify what underlies the differences between (two) alternative expressions. Thus, (binomial) logistic regression makes it clear which variables influence grammatical choice and to what extent, and its high interpretability is one reason why many linguists have preferred this method. In fact, compared with methods like logistic regression, the current analysis method using (large) language models can be said to be black-box-like (Baayen, 2024: 621).

Although language models may lose a certain degree of interpretability that methods like regression analysis would have, it is also true that language models are rapidly gaining traction in linguistics in recent years. Morger and Berdicevskis (2026) analyze English and Swedish constructions using BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers). Despite their lower interpretability, language models offer important advantages over traditional regression-based approaches. As noted above, by using a language model, our study aims to clarify similarities and differences between distinct constructions that are not accessible to human intuition alone.

2.4 Language models and linguistics

As mentioned in 2.3, while some researchers are against using language models on the grounds that they are black boxes, others argue that language models should be incorporated positively into language research (e.g., Futrell and Mahowald, 2025). As pointed out in Anthoney (2025), researchers in the latter group are convinced that language models inherit the ideas of traditional corpus linguistics (Sinclair, 1991) and distributional semantics proposed by Firth (1957). In what follows, we explain the mechanism of GPT-2 used in this study, introduce related studies that apply GPT-2 to linguistic research, and demonstrate how useful language models can be for language studies.

Since 2022, the term 'language model' or 'AI' has come to be associated with chatbots like ChatGPT and Gemini. In reality, however, language models, or large language models, are systems that assign probabilities to words or contexts given contexts of several or even tens of thousands of words. The probability distributions of language models are obtained by training on vast amounts of text. For example, GPT-2 (Radford et al., 2019) was pre-trained on a dataset of 8 million web pages.

To get a sense of how language models predict the next word given context, consider the following example.

(1) a. It rained yesterday, so let's go on a picnic.

b. It rained yesterday, so I'm gonna stay home today.

In these sentences, the pre-contexts are exactly the same, but the words that follow *so* are completely different. Given *it rained yesterday*, language models can assign probabilities to *let's*, *go*, *on*, *a*, and *picnic* in (1a) and to *I'm*, *gonna*, *stay*, *home*, *today* in (1b). The probability of each word was calculated using the following formula because GPT-2 is a causal/autoregressive language model that iteratively predicts word probabilities left-to-right.

(2) a. *P* (***so*** | *It rained yesterday,*)

b. *P* (***let's*** | *It rained yesterday, so*)

…

f. *P* (***picnic*** | *It rained yesterday, so let's go on a*)

(3) a. *P* (***so*** | *It rained yesterday,*)

b. *P* (***I'm*** | *It rained yesterday, so*)

…

f. *P* (***today*** | *It rained yesterday, so let's go on a*)

For example, the probability of *picnic* is 0.0038, while that of *stay* is 0.017. This result indicates that, given the pre-context *it rained yesterday*, the probability of the word 'picnic' is very low; given the pre-context, the probability of the word 'stay' is much higher. As a side note, the reason this study uses GPT-2, a relatively small, older language model rather than newer models like GPT-5, is that GPT-2 is fully open source, enabling highly reproducible research. Additionally, GPT-2 can be analyzed on standard GPUs, which keeps computational costs low even when experiments are repeated many times.

In recent years, there has been a growing number of attempts to empirically validate linguistic theories (including cognitive/functional linguistics and generative linguistics) using the language models illustrated above. For example, Warstadt et al. (2020) use minimal pairs, a traditionally employed method in linguistics, to examine how language models capture morphosyntactic knowledge. They use a total of 67,000 pairs of grammatical and ungrammatical sentences such as the following.

(4) a. The cats licked themselves.

b. *The cats licked itself.

Comparing the probability that *themselves* occurs after *the cats licked* with the probability that *itself* occurs after *the cats licked*, the model predicts that *themselves*, the grammatical expression, occurs with a higher probability; then it can be judged that the language model possesses grammatically correct knowledge. This study is suggestive in that it reveals that language models that have not undergone explicit grammatical learning nonetheless possess sufficient morphosyntactic features that have been investigated in the theoretical framework of generative grammar. Furthermore, Potts (2018) argues that advances in deep learning research have enabled semantics to be more closely tied to concrete usage than ever before. Such claims are also consistent with the usage-based model (Langacker, 1999) and prototype semantics (Lakoff, 1987). Potts suggests the potential for research that treats meaning as a vector in order to further advance the study of meaning.

3. Methods

3.1 Corpus data

In this study, we extracted all examples from the offline text edition of the *Corpus of Contemporary American English* (COCA; Davies, 2008), a balanced corpus of American English comprising approximately one billion words across eight registers (spoken, fiction, magazines, newspapers, academic writing, TV/movie, blogs, and others). Candidate examples of personal necessity construction and impersonal necessity construction were drawn from the corpus files with regular expression searches. The raw text was first segmented into sentences, with colons and semicolons treated as sentence delimiters in addition to sentence-final punctuation, so that a search hit never spanned two independent clauses. For the impersonal construction, the pattern matched *there* followed by a form of *be* (including contracted, perfect, and modalized forms such as *there's*, *there has been*, *there may be*) plus *no need to* or *no need for NP to*. For the personal construction, the pattern matched a subject followed by *do not* / *does not* / *did not* (full or contracted) plus *need to*. Because an adverb frequently intervenes between the negated auxiliary and the verb (e.g., *don't* ***even*** *need to*, *don't* ***really*** *need to*), the extraction was rerun with patterns that allowed such intervening material. The extracted examples were then screened manually. We jointly inspected every candidate and excluded examples such as the ones in (5).

(5) I only wish you didn't have to leave us so soon.

Only examples on which both annotators agreed were retained. The items used in the present study were randomly sampled from these vetted pools.

3.2 Stimuli

The stimuli for this study consisted of 100 original target sentences, 50 containing the personal necessity construction and 50 containing the impersonal necessity construction, each accompanied by its immediately preceding sentence in the corpus, which served as the preceding context. For each original sentence, we constructed a minimally different counterpart by swapping the verb phrases while leaving the context untouched. For example, the corpus-original impersonal sentence in (6b) was paired with the constructed personal counterpart in (6c), and both were evaluated against the same corpus context in (6a).

(6) a. CONTEXT: Paul clambered to the top, then stepped out onto the very narrow ledge and looked toward the west.

b. ORIGINAL (impersonal): There was no need to use his telescope.

c. SWAPPED (personal): You did not need to use his Telescope.

This yielded 100 item pairs (200 target sentences in total). 50 *personal-original* pairs, in which the corpus data is personal and the swapped variant impersonal, and 50 *impersonal-original* pairs, in which the reverse holds. Every target sentence was evaluated under two presentation conditions: a with-context condition, in which the target was preceded by its corpus context, and a without-context condition, in which the target was presented in isolation. The logic of the design is that if each construction is associated with a preferred discourse, the corpus-original member of a pair should be more predictable than its swapped counterpart when the original context is present, whereas differences in the without-context condition can reflect context-independent properties of the sentences themselves.

3.3 Cross-entropy computation

In this study, we defined predictability as the mean cross-entropy under GPT-2 (base model, 124M parameters; Radford et al., 2019), accessed through the transformers library (Wolf et al., 2020). For each sentence, the target sentence was tokenized with the model's byte-pair-encoding tokenizer and its mean negative log-probability per token, in nats, was computed:

$$CE(t_{1:n} \mid c) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(t_i \mid c, t_{1:i-1})$$

Where $t_{1:n}$ are the tokens of the target sentence, and c is the preceding context. In the with-context condition, the context was fed to the model as a prefix but excluded from the loss, so that the cross-entropy reflects the predictability of the target tokens only; in the without-context condition, c was empty. Lower cross-entropy indicates higher predictability.

Three directional hypotheses were tested here. First, in the with-context condition, corpus-original sentences were predicted to have lower cross-entropy than their swapped counterparts. Second, no such prediction was made for the without-context condition. Third, cross-entropy with context was predicted to be lower than cross-entropy without context, reflecting the general facilitative effect of discourse context on prediction (Hale, 2001; see also Levy, 2008).

4. Results

4.1 Descriptive statistics

For each comparison, we report descriptive statistics, a one-tailed paired *t*-test, and Cohen's d_z for paired data ($d_z = M_{diff}/SD_{diff}$). The Shapiro-Wilk test indicated non-normality of the paired differences in several comparisons, so a one-tailed Wilcoxon signed-rank test is reported alongside each *t*-test as a non-parametric check. All

analyses were conducted in R version 4.3.1 (R Core Team, 2023); the data and analysis scripts are available here (https://github.com/ymorishi-source/necessity-constructions-gpt2).

Comparison	*n*	M_1 (*SD*)	M_2 (*SD*)	*t*	*p*	d_z	%
With context: original vs. swapped							
Personal-original	50	2.74 (0.66)	2.96 (0.56)	−3.95	<.001	−0.56	78
Impersonal-original	50	2.91 (0.64)	3.09 (0.68)	−4.69	<.001	−0.66	70
Without context: original vs. swapped							
Personal-original	50	3.21 (0.57)	3.24 (0.51)	−0.50	.310	−0.07	56
Impersonal-original	50	3.13 (0.63)	3.30 (0.66)	−4.87	<.001	−0.69	84
Context effect: with vs. without context							
All targets	200	2.92 (0.65)	3.22 (0.59)	−8.91	<.001	−0.63	76

Table 1: Summary of the paired cross-entropy comparisons (nats/token). CE = mean cross-entropy; d_z = Cohen's *d* for paired data; % = percentage of pairs in the predicted direction (original < swapped, or with < without context). All tests one-tailed.

4.2 Personal-original pairs in context

We first examined the 50 pairs whose original corpus examples instantiate the personal construction. When the original context was retained, the corpus-original personal sentences were more predictable (M = 2.74, SD = 0.66, Mdn = 2.63) than their swapped impersonal counterparts (M = 2.96, SD = 0.56, Mdn= 2.98), a mean paired difference of −0.21 nats/token (SD = 0.38). The Shapiro-Wilk test did not reject normality of the differences (W = 0.96, p = 0.72). The difference was significant in the expected direction, $t(49) = -3.95$, $p < .001$ (one-tailed), $d_z = -0.56$, and the Wilcoxon signed-rank test also indicated a significant difference between the two conditions, V = 229, $p < .001$. The original sentence yielded lower cross-entropy than its swapped counterpart in 39 of the 50 pairs (78%). Figure 1 summarizes these results.

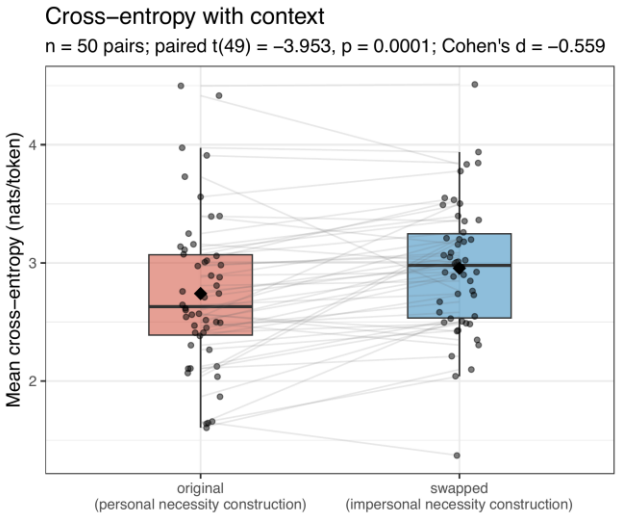


Figure 1: Cross-entropy (nats/token) of original personal sentences from the corpus and their swapped impersonal counterparts, with the original corpus context retained (n = 50 pairs). Gray lines connect paired sentences, and diamonds indicate condition means.

4.3 Impersonal-original pairs in context

The same pattern was obtained for the 50 sentence pairs extracted from corpus examples of the impersonal construction. When the original context was retained, the corpus-original impersonal sentences (M = 2.91, SD = 0.64, Mdn = 2.91) were more predictable than their personal swapped counterparts (M = 3.09, SD = 0.68, Mdn = 3.07), a mean paired difference of −0.19 nats/token (SD = 0.28). The Shapiro-Wilk test did not reject normality of the differences (W = 0.93, p = .006). Both the paired *t*-test, $t(49) = -4.69$, $p < .001$ (one-tailed), $d_z = -0.66$, and the Wilcoxon signed-rank test, V = 216, $p < .001$. The original impersonal sentence yielded lower cross-entropy than its swapped counterpart in 39 of the 50 pairs (78%). Thus, for both construction types, the corpus-attested construction was more predictable in its original discourse context than the minimally different alternative, consistent with the hypothesis that the two constructions are associated with distinct contextual environments. Figure 2 summarizes these results.

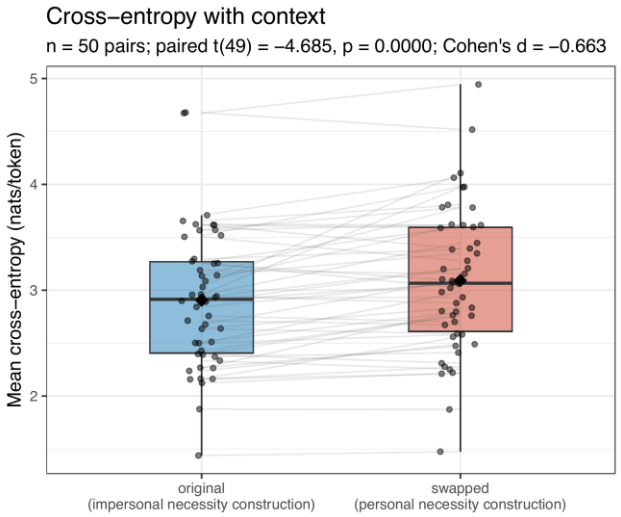


Figure 2: Cross-entropy (nats/token) of original impersonal sentences from the corpus and their swapped impersonal counterparts, with the original corpus context retained (n = 50 pairs). Gray lines connect paired sentences, and diamonds indicate condition means.

4.4 Comparisons without context

To determine whether the original sentences are more appropriate for the predicates and are also better suited to the context, it is necessary to compare the cross-entropy of the original sentences with those of the swapped ones in the absence of context. For the personal-original pairs, the original sentences (M = 3.21, SD = 0.57) and their impersonal counterparts (M = 3.24, SD = 0.51) did not differ despite our expectations, with a negligible mean difference of −0.03 nats/token (SD = 0.37), $t(49) = -0.50$, p = .310 (one-tailed), $d_z = -0.07$; Wilcoxon V = 557, p = .220. Only 28 of 50 pairs (56%) went in the direction of the original.

For the impersonal-original pairs, by contrast, the original sentences remained more predictable even in isolation (M = 3.13, SD = 0.63) than their personal counterparts (M = 3.30, SD = 0.66), a mean difference of −0.18 nats/token (SD = 0.26), $t(49) = -4.87$, $p < .001$ (one-tailed), $d_z = -0.69$; Wilcoxon V = 174, $p < .001$ (Shapiro-Wilk W = 0.95, p = 0.30); 42 of 50 pairs (84%) are as predicted.

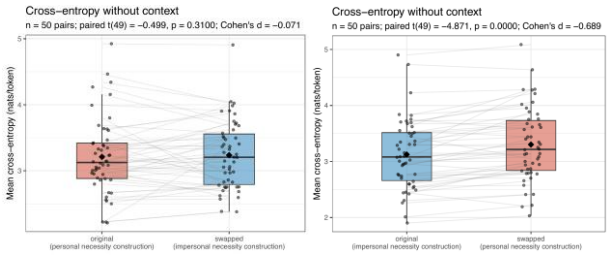
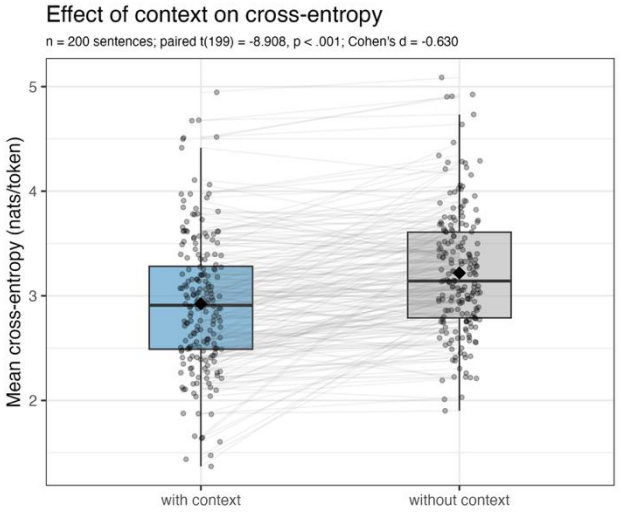


Figure 3: Cross-entropy (nats/token) without context for personal-original pairs (left) and impersonal-original (right), n = 50 pairs each.

This asymmetry is informative in that, for personal-original, the predictability advantage of the attested construction is observed only when the context is available. This indicates that the advantage is genuinely contextual in original rather than a by-product of sentence-internal properties such as lexical frequency. For impersonal-original, part of the advantage is context-independent. That is, verb phrases originally in the impersonal construction yield lower cross-entropy than when swapped with ones in the personal construction, even with no context, indicating each construction preferentially expresses verb phrases that are particularly well suited to it.

4.5 Effects of context

Finally, we examined whether the original corpus contexts generally facilitated prediction of the target sentences. Across all 200 target sentences, cross-entropy was lower when the original corpus context was provided (M = 2.92, SD = 0.65, Mdn = 2.91) than when no context was available (M = 3.22, SD = 0.59, Mdn = 3.14), yielding a mean difference of −0.30 nats/token (SD = 0.47). Although the paired differences deviated from normality (Shapiro-Wilk W = 0.92, $p < .001$), both the paired t-test and the Wilcoxon signed-rank test showed a robust facilitative effect of context, $t(199) = -8.91$, $p < .001$ (one-tailed), $d = -0.63$; Wilcoxon V = 3473, $p < .001$. Context reduced cross-entropy for 152 of the 200 sentences (76%; Figure 4).



The facilitative effect was comparable across the two constructions. For the impersonal construction, cross-entropy was lower with context than without it, $t(99) = -6.08$, $p < .001$, $d = -0.61$, with facilitation observed for 74% of the sentences. A similar pattern was found for the personal construction, $t(99) = -6.57$, $p < .001$, $d = -0.66$, with facilitation observed for 78% of the sentences.

5. Discussion

Overall, the results of our experiment might suggest that the two constructions correspond to different regions in the semantic/functional space, supporting the principle of isomorphism (see 2.1). From the results given in 4.2, 4.3 and 4.5 it follows that each construction can only be properly characterized by its preceding linguistic context (or "cotext"); we could say, in analogy to Firth's (1957) well-cited statement about word meaning, that "a grammatical construction is characterized by the company it keeps."

On the other hand, the results given in 4.4 indicate something intriguing about the relation between the two constructions that may fall outside of the principle of isomorphism; the notable asymmetry observed between the two test results under the "without context" condition is totally unexpected for the principle which only assumes one-to-one mappings between meaning and form. A possible account can be found in the notion of markedness as established in structuralist linguistics; the observed asymmetry can be attributed to the status of the personal necessity construction as an *unmarked* expression of "absence of necessity," which is in contrast to that of the impersonal necessity construction as a marked expression of the notion. If this is the case, attested examples of personal necessity construction may include ones that could have been expressed with the impersonal necessity construction. In those cases, the personal necessity construction will be chosen over the impersonal counterpart *only because* the former is the unmarked option. Examples of the impersonal necessity construction, on the other hand, are not supposed to involve such "both could have worked" cases since the marked item, by definition, needs to be as clearly distinguished from the unmarked one as possible. It could be concluded that the result of the former test (i.e., "personal-original, without context") arguably represents the semantic overlap (i.e., synonymity) between the two constructions.

Taken together, the experiment results given in section 4 crucially point to the possibility that as for near-synonymous constructions, language users look *inside* their internal structures (e.g., predicates and arguments) separately from what is *outside* of them (i.e., context). This means that the relation of synonymy is more complex at the clause-level than it is at the word-level.

Before closing, it should also be stressed that the results given in section 4 would have been almost impossible to produce through any type of qualitative analysis based on speakers' intuitive judgments. In particular, speakers' intuition alone would never have reached what the experiments given in 4.4 obtained, presumably because speakers, when asked to work as "analysts" of the language, are strongly biased to seek differences in meaning that correspond to formal distinctions (cf. the principle of isomorphism; see 2.1). In order to overcome this disadvantage, the present study illustrates, LLMs will be a promising tool, as they are totally free from any such biases.

6. Conclusion and outlook

In this paper, we have presented a case study demonstrating that a synonymous relation between two formally distinct grammatical constructions can be captured through a large language model. It was shown that, while the personal and impersonal necessity constructions are not interchangeable when preceding context is counted, they are partially interchangeable when it is excluded. To highlight the limitation of the method that relies solely on researchers' intuitive judgments, we chose to focus on two constructions that express a modal concept that are not established in the literature as constituting a synonymous pair. To further consolidate our argumentation, however, we will need to see how our method will be applicable to more well-established synonymous pairs of modal expressions.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26K16050. We would like to thank Jon-Patrick Garcia Fajardo for his help with the stylistic improvement of the paper. All usual disclaimers apply.

8. Bibliographical References

Anthony, L. (2025). Stochastic parrots meet corpus linguists: Understanding language in the age of AI. Plenary talk presented at the Corpus Linguistics 2025 conference, Birmingham, UK.

Baayen, R. H. (2024). The wompom. *Corpus Linguistics and Linguistic Theory*, 20(3): 615–648.

Bolinger, D. W. (1977). *Meaning and form*. London: Longman.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Royal Netherlands Academy of Arts and Sciences.

Depraetere, I & Cappelle, B. (2023). English modals: An outline of their forms, meanings and uses. In I. Depraetere, I., Cappelle, B., Hilpert, M., De Cuypere, L., Dehouck, M., Denis, P., Flach, S., Grabar, N., Grandin, C., Hamon, T., Hufeld, C., Leclercq, B., & Schmid, H. (Eds.), *Models of modals* (pp. 14–59). Berlin: Mouton de Gruyter.

Goldberg, A. E. (1995). *Constructions*. Chicago: Chicago University Press.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64(3):501–38.

Fillmore, C. & Kay, P. (1999). Grammatical constructions and linguistic generalization: The *What's X doing Y? construction*. *Language*, 75(1):1–33.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis* (pp. 1–32). Blackwell.

Futrell, R., & Mahowald, K. (2025). *How linguistics learned to stop worrying and love the language models*. arXiv.

Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language*, 56(3):515–540.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Association for Computational Linguistics.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Langacker, R. W. (1999). *Grammar and conceptualization*. Berlin: Mouton de Gruyter.

Levin, B. (1993). *English Verb Classes and Alternations*. Chicago: The University of Chicago Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Morger, F., & Berdicevskis, A. (2026). Not all linguistic variation is equally predictable: Evidence from two case studies with large language models. *International Journal of Corpus Linguistics*, 31(3):333–366.

Perek, F. (2015). *Argument structure in usage-based Construction Grammar: Experimental and corpus-based perspectives*. Amsterdam: John Benjamins.

Pinker, S. (1989). *Learnability and cognition*: The acquisition of argument structure. MIT Press.

Potts, C. (2018). *A case for deep learning in semantics*. arXiv.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI.

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

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S.R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp.38–45). Association for Computational Linguistics.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

9. Language Resource References

Davies, M. (2008–). *The Corpus of Contemporary American English (COCA): 1+ billion words, 1990–2019* [Text edition]