

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

Zixin Tang¹, Janet G. van Hell^{2,3}

¹College of Information Sciences and Technology,

²Department of Psychology, ³Center for Language Science
The Pennsylvania State University
{zxtang, jvg3}@psu.edu

Abstract

People tend to distribute information evenly during language production, such as when writing an essay, to improve clarity and communication. However, this may pose challenges to non-native speakers. 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 written essays. We used information-based metrics, *i.e.*, word surprisal, word entropy, and uniform information density, to estimate how writers distribute information throughout the essay to deliver information. The surprisal and constancy of entropy metrics showed that as writers' L2 proficiency increases, their essays show more native-like patterns will be in the essay, indicating more native-like mechanisms in delivering informative but less surprising content. In contrast, the uniformity of information density metric showed fewer differences across L2 speakers, regardless of their L1 background and L2 proficiency, suggesting that distributing information evenly is a more universal mechanism in human language production mechanisms. This work provides a computational approach to investigate language diversity, variation, and L2 acquisition via human language production.

1 Introduction

With the progress of globalization, more people have started acquiring new 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 for the multilingual population: 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 (Leow, 1998), the length and styles of their acquisition process (Legault et al., 2019), and their native language (L1) backgrounds and experiences (Zdorenko and Paradis, 2012). The cognitive mechanisms underlying language use in multilingual speakers may differ from those of native speakers, not only due to variations in proficiency but also because of diverse language backgrounds and experiences (Bates and MacWhinney, 1989; Hernandez et al., 2005).

Many previous studies have explored whether and how speakers with different language backgrounds comprehend and produce languages differently. For example, Spanish-English speakers may produce “Spanish-like” sentences in their English production, where such types of grammar are rarely used or even prohibited in English. Most of these studies have reached a similar conclusion: for multilingual speakers, representations are integrated across languages, forming a unified system for human language processing (Putnam et al., 2018; Hartsuiker et al., 2004). Consequently, for individuals who know more than one language, the language(s) that are not seemingly involved in the target language production task, can also contribute to and influence comprehension and production processes in the target language, leading to unique patterns in human language processing that can reveal information and knowledge from other languages.

Despite variations in language production among multilingual speakers, the overarching goal of speaking and writing remains the same: to deliver information effectively. To achieve this goal, people distribute information evenly across language production, maintaining relatively equal predictability for each upcoming word (Genzel and Charniak, 2002; Frank and Jaeger, 2008; Meister et al., 2021). Furthermore, the information carried by a unit of production (*e.g.*, a word) can be quantified in several ways, including surprisal (Shannon,

1948), entropy (Shannon, 1948; Genzel and Charniak, 2002), and the uniformity of information distribution (UID) (Frank and Jaeger, 2008; Meister et al., 2021). These metrics help characterize the underlying rules of human language production, which can be summarized as follows:

- **Surprisal Effect:** Processing unexpected information in the produced signal takes longer.
- **Entropy Rate Constancy (ERC):** The rate of information transmitted in a produced unit remains relatively constant across language production.
- **Uniform Information Density (UID):** People prefer to avoid sudden and rapid changes in information density by evenly distributing information across language production.

These rules have been substantiated by a wealth of empirical studies. For instance, people need longer time to process unexpected words during comprehension (Smith and Levy, 2013; Wilcox et al., 2023); during production, people maintain uniformity of information and constancy of predictability by selecting shorter words (Mahowald et al., 2013), repetitive/familiar syntactic structures (Xu and Reitter, 2016, 2018), or faster speech rate (Priva, 2017). Using information-based metrics, prior studies also explored how the complexity of language production changes across language acquisition, and whether we can predict learners' proficiency based on those changes (Kharkwal and Muresan, 2014; Sánchez et al., 2024; Sun and Wang, 2021).

What remains unknown, despite numerous studies exploring how individuals use these rules to enhance language production, is how L2 speakers apply these rules to distribute information in their L2 production—a topic that remains under-researched. Given that L2 speakers often exhibit different preferences in lexical selection and syntactic structures compared to native speakers (Hartsuiker et al., 2004; Van Gompel and Arai, 2018)—variations influenced by their language backgrounds—it is reasonable to assume that these differences may result in distinct patterns in their L2 output. In this paper, we use several well-established metrics from psycholinguistics and information science to investigate how speakers with diverse L1 backgrounds and varying levels of L2 proficiency distribute information in their written production.

2 Related Work

The cognitive mechanisms underlying multilingual language processing represent a significant research topic spanning multiple fields, including psychology (Kroll and De Groot, 2009; Schwiter, 2015), linguistics (Bhatia and Ritchie, 2014), and cognitive neuroscience (Morgan-Short and van Hell, 2023; van Hell, 2023). As an integrated mechanism covering multiple languages, multilingual speakers demonstrate several typical cognitive and language patterns, such as cross-lingual priming effect (Hartsuiker et al., 2004; Sung et al., 2016), cross-lingual cognate effects (Dijkstra et al., 2019), and code-switching effect (Green and Wei, 2014). Other studies explored how multilingualism impacts general cognitive capabilities and neural structures (Baum and Titone, 2014; Birdsong, 2018). Recently, studies also started involving artificial intelligence to explore multilingualism and potential applications for multilingual populations (Zhai and Wibowo, 2023), which provides new opportunities to explore and simulate multilingual processes and potential new methods for language education and proficiency assessment.

While some prior studies take an information-based approach to investigate human language production, few of them specify the nature of their multilingual sample. Some studies offer intriguing evidence regarding cross-lingual production, such as the observation that multilingual speakers switch languages to avoid using uncommon words, demonstrating the surprisal effect (Calvillo et al., 2020). Specifically, bilingual speakers are more likely to switch languages when the coming words are difficult to predict, leading to a reduction of information density (Myslín and Levy, 2015). Even though some previous works proposed that different mechanisms may exist to help L2 speakers better deliver information in communication (Costa et al., 2008), details regarding these mechanisms remain under-researched.

3 Methods

3.1 Materials and Models

Corpus. 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. Each L1 category has 1,000 essays, making a total of 11,000 essays in the corpus. Speakers are grouped into 3 proficiency groups based on their

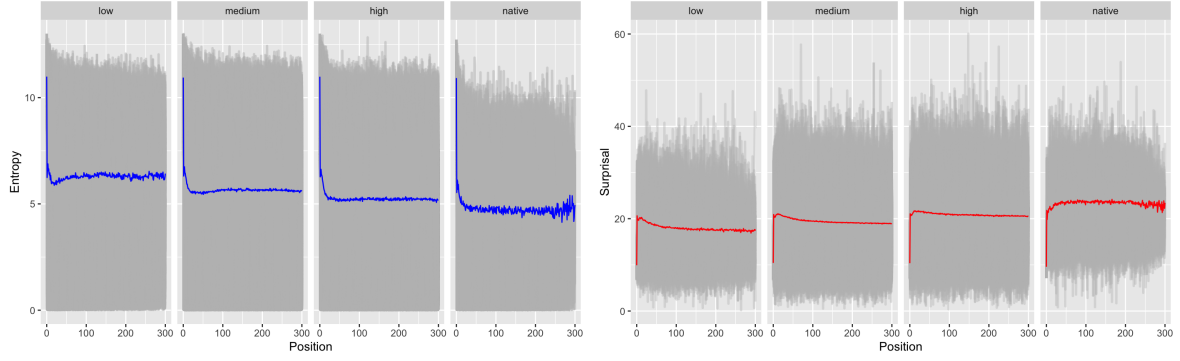


Figure 1: Entropy (left) and surprisal (right) values within written essays, categorized by speaker proficiency. The mean values of both metrics are represented by lines.

essay scores. Detailed information can be found in Appendix A.1. Since native English speakers do not typically take the TOEFL exam, we also included 400 essays written by native English speakers from the ICNALE corpus (Ishikawa, 2013), which is fewer than any group of L2 learners in the TOEFL11 dataset (Blanchard et al., 2013). We specifically selected this dataset as a comparison due to its similar setup and data collection process as TOEFL11 corpus: the essays in ICNALE corpus are short essays related to discussion-based topics, written within a short time (20-40 minutes). Given the similar setup and nature of the written instructions, we used these native speakers’ essays to illustrate 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.

Model. Previous 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’ written essays, using this method for extracting information measures potentially introduces biases toward non-native-like syntactic structures or lexical selections. To minimize such biases, we extracted information metrics using pre-trained large language models (LLMs), as these models provide more general and universal estimation regarding tokens’ conditional probabilities. In this study, we used GPT-2 (Radford et al., 2019) to tokenize the original essays and convert token-based probability sequences. We selected GPT-2 as it is an open-access language model without a usage limit. Since GPT-2 is trained based on large-scale web-based materials, it provides a convenient process in capturing general language probability distribution patterns of mainstream lan-

guage users (English native speakers in our case). Because of its openness and transparency, GPT-2 has been used to investigate biases in text generation (Narayanan Venkit et al., 2023) and information distribution patterns on natural language generation (Venkatraman et al., 2024). Other studies also involved the probability sequences from GPT-2 to predict human behavioral performance (Shain et al., 2024; Oh and Schuler, 2022) and neural behaviors (Michaelov et al., 2022; Goldstein et al., 2022).

Data Pre-Processing. Using the GPT-2 model, we first tokenized the original essays using the GPT-2 tokenizer. The statistic description can be found in Appendix A.1. Each essay had 2 token-based metrics and 3 essay-level metrics to represent the information distribution patterns, detailed extraction processes are introduced in Section 3.2. Due to the shorter native speakers’ essays (250 words, see Table A.1) and the positively skewed distribution of essay length in the TOEFL11 corpus, the token-based sequences included the first 300 tokens in each essay to balance data sparsity, maintain data completeness, and eliminate less reliable results.

3.2 Information-Based Metrics

We extracted five metrics from three widely used information-based metrics as follows. First, using the token sequences, we obtained the conditional probability $p(w|C)$ for each token w given all previous context C , using GPT-2 (Radford et al., 2019). We then calculated three following three metrics using the probability sequences:

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

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

In our study, surprisal measures the information density of each token, given the previous context: a lower value indicates a more predictable word. In this study, the surprisal sequence of the first 300 tokens and the mean value of surprisal among all tokens in each essay are extracted as two measures.

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

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

Unlike surprisal, entropy calculates the average expectancy of the next word before it is produced: a lower value represents a higher certainty regarding the upcoming word. In this study, the entropy sequence of the first 300 tokens and the mean value of entropies among all tokens in each essay are extracted.

- **UID score:** Following previous work on information distribution (Frank and Jaeger, 2008; Meister et al., 2021), the UID score is measured as the variance of token surprisal, which indicates the information density of each token in the essay. Given the human written production y , the UID score represents how uniform the information is distributed across the written production.

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

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

4 Results

4.1 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 (Figure 1 and Table 1). Such a pattern was also observed in the following essay-level analysis (Figure 2). These

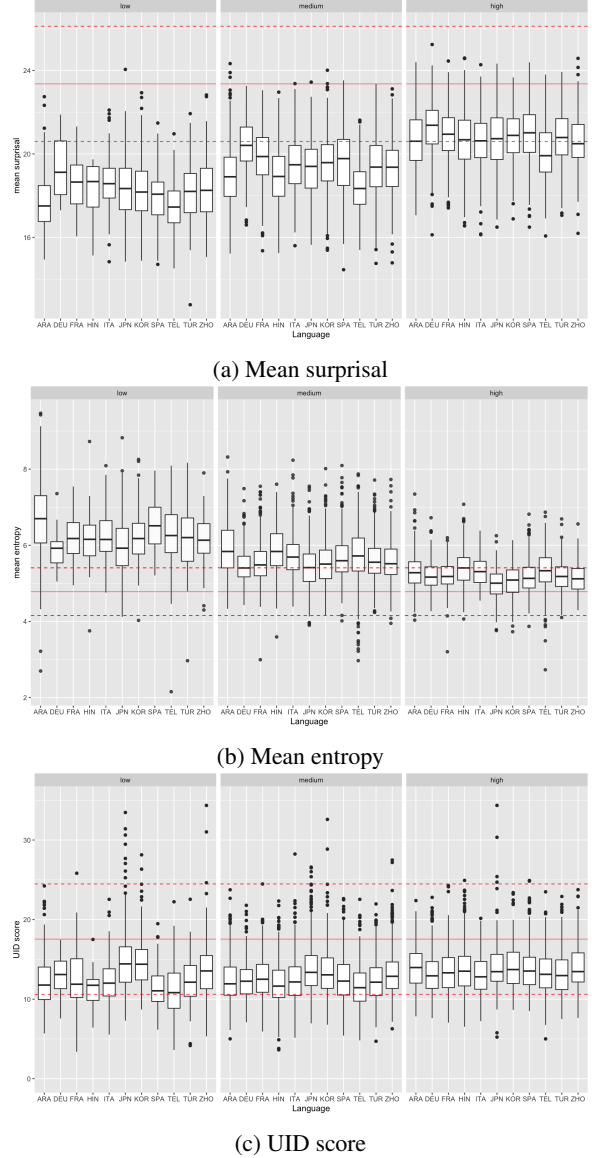


Figure 2: Boxplots of information metrics among non-native speakers’ essays. Red lines indicate the mean and 95% distribution among native speakers.

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.

Due to the lack of predictive power ($\eta = 0.07$), there are no significant differences in UID scores regarding speakers with different proficiency levels. Such a pattern can also be observed in Figure 2, and will be further discussed in Sec. 5.

4.2 L1 Background vs. Information Distribution

Using only L2 speakers’ data and essay-based metrics, a one-way analysis of variance (ANOVA) in-

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

*** p -value < 0.001

Table 1: β values of proficiency (native speakers as reference level) of linear mixed effects models.

Proficiency	Surprisal	Entropy	UID
low	9.37	13.69	12.11
medium	70.57	34.17	21.74
high	21.74	26.64	4.89

Table 2: F-scores regarding each metric in ANOVA analysis with proficiency control.

icated a significant effect of L1 backgrounds on (***) indicates $p < 0.001$):

- **Mean surprisal**, $F(10, 10989) = 143.1^{***}$,
- **Mean entropy**, $F(10, 10989) = 82.14^{***}$, and
- **UID**, $F(10, 10989) = 28.22^{***}$.

These effects remained significant when controlling for proficiency (Figure 2), indicating that speakers’ information distribution patterns are influenced by their L1 background.

Controlling for proficiency, Table 2 summarized the variations of essay-level metrics as F-scores in ANOVA analysis. Medium-proficient L2 speakers show the largest variation in distributing information in terms of all three metrics, while UID showed less variations compared to the other two metrics (Table 2). This pattern is further discussed in the following sections.

5 Discussion

This study explored how speakers with different L1 backgrounds distribute information in their L2 written production. Our results revealed more “native-like” trends in metrics such as surprisal and entropy as the speakers’ L2 proficiency increased. In contrast, metrics such as the UID score indicated that L2 writers tend to adhere to the fundamental principles of information distribution, even when they are less proficient in L2. These results provide additional insights regarding the learning progress among L2 speakers in language production and communication.

Language surprisal and entropy emphasize language production from different aspects: Surprisal measures the exact information carried by the in-

coming word, while entropy estimates the expected certainty about upcoming words. As demonstrated by native speakers in Figure 1, speakers want to maximize the information in each word while minimizing the overall expected uncertainty for effective and clearest communication. As learners’ proficiency in L2 increases, they develop more native-like language production. With increased L2 proficiency, 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 production outcomes (Crossley, 2020; Lu, 2010, 2011). Our results provide additional insights through the information distribution among L2 speakers, 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 scores among L2 speakers with different L2 proficiency levels and L1 backgrounds, the UID scores showed a different pattern with fewer variations and a more native-like distribution across all proficiency groups (see Figure 2c and Table 2). Since UID is associated with the variance of surprisal in language production, the UID variations might indicate that the ability to distribute information evenly might be acquired and generalized as a universal production skill across languages, regardless of how proficient a speaker is in the target language.

6 Conclusion and Future Work

This paper studies how information is distributed in written essays from native and non-native English speakers using information-based metrics. The increasing surprisal and decreasing entropy values showed that proficient L2 speakers distribute information in a more native-like style by maximizing the usage of information channels while reducing the uncertainty of upcoming words. In contrast, the UID score showed fewer differences among proficiency groups, indicating that maintaining smooth communication channels is a more general skill among human language users. Future studies can investigate the relationship between linguistic features and information-based metrics regarding speakers’ language production, as well as how prior language experiences impact the information distribution patterns.

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 provide 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 investigate the diversity of language production in depth. We also only explored the information distribution patterns across L2 English learners' written products, which may restrict the generalizability when dealing with languages from other language families. Future studies may use datasets that include more details regarding language history and the L2 acquisition process, and/or corpora in other languages, to further explore variations in speakers' language production and information distribution patterns and to better understand the language learning trajectories and language representations in multilingual speakers.

Secondly, our metric calculations may underestimate local changes and fluctuations in information distribution. The essay-level metrics can 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 the first 300 tokens in the essays for position-based models. However, this method has a hard cut-off of the essays, 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.

Thirdly, this study assumes that the probability sequences estimated by LLMs can represent human-like psycholinguistics patterns, which is

supported by several studies (Michaelov et al., 2022; Goldstein et al., 2022; Michaelov et al., 2024). However, several studies showed that LLMs may not directly represent humans' mechanisms regarding language comprehension (McCoy, 2019; Oh et al., 2022; Oh and Schuler, 2023). The differences in "language acquisition" processes between humans and machines can lead to fundamental differences in language representations and mechanisms, even if their final outputs appear similar. Future studies should further investigate the differences in language representations and mechanisms across humans and machines, and examine how such differences can impact the usage of modern computational models in traditional language science research areas.

Lastly, our work focused on computational-based metrics (surprisal, entropy, and UID) and we did not examine more traditional linguistic features, such as specific syntactic constructions. Research has shown that to maintain UID, speakers select specific types of lexical items and syntactic structures when producing languages (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 (Crossley, 2020; Lu, 2011). Future studies could examine the relationships between computational linguistics metrics and traditional linguistic features for a more complete and detailed understanding of L2 speakers' acquisition and language production.

Acknowledgements

This study is supported by the National Science Foundation (DGE NRT 2125865 and BCS 1734304). We thank Ting-Hao 'Kenneth' Huang for his generous support and suggestions on revising the manuscript. We also thank the anonymous reviewers and the meta reviewer for their helpful comments on the earlier draft.

References

- Elizabeth Bates and Brian MacWhinney. 1989. Functionalism and the competition model. *The Crosslinguistic Study of Sentence Processing*, 3:73–112.
- Shari Baum and Debra Titone. 2014. Moving toward a neuroplasticity view of bilingualism, executive con-

- trol, and aging. *Applied Psycholinguistics*, 35(5):857–894.
- Tej K Bhatia and William C Ritchie. 2014. *The handbook of bilingualism and multilingualism*. John Wiley & Sons.
- David Birdsong. 2018. Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, 9:81.
- 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.
- Albert Costa, Martin J Pickering, and Antonella Sorace. 2008. Alignment in second language dialogue. *Language and Cognitive Processes*, 23(4):528–556.
- 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*.
- TON Dijkstra, Alexander Wahl, Franka Buytenhuijs, Nino Van Halem, Zina Al-Jibouri, Marcel De Korte, and Steven Rekké. 2019. Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4):657–679.
- 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.
- Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nas-tase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.
- David W Green and Li Wei. 2014. A control process model of code-switching. *Language, Cognition and Neuroscience*, 29(4):499–511.
- Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414.
- Arturo Hernandez, Ping Li, and Brian MacWhinney. 2005. The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, 9(5):220–225.
- 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.
- Gaurav Kharkwal and Smaranda Muresan. 2014. Surprisal as a predictor of essay quality. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–60.
- Judith F Kroll and Annette MB De Groot. 2009. *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.
- Jennifer Legault, Jiayan Zhao, Ying-An Chi, Weitao Chen, Alexander Klippel, and Ping Li. 2019. Immersive virtual reality as an effective tool for second language vocabulary learning. *Languages*, 4(1):13.
- Ronald P Leow. 1998. The effects of amount and type of exposure on adult learners’ l2 development in sla. *The Modern Language Journal*, 82(1):49–68.
- 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.
- RT McCoy. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.
- Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.
- James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, 5(1):107–135.
- James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2022. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, 15(3):1033–1042.

- Kara Morgan-Short and Janet G van Hell. 2023. *The Routledge Handbook of Second Language Acquisition and Neurolinguistics*. Taylor & Francis.
- Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, pages 871–905.
- Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. [Nationality bias in text generation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.
- Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.
- Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. *arXiv preprint arXiv:2212.11185*.
- Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.
- 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.
- Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024. Harnessing gpt to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 414–427.
- John W Schwieter. 2015. *The Cambridge handbook of bilingual processing*. Cambridge University Press.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- 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.
- Kun Sun and Rong Wang. 2021. Using the relative entropy of linguistic complexity to assess l2 language proficiency development. *Entropy*, 23(8):1080.
- Yao-Ting Sung, Jung-Yueh Tu, Jih-Ho Cha, and Ming-Da Wu. 2016. Processing preference toward object-extracted relative clauses in mandarin chinese by l1 and l2 speakers: an eye-tracking study. *Frontiers in Psychology*, 7:4.
- Roger PG Van Gompel and Manabu Arai. 2018. Structural priming in bilinguals. *Bilingualism: Language and Cognition*, 21(3):448–455.
- Janet G van Hell. 2023. The neurocognitive underpinnings of second language processing: knowledge gains from the past and future outlook. *Language Learning*, 73(S2):95–138.
- Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. Gpt-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115.
- 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.
- Tatiana Zdorenko and Johanne Paradis. 2012. Articles in child l2 english: When l1 and l2 acquisition meet at the interface. *First Language*, 32(1-2):38–62.
- Chunpeng Zhai and Santoso Wibowo. 2023. A systematic review on artificial intelligence dialogue systems for enhancing english as foreign language students’ interactional competence in the university. *Computers and Education: Artificial Intelligence*, 4:100134.

A Appendix

A.1 Corpus Description

We included the TOEFL11 corpus (Blanchard et al., 2013) and 400 native speakers’ essays from the ICNALE corpus (Ishikawa, 2013) for this study. The detailed information regarding the dataset is listed below, where essay length is measured as GPT-2 tokens.

Language	Language family	Portion of essays ^a	Mean (SD) of essay length ^b
Arabic	Afro-Asiatic	0.274, 0.545, 0.181	341.87 (95.21)
German (DEU)	Germanic	0.014, 0.371, 0.615	392.06 (73.51)
French	Romance	0.060, 0.526, 0.414	372.04 (78.23)
Hindi	Indo-Iranian	0.025, 0.399, 0.576	417.42 (86.96)
Italian	Romance	0.145, 0.569, 0.286	340.37 (78.90)
Japanese	Altaic	0.207, 0.617, 0.176	335.33 (99.16)
Korean	Altaic	0.154, 0.617, 0.229	356.48 (97.00)
Spanish	Romance	0.073, 0.502, 0.425	382.84 (77.35)
Telugu	Dravidian	0.086, 0.595, 0.319	418.69 (95.22)
Turkish	Altaic	0.073, 0.561, 0.366	373.41 (88.03)
Chinese (ZHO)	Sino-Tibetan	0.090, 0.662, 0.248	384.87 (84.44)

^aof low, medium, and high proficiency speakers.

^bmean (SD) of native speakers: 250.72 (30.92).

Table 3: Corpus description.

A.2 Post-hoc Analysis of Essay-level metrics

Besides the F-scores from ANOVA analysis, we also conducted the post hoc analysis to investigate the variations of information distribution among L2 English learners with different L1 backgrounds and proficiency. The following tables showed the post hoc analysis results for the surprisal metric (Table 4), the entropy metric (Table 5), and the UID metric (Table 6). Similar to the F-scores result in Section 4.2, we found more significantly different L1 pairs among medium proficiency speakers, indicating these speakers have more variation in terms of information distribution patterns than less and more proficient speakers. Measured as the number of significantly different L1 pairs, the UID metric shows less variation than the surprisal and entropy metrics, suggesting that distributing information evenly when producing written language is a more universal mechanism for human language users.

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	1.688	-								
FRA	0.860	-0.828	-							
HIN	0.572	-1.116	-0.288	-						
ITA	0.922	-0.766	0.062	0.350	-					
JPN	0.685	-1.003	-0.175	0.113	-0.237	-				
KOR	0.572	-1.116	-0.288	<0.001	-0.350	-0.113	-			
SPA	0.247	-1.441	-0.613	-0.325	-0.675	-0.438	-0.325	-		
TEL	-0.176	-1.864	-1.036	-0.748	-1.098	-0.861	-0.748	-0.423	-	
TUR	0.465	-1.223	-0.395	-0.107	-0.457	-0.220	-0.108	0.217	0.641	-
ZHO	0.690	-0.998	-0.170	0.118	-0.232	0.005	0.118	0.443	0.866	0.226

(a) Low proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	1.452	-								
FRA	0.902	-0.550	-							
HIN	0.021	-1.431	-0.881	-						
ITA	0.526	-0.926	-0.376	0.505	-					
JPN	0.359	-1.092	-0.543	0.339	-0.166	-				
KOR	0.604	-0.848	-0.298	0.583	0.078	0.244	-			
SPA	0.662	-0.790	-0.240	0.641	0.136	0.302	0.058	-		
TEL	-0.545	-1.997	-1.447	-0.566	-1.071	-0.904	-1.148	-1.207	-	
TUR	0.441	-1.010	-0.460	0.421	-0.084	0.082	-0.162	-0.220	0.986	-
ZHO	0.381	-1.071	-0.521	0.360	-0.145	0.022	-0.222	-0.281	0.926	-0.060

(b) medium proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	0.637	-								
FRA	0.286	-0.350	-							
HIN	-0.012	-0.649	-0.299	-						
ITA	-0.035	-0.672	-0.322	-0.023	-					
JPN	0.084	-0.553	-0.202	0.097	0.119	-				
KOR	0.194	-0.443	-0.092	0.207	0.229	0.110	-			
SPA	0.314	-0.323	0.027	0.326	0.349	0.229	0.120	-		
TEL	-0.582	-1.219	-0.869	-0.570	-0.547	-0.667	-0.776	-0.896	-	
TUR	0.155	-0.482	-0.131	0.167	0.190	0.071	-0.039	-0.159	0.737	-
ZHO	-0.028	-0.665	-0.315	-0.016	0.007	-0.112	-0.222	-0.342	0.554	-0.183

^a A negative number indicates a smaller mean value for the row L1.

^b A bold value indicates a significant difference between row and column L1 (p -value < 0.05).

(c) High proficiency

Table 4: Post-hoc group difference of surprisal metric regarding L1, with proficiency control.

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	-0.777	-								
FRA	-0.519	0.258	-							
HIN	-0.535	0.242	-0.016	-						
ITA	-0.456	0.321	0.063	0.080	-					
JPN	-0.718	0.059	-0.199	-0.183	-0.262	-				
KOR	-0.523	0.254	-0.004	0.013	-0.067	0.196	-			
SPA	-0.134	0.643	0.385	0.401	0.322	0.584	0.388	-		
TEL	-0.434	0.343	0.085	0.101	0.022	0.284	0.089	-0.300	-	
TUR	-0.543	0.234	-0.023	-0.007	-0.087	0.176	-0.020	-0.408	-0.109	-
ZHO	-0.559	0.218	-0.040	-0.023	-0.103	0.159	-0.036	-0.425	-0.125	-0.016

(a) Low proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	-0.450	-								
FRA	-0.342	0.108	-							
HIN	-0.035	0.415	0.307	-						
ITA	-0.172	0.278	0.170	-0.137	-					
JPN	-0.453	-0.003	-0.111	-0.418	-0.281	-				
KOR	-0.363	0.087	-0.021	-0.328	-0.191	0.090	-			
SPA	-0.225	0.225	0.117	-0.190	-0.053	0.228	0.138	-		
TEL	-0.184	0.266	0.158	-0.149	-0.012	0.269	0.179	0.041	-	
TUR	-0.282	0.168	0.060	-0.247	-0.109	0.171	0.081	-0.057	-0.098	-
ZHO	-0.334	0.116	0.008	-0.299	-0.161	0.119	0.029	-0.109	-0.150	-0.052

(b) Medium proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	-0.141	-								
FRA	-0.119	0.022	-							
HIN	0.075	0.216	0.194	-						
ITA	-0.019	0.122	0.100	-0.094	-					
JPN	-0.347	-0.206	-0.228	-0.422	-0.328	-				
KOR	-0.256	-0.115	-0.137	-0.331	-0.237	0.091	-			
SPA	-0.143	-0.002	-0.024	-0.218	-0.123	0.204	0.113	-		
TEL	0.014	0.155	0.133	-0.061	0.033	0.361	0.270	0.156	-	
TUR	-0.135	0.007	-0.015	-0.209	-0.115	0.213	0.122	0.008	-0.148	-
ZHO	-0.184	-0.043	-0.065	-0.259	-0.165	0.163	0.072	-0.041	-0.198	-0.050

^a A negative number indicates a smaller mean value for the row L1.

^b A bold value indicates a significant difference between row and column L1 (p -value < 0.05).

(c) High proficiency

Table 5: Post-hoc group difference of entropy metric regarding L1, with proficiency control.

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	0.617	-								
FRA	0.417	-0.200	-							
HIN	-0.032	-0.649	-0.449	-						
ITA	-0.054	-0.671	-0.471	-0.023	-					
JPN	2.652	2.035	2.235	2.684	2.707	-				
KOR	2.436	1.819	2.019	2.468	2.491	-0.216	-			
SPA	-0.749	-1.366	-1.166	-0.717	-0.694	-3.401	-3.185	-		
TEL	-1.303	-1.920	-1.720	-1.271	-1.248	-3.955	-3.739	-0.554	-	
TUR	1.469	0.852	1.052	1.501	1.523	-1.183	-0.967	2.218	2.772	-
ZHO	2.270	1.653	1.853	2.302	2.324	-0.382	-0.166	3.019	3.573	0.801

(a) Low proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	0.032	-								
FRA	0.341	0.308	-							
HIN	-0.232	-0.265	-0.573	-						
ITA	0.070	0.037	-0.271	0.302	-					
JPN	1.433	1.400	1.093	1.665	1.363	-				
KOR	1.141	1.109	0.801	1.373	1.071	-0.292	-			
SPA	0.202	0.170	-0.138	0.435	0.132	-1.231	-0.939	-		
TEL	-0.576	-0.609	-0.917	-0.344	-0.646	-2.009	-1.717	-0.779	-	
TUR	-0.046	-0.079	-0.387	0.186	0.116	-1.480	-1.188	-0.249	0.530	-
ZHO	0.886	0.854	0.546	1.118	0.816	-0.547	-0.255	0.684	1.462	0.933

(b) Medium proficiency

Language	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	-									
DEU	-0.778	-								
FRA	-0.450	0.328	-							
HIN	-0.246	0.532	0.204	-						
ITA	-0.905	-0.127	-0.456	-0.659	-					
JPN	0.182	0.961	0.632	0.429	1.088	-				
KOR	0.036	0.815	0.486	0.283	0.942	-0.146	-			
SPA	-0.038	0.740	0.411	0.208	0.867	-0.221	-0.075	-		
TEL	-0.541	0.237	-0.091	-0.295	0.364	-0.723	-0.578	-0.503	-	
TUR	-0.672	0.106	-0.222	-0.426	0.233	-0.854	-0.709	-0.634	-0.131	-
ZHO	-0.051	0.727	0.399	0.195	0.854	-0.233	-0.088	-0.013	0.490	0.621

^a A negative number indicates a smaller mean value for the row L1.

^b A bold value indicates a significant difference between row and column L1 (p -value < 0.05).

(c) High proficiency

Table 6: Post-hoc group difference of UID metric regarding L1, with proficiency control.