Information-theoretic conditioning in terminological alternations in specialized domains: The cases of Taiwan Mandarin legal language and English biomedical language

Po-Hsuan Huang
Department of Linguistics
University of Southern California
3601 Watt Way, Los Angeles, CA 90089
pohsuan@usc.edu

Hsuan-Lei Shao
Institute of Medical and
Biotechnology Law
Taipei Medical University
301 Yuantong Rd. Zhonghe Dist.
New Taipei City, Taiwan, 235603
hlshao@tmu.edu.tw

Abstract

This study examines how information-theoretic correlates, specifically contextual surprisal, condition terminological alternations in specialized domains, where both domain-specific and general terms express similar concepts. Specifically, two competing theories exist. The Uniform Information Density (UID) theory proposes that the speaker would avoid abrupt information rate changes. This predicts the use of more specific variants when the surprisals are higher. Conversely, availability-based production suggests the use of more readily-accessible items with higher surprisals. This study examines the dynamics between these two potential mechanisms in the terminological use in specialized domains. Specifically, we argue that, in specialized language, due to the higher frequency of domain-specific terms, both accounts predict the use of specific items in higher-surprisal contexts. The cases of Taiwan Mandarin legal language and English biomedical language were, therefore, examined. Crucially, a current popular method for probability estimation is through large language models (LLMs). The linguistic distribution in specialized domains, however, may deviate from the general linguistic distribution on which the LLMs are trained. Thus, we propose a novel semantics-based method of estimating the token probability distribution in a given corpus that avoids the potentially different linguistic distribution and the issue of word segmentation. As expected, results indicated a positive correlation between a variable's surprisal and the use of domain-specific variants in both cases. This supports UID-based production, and arguably also availability-based production, since more specific and frequent variants are preferred in high-surprisal contexts. Specifically, our semantics-based probability estimation outperformed LLM-based estimation and the baseline in both cases. This suggests the feasibility of semantics-based probability estimation in specialized domains.[1]

Keywords: domain-specific variation, information theory, surprisal calculation, semantics

1 Introduction

A growing number of studies have come to emphasize the role of information-theoretic (Shannon, 1948) constraints in communication and the conditioning of these constraints on linguistic distributions. Especially, lexical and syntactic production and processing are attested to be conditioned by information-theoretic correlates, including word frequency and contextual surprisal. Zhan and Levy (2018), for example, investigated the choice of classifiers in Mandarin and found that while frequency did not have an effect, the surprisal of the following noun could predict the language user's choice of classifiers. When the following noun had a higher contextual surprisal, the language user was more likely to opt for the general classifier *ge*, as opposed to the other specific classifiers. Likewise, in Wilcox et al.'s (2023) reading time study across 11 languages, it was found that both contextual surprisal and contextual entropy were positively correlated with the subjects' reading time.

[1]The code implementation of this study is available at: https://github.com/Peh-Suan/information_theoretic_conditioning_domain_specific.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 103–107

1.1 Speaker-centric vs. listener-centric production

Importantly, two competing mechanisms have been put forth. The Uniform Information Density (UID) theory (Levy and Jaeger, 2007; Jaeger, 2010) proposes that during communication, the speaker would prevent abrupt information rate changes to facilitate better speech comprehension. Conversely, a more speaker-centric account, availability-based production (Bock, 1987; Ferreira and Dell, 2000), predicts that the speaker would prefer more readily accessible items. These two mechanisms, therefore, make opposite predictions: While UID would predict the use of more specific variants when the variable is contextually surprising, availability-based production would expect more general items to be used, as they are more accessible than specific ones.

In this study, we examine these two potential mechanisms in the terminological alternations in specialized domains. Specifically, several studies have suggested availability-based production in lexical-syntactic alternations. For example, Zhan and Levy (2018) examined how the contextual surprisal of a noun might influence the use of the general classifier *ge* vs. specific classifiers in Mandarin. It was found that when the noun had a higher surprisal, there was a higher tendency for the speaker to use the general classifier. Likewise, such availability-based production was also attested in Russian comparative constructions (Clark et al., 2022). In Russian, there are two options for comparative construction. The first one is the explicit option, where "than" is used. The other is the genitive option, where the target noun phrase being compared is marked with the genitive case, and "than" is omitted. In the first construction, there is an additional morpheme before going into the target noun phrase, while in the second construction, there is no such buffer. The first construction thus provides a higher availability for the speaker's speech planning. Indeed, it was also found that when the target noun phrase was more complex, the explicit option was preferred.

1.2 Terminological alternations in specialized domains

All the previous studies, however, focused on general language use. It therefore remains unknown whether style differences exist between general and domain-specific language.

Crucially, it is likely that both accounts may favor the domain-specific terms in high-surprisal contexts in domain-specific language. In specialized domains, the same concepts may be expressed through different terms. In English biomedical language, *dermis* or *epidermis* can be used instead of *skin*. Similarly, in Taiwan Mandarin legal language, *zhi.yan.zhi* "in sum" can be used instead of the more colloquial *jian.yan.zhi*.

In the general context, the general terms are without doubt more frequently used. In the specialized domains, however, the respective domain-specific terms may actually be more frequent than the general counterparts. Indeed, in the corpora in this study, the domain-specific terms are 2.18 and 3.27 times more frequent than the general terms in Taiwan Mandarin legal language and English biomedical language, respectively. This, therefore, suggests that both the availability-based production and UID may support the use of domain-specific items when the surprisals are higher.

Therefore, in this study, we examine the information-theoretic conditioning, specifically the effects of surprisal, on terminological alternations in Taiwan Mandarin legal language and English biomedical language.

2 Methods

To answer how contextual surprisal interacts with terminological alternations, two corpora were examined. The contextual surprisals of the terminological variables were calculated based on the popular LLM-based probability estimation and our proposed semantics-based estimation. Linear-mixed effects models were used for statistical analysis.

2.1 Corpora

2.1.1 Taiwan Mandarin legal corpus

The Taiwan Mandarin legal corpus was built from 383,733 legal judgments made in 2024 obtained from the Governmet OpenData platform (`http://data.gov.tw`). Sentence segmen-

tation was performed based on punctuation. 580,593 sentences were collected. 100,000 sentences were then randomly selected as the final corpus.

2.1.2 English biomedical corpus

A subset of the PMC corpus (National Library of Medicine, 2024) was used to build the English biomedical corpus. 1,029,191 sentences were collected. 100,000 sentences were then randomly selected as the final corpus.

2.1.3 Terminological variable selection

The terminological variables were manually inspected and selected by the authors. Only variables with higher frequencies were included. 15 general-vs.-legal and 25 general-vs.-biomedical terminological variables were chosen. An example of such variables is the *skin* vs. *dermis*/*epidermis* alternation mentioned previously. In this example, *skin*, *dermis*, and *epidermis* are all variants of this variable.

2.2 Surprisal estimation

The contextual surprisal of a token w given the context c is $-logP(w|c)$. To calculate a token's contextual surprisal, therefore, its probability in the corpus has to be estimated.

A conventional method of calculating probability is to calculate the raw frequency of the token. This is, however, not ideal for contextual surprisal estimation, since the likelihood of the exact context sentence happening more than once is low.

A more popular alternative is to directly estimate $P(w|c)$ through trained large language models (LLMs). This, however, may also not be ideal since the style differences between general and specialized language may lead to different linguistic distributions.

Therefore, in this study, we propose a novel semantics-based probability estimation based on the "semantic bit count" instead of the raw frequency of the tokens. We propose that, since information-theoretic correlates are essentially based on the amount of information, the semantics of the word token could be more revealing than pure token counts.

2.2.1 Semantics-based probability estimation

In this study, we propose counting a token's semantic bit occurrences in the corpus to estimate the probability of the token. Given the word embedding of a token w, and the embedding of a context sentence c (calculated as the mean of all the token embeddings in the sentence), the number of semantic bits of w in c can be approximated as the cosine similarity between the two vectors.

To convert this cosine similarity to a semantic bit count (sb), it is then rescaled from -1 to 1 to 0 to 1. This semantic bit count is then used instead of the raw frequency. The final semantics-based probability estimation of a token w in the context c is shown in Eq. 1, where C is all the context sentences in the corpus and C_w is all the context sentences where w occurs.

$$\hat{P}_{semantics}(w|c) = \frac{\sum_{c_j \in C_w} sb(w, c_j)}{\sum_{c_i \in C} sb(c, c_i)} \quad (1)$$

To compare the performance of semantics-based surprisal ($I_{semantics}$), LLM-based[2] surprisal (I_{LLM}) and baseline surprisal ($I_{baseline}$), which were calculated with direct 5-gram context counts, were also calculated.

2.3 Statistical analysis

Logistic-mixed effects models (LMMs) were used to test statistical significance through Satterthwaite's method. A model was fitted for each of the three kinds of surprisals for each of the two corpora.

The use of general vs. domain-specific was contrast coded as -0.5 (general) and 0.5 (domain-specific). Surprisal was standardized and taken as the predictor. Standardized frequency was included as a control variable. Random intercepts were grouped by terminological variable.

To compare the performance of the three types of surprisals, Akaike Information Criterion (AIC) was also used to test the relative quality of the fitted models.

3 Results

3.1 Taiwan Mandarin legal language

For both $I_{semantics}$ and I_{LLM}, positive correlations between the use of the domain-specific variants and the variable's contextual surprisal were found ($I_{semantics}$: $\hat{\beta} = 4.37$; $p < 0.001$;

[2]LLAMA-2-7B were used.

I_{LLM}: $\hat{\beta} = 0.11$; $p = 0.03$). On the flip side, $I_{baseline}$ was found to have insignificant effects ($\hat{\beta} = -0.06$; $p = 0.10$).

Crucially, the $I_{semantics}$ model had the lowest AIC ($I_{semantics}$: 6164.80; I_{LLM}: 6185.45; $I_{baseline}$: 6623.24), suggesting it is the most ideal model among the three.

3.2 English biomedical language

Similar positive effects were found between $I_{semantics}$ and domain-specific vs. general term use ($\hat{\beta} = 0.27$; $p < 0.001$). However, a negative correlation between I_{LLM} and domain-specific vs. general term use was found ($\hat{\beta} = -1.51$; $p < 0.001$). On the other hand, $I_{baseline}$ was once again found to have insignificant effects ($\hat{\beta} = 0.02$; $p = 0.48$).

In terms of the model quality based on AIC, the $I_{semantics}$ model was once again suggested to be the most ideal ($I_{semantics}$: 654.47; I_{LLM}: 2100.33; $I_{baseline}$: 5789.62).

4 Discussion

4.1 Information-theoretic conditioning in specialized language and speaker vs. listener-centric production

The main focus of this study is to examine how the style differences between general language and specialized language may interplay with speaker vs. listener-centric production. As discussed in Section 1.1, the UID theory and availability-based production are put forth as two competing mechanisms in previous studies (Zhan and Levy, 2018; Clark et al., 2022). It is suggested that, from a listener-centric perspective, the UID theory would predict more specific language use when the variable is more unpredictable/informative, in order to reduce abrupt information rate changes. From a speaker-based angle, on the other hand, the speaker would prefer more readily accessible variants. This would thus predict the use of more general items, which are presumably more accessible, when the unpredictability is higher.

These studies, however, focus on general language use. We argue that while such competition may hold in general/colloquial language, the two mechanisms may be compatible in specialized language. This is because the domain-specific terms may in fact be more frequent, and thus more accessible, than the general terms in specialized domains. Thus, both accounts would predict the use of domain-specific terms in higher-surprisal contexts, since they are at the same time more informative and readily accessible.

Indeed, the results in this study support our hypothesis. Positive correlations were attested between the semantics-based surprisal and the use of domain-specific terms in both cases. Indeed, opposite effects were found for the LLM-based surprisal, and no effects were found for the baseline surprisal. We argue, however, as will be discussed in the next section, that the semantics-based surprisal is the more appropriate estimation.

4.2 Semantics-based probability estimation for specialized language

The other contribution of this study is the proposal of a novel semantics-based probability estimation for specialized language. As argued in Section 2.2, contextual surprisal cannot be ideally calculated through raw token frequency, nor is it appropriate to use pretrained LLMs, as the linguistic distributions of specialized language may differ from that of general language.

In this study, we propose that the semantics of a token may be more information-theoretically relevant than pure occurrence frequency. A probability estimation based on the semantic bit count of the token was proposed. It was found that in both test cases, our method outperformed the LLM-based method and the baseline. Our results, therefore, suggest the feasibility of semantics-based probability estimation for specialized language in future studies.

Limitations

While the examination of competing surprisal candidates, i.e., the LLM-based surprisal and the baseline, allowed for a general investigation of the performance of our semantics-based method, it remains possible that different LLMs may lead to better or worse performances. In this study, we only examined one LLM (LLAMA-2-7B). To make the findings more grounded, a comparison between our

method with a wider array of models may be ideal.

References

Kathryn Bock. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, 26(2):119–137.

Thomas Hikaru Clark, Ethan Gotlieb Wilcox, Edward Gibson, and Roger P. Levy. 2022. Evidence for availability effects on speaker choice in the Russian comparative alternation. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Victor S. Ferreira and Gary S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 849–856.

National Library of Medicine. 2024. Pubmed central open access subset. `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`. Accessed: 2024-11-01.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.