Learning to Write Rationally: How Information Is Distributed in Non-native Speakers' Essays

Anonymous ACL submission

Abstract

People tend to distribute information evenly in language production for better and clearer communication. In this study, we compared essays written by second language (L2) learners with various native language (L1) backgrounds to investigate how they distribute information in their non-native L2 production. Analyses of surprisal and constancy of entropy rate indicated that writers with higher L2 proficiency can reduce the expected uncertainty of language production while still conveying informative content. However, the uniformity of information distribution showed less variability among different groups of L2 speakers, suggesting that this feature may be universal in L2 essay writing and less affected by L2 writers' variability in L1 background and L2 proficiency.

1 Introduction

With increased globalization, more people have started acquiring and using multiple languages. For instance, the proportion of individuals who speak multiple languages daily in the United States has doubled over the past four decades, rising from about one in ten speakers to about one in five (Dietrich et al., 2022). These rapid changes in linguistic diversity offer unique opportunities, but also present challenges: Not all speakers achieve perfect or proficient levels in their non-native languages (L2s) due to various factors, including the quantity and quality of exposure to L2s, the duration and nature of their acquisition process, and their prior language experiences and native language (L1) backgrounds. The language processing mechanisms of multilingual speakers may differ from those of native (monolingual) speakers, not only due to variations in proficiency but also because of diverse language backgrounds and experiences.

The cognitive mechanisms underlying multilingual language processing represent a vibrant research topic spanning multiple fields, including psychology, linguistics, cognitive sciences, and artificial intelligence. Many previous studies have explored whether and how speakers with different language backgrounds comprehend and produce languages differently, using various approaches (e.g. Bernolet et al., 2007; Hartsuiker et al., 2016; Hsiao and Gibson, 2003 for behavioral studies, and Gries and Kootstra, 2017; Putnam et al., 2018 for corpus-based studies). Most of these studies have reached a similar conclusion: the multiple language systems of multilingual speakers are highly interactive, and phonological, lexical, and syntactic representations are integrated across languages. Consequently, multilingual speakers can't just turn off the other language(s) when they use a particular language. This other language(s) can influence the comprehension and production processes of the language currently in use, leading to unique patterns in target language processing that can reveal information and knowledge from other languages.

Even though there are variations in language production among multilingual speakers, the goal remains the same: to deliver information effectively and efficiently. To achieve this goal, people distribute information evenly across language production, maintaining relatively equal predictability for each upcoming word. More specifically, the information carried by a unit of production can be quantified by several features, such as surprisal (Shannon, 1948), entropy (Shannon, 1948; Genzel and Charniak, 2002), and the uniformity of information distribution (Frank and Jaeger, 2008). Using these features, the goal of language production can be described by the following principles:

- The surprisal effect (Levy, 2008): Processing unexpected information in the produced signal takes longer.
- The constancy of entropy rate (ERC, Genzel and Charniak, 2002): The rate of information

transmitted in a produced unit remains relatively constant across language production.

- The uniform information density theory (UID, Frank and Jaeger, 2008): People prefer to avoid sudden and rapid changes in information density by evenly distributing information across language production.

Numerous empirical studies substantiated these principles. For instance, people need longer time to process unexpected words during comprehension (e.g. Smith and Levy, 2013; Wilcox et al., 2023) and speakers maintain uniformity of information and constancy of predictability by selecting shorter words (e.g. Mahowald et al., 2013), repetitive/familiar syntactic structures (e.g. Xu and Reitter, 2016, 2018), or faster speech rate (e.g. Priva, 2017). The surprisal effect can also be observed in cross-lingual production: multilingual speakers will switch languages to avoid uncommon words in their production that will take longer to process (Calvillo et al., 2020).

While numerous studies, including those mentioned above, have explored how individuals use these rules to enhance language production, how L2 speakers acquire and utilize those rules to distribute information in their L2 production remains under-researched. Considering that L2 speakers' preferences in lexical selection and syntactic structures can differ from native speakers and can vary based on their L1 backgrounds, we hypothesize that L2 production varies across multilingual speakers. In this study, we employ well-attested features from psycholinguistics and information science to examine how L2 speakers of English with diverse native language (L1) backgrounds and varying levels of L2 English proficiency distribute information in their written English output.

2 Method

2.1 Corpus and data pre-processing

We used the TOEFL11 corpus (Blanchard et al., 2013) for this study. The TOEFL11 corpus contains written essays from actual TOEFL exam takers from 11 different L1 backgrounds which are from 7 language families; speakers are grouped into 3 proficiency groups based on their essay scores. Detailed information is presented in Table 1. We also included essays written by native English speakers from the ICNALE corpus[1] (Ishikawa, 2013) as native-like information distribution patterns. This inclusion helps in understanding whether and how information distribution varies with changes in speakers' L2 proficiency and L1 backgrounds. Due to the size of the dataset and shorter essay length in the low proficiency group and the native speaker group, only the first 300 tokens in each essay were used for position-based analyses.

[1]The ICNALE corpus: http://language.sakura.ne.jp/icnale/

2.2 Information-based feature extraction

To extract information features, corpus-based studies typically analyze the information and language resources within the target corpora. However, since the TOEFL11 corpus consists entirely of non-native speakers' language production, using this method for extracting information features potentially introduces biases toward non-native-like syntactic structures or lexical selections. To minimize such biases, we extracted information features using pre-trained large language models (LLMs), as these models are more robust and generalized due to their extensive and diverse corpora resources.

We extracted three widely used information-based features (Frank and Jaeger, 2008; Genzel and Charniak, 2002; Wilcox et al., 2023) as follows: First, we converted each essay into tokens and obtained the conditional probability p for each token w using a pre-trained LLM (GPT-2, Radford et al., 2019). We then converted the probability sequences into the following features:

- **Surprisal**: The surprisal feature (Shannon, 1948; Wilcox et al., 2023) measures how much information a signal carries. Given the context history (C), the surprisal of the i-th token is calculated as:

$$S_i = -log_2(p(w_i|C_{t<i})) \tag{1}$$

  In our study, surprisal measures how unpredictable the exact token is given the previous context. The lower surprisal value indicates a more predictable upcoming word.

- **Entropy**: The entropy feature measures the expected predictability of the upcoming token (Shannon, 1948) through the following equation, given the history of context C.

$$H_i = -\sum_{w \in vocab}(p(w|C_{t<i})log(p(w|C_{t<i}) \tag{2}$$

  Unlike surprisal, entropy calculates the expected predictability of the next word before

Language family	Language(s)	Number of essays[a]	Mean (SD) of essay length[b]
Afro-Asiatic	Arabic	274, 545, 181	342.72 (96.56)
Altaic	Japanese, Korean, Turkish	434, 1795, 771	355.42 (94.62)
Dravidian	Telugu	86, 595, 319	417.42 (94.03)
Germanic	German	14, 371, 615	391.23 (73.01)
Indo-Iranian	Hindi	25, 399, 576	418.04 (88.62)
Romance	French, Italian, Spanish	278, 1597, 1125	365.07 (79.33)
Sino-Tibetan	Chinese	90, 662, 248	384.53 (84.17)

[a]*of low, medium, and high proficiency speakers.* [b]*mean (SD) of native speakers: 250.72 (30.92).*

Table 1: Corpus description.

it is produced. Therefore, a lower value indicates higher certainty in the selection of the next word.

- **UID score**: Given the language production y, the UID score measures the variance of the surprisal, representing how uniformly information is distributed across the language production.

$$UID(y) = \frac{1}{|y|} \sum_i (y_i - \overline{y})^2 \qquad (3)$$

Based on this equation, a signal with a perfectly even distribution of information receives a 0 UID score.

For surprisal and entropy features, both token-based values and document-based mean values were extracted for further analysis.

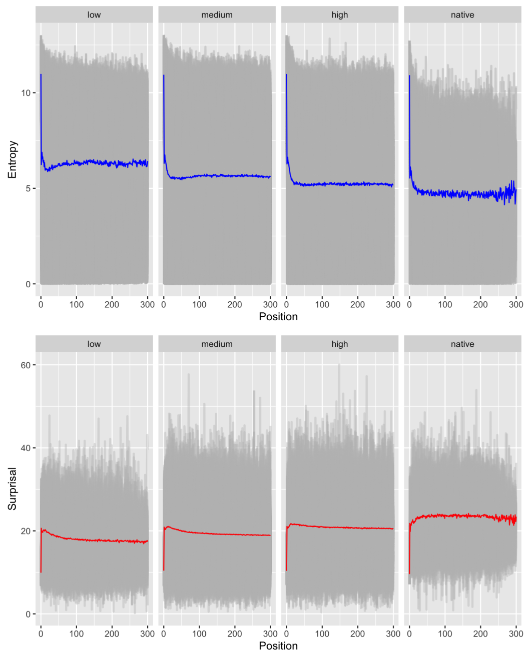


Figure 1: Entropy (top) and surprisal (bottom) against token position, group by speaker proficiency. Shaded area: actual entropy/surprisal values.

3 Results

3.1 Proficiency vs. information distribution

We fitted two linear mixed-effect models using token-based surprisal and entropy as response variables, token positions and proficiency as fixed effects, and individual essays as random effects. We observed a trend towards more native-like patterns, with decreasing entropy values and increasing surprisal values in position-based results as the speaker's proficiency increases (see Figure 1 & Table 2). Such a pattern was also observed in the following document-level analysis (see Figure 2). These findings indicate the significance of L2 proficiency in predicting how native-like the information distribution pattern is in L2 production: a higher L2 proficiency is associated with lower uncertainty, but a higher level of informative content.

3.2 L1 background vs. information distribution

Using only L2 speakers' data and document-based features, a one-way ANOVA analysis indicated a significant effect of L1 backgrounds on mean surprisal, $F(10, 10989) = 143.1^{***}$, mean entropy, $F(10, 10989) = 82.14^{***}$, and UID, $F(10, 10989) = 28.22^{***}$ (*** indicates $p < 0.001$). These differences were also observed when controlling for proficiency (see Figure 2), indicating that speakers' information distribution patterns are influenced by L1 background. Table 3 summarized the number of significant pairs regarding all three features mentioned above. Medium-proficient L2 speakers show the largest variation in distributing information, while low-proficient speakers have the least

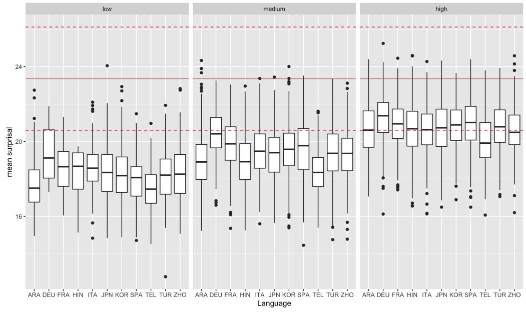
Proficiency	Surprisal	Entropy
low	-3.974***	1.256***
medium	-2.739***	0.696***
high	-1.703***	0.391***

***$p < 0.001$

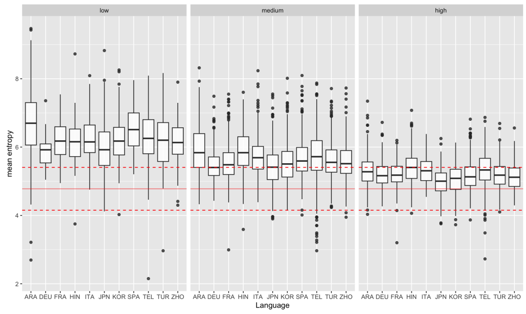
Table 2: β values of proficiency (native speakers as reference level) of linear mixed effects models.

Proficiency	Surprisal	Entropy	UID
low	14	13	14
medium	40	35	26
high	23	36	9

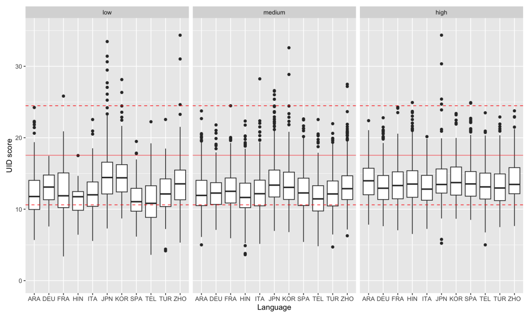
Table 3: Numbers of significant pairs of group differences in post hoc ANOVA analysis.



(a) Mean surprisal



(b) mean entropy



(c) UID score

Figure 2: Boxplots of information features among non-native speakers' essays. Red lines indicate the mean and 95% distribution among native speakers.

variation. A further discussion of this pattern follows in the next sections.

4 Discussion

This study explored how multilingual speakers with different L1 backgrounds distribute information in their L2 written production. Our results revealed more "native-like" trends in surprisal and entropy as the speakers' L2 proficiency increased. In contrast, the UID score indicated that all multilingual speakers tend to hold the fundamental principles of information distribution in their L2 writing, even when they are less proficient in L2. These results provide additional insights regarding specific effects of L2 proficiency on L2 speakers' language production and communication.

Language surprisal and entropy emphasize incoming production from different perspectives: Surprisal measures the exact information carried by the incoming word, while entropy estimates the expected certainty about upcoming words. As shown in Figure 1, native speakers seek to maximize the information in each word (surprisal) while minimizing the overall expected uncertainty (entropy) for effective and clearest communication. As shown in our analyses of surprisal and entropy features, as L2 speakers' proficiency in a second language increases, they develop more native-like language production. Presumably, they have more L2 resources, which further lead to more advanced, sophisticated, and coherent lexical selection, longer production units, and more complex syntactic structures in their L2 production (Crossley, 2020; Lu, 2010, 2011). Our analyses of information distribution among L2 speakers further support this by showing that higher L2 proficiency enables learners to produce language more effectively and efficiently by carrying more information and reducing expected uncertainty in their production.

Even though we observed significant group differences in mean surprisal and entropy measures among speakers with different L2 proficiency levels and L1 backgrounds, the UID scores showed a slightly different pattern with fewer variations and a more native-like distribution across all L2 proficiency groups (see Figure 2c and Table 3). Since UID is associated with the variance of surprisals in language production, the UID score suggests that the ability to distribute information evenly might be a generalized effect across L2 speakers, regardless of their L1 background and L2 proficiency in the target language.

5 Limitations

Our study is among the first to explore surprisal, entropy, and uniform information density in L2 English writing in a large group of L2 English speakers with a wide variety of L1 backgrounds and with varying levels of L2 English proficiency. Here we outline several limitations of the present work and directions for future research.

Firstly, the dataset contained only basic information regarding speakers' language background and experience. The only information available in the TOEFL11 dataset is the speakers' L1. Other crucial details, such as the frequency of L2 usage, duration of L2 acquisition, and the amount of exposure to language(s) other than their L1 and L2 English, are missing. This lack of information restricts the analysis and discussions of underlying causes of the observed variations within each subgroup in the data set, making it challenging to deeply investigate the diversity of language production. Future studies may use datasets that include more details regarding language history and the L2 acquisition process to further explore variations in speakers' language production and information distribution patterns.

Secondly, we only applied informatics features at the document level, which may underestimate local changes and fluctuations in information distribution. Document-level features can also ignore or underestimate the impact of production length, as longer texts may exhibit larger variations in information density due to the larger number of produced words. In our study, we addressed this issue by analyzing language production within a finite length in some models, but this method involves a hard slicing of language production, potentially leading to incomplete representations of information density distribution. Future studies could address this issue by analyzing shorter production units, such as sentences or paragraphs, to better investigate how information is distributed among L2 learners' written production.

Lastly, our work focused on computational-based features (surprisal, entropy, and UID) and we did not examine more traditional linguistic features, such as specific syntactic constructions. Research has shown that for better communication, speakers select specific types of lexical items and syntactic structures when producing languages (e.g. Xu and Reitter, 2016). In the L2 acquisition process, as proficiency increases, learners have more language resources available to produce language, which leads to more complex, richer, and more appropriate lexical selections and syntactic structures in their language production (e.g. Crossley, 2020; Lu, 2011). For a more complete and detailed understanding of L2 speakers' acquisition and language production, future studies could examine the relationships among computational linguistics features and traditional linguistic features.

References

Sarah Bernolet, Robert J Hartsuiker, and Martin J Pickering. 2007. Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):931.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Jesús Calvillo, Le Fang, Jeremy Cole, and David Reitter. 2020. Surprisal predicts code-switching in chinese-english bilingual text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4029–4039.

Scott A Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Sandy Dietrich, Erik Hernandez, et al. 2022. Language use in the united states: 2019. *American Community Survey Reports*.

Austin F Frank and T Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.

Stefan Th Gries and Gerrit Jan Kootstra. 2017. Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, 20(2):235–250.

Robert J Hartsuiker, Saskia Beerts, Maaike Loncke, Timothy Desmet, and Sarah Bernolet. 2016. Cross-linguistic structural priming in multilinguals: Further evidence for shared syntax. *Journal of Memory and Language*, 90:14–30.

Franny Hsiao and Edward Gibson. 2003. Processing relative clauses in chinese. *Cognition*, 90(1):3–27.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, 1:91–118.

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

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, 45(1):36–62.

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Uriel Cohen Priva. 2017. Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, 160:27–34.

Michael T Putnam, Matthew Carlson, and David Reitter. 2018. Integrated, not isolated: Defining typological proximity in an integrated multilingual architecture. *Frontiers in psychology*, 8:2212.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.