# Lexical Sophistication in Advanced L1 and L2 Writing in a Standardized Language Proficiency Test

You-Min Lin & Michelle Y. Chen
Paragon Testing Enterprises

## Summary

- Drawing on recent advances in natural language processing (NLP) tools, this study examines whether the lexical features differ in advanced English writing performances by English and Spanish L1 test takers of the Canadian English Language Proficiency Index Program (CELPIP)-General.
- Statistical analysis shows that the high proficiency samples by English and Spanish L1 writers demonstrate largely similar lexical profiles. However, linear discriminant analysis (LDA) shows that psycholinguistic word properties (lexical decision response time, word naming accuracy) and range indices are among the most discriminating features for advanced English and Spanish L1 writing samples.

## Background

- Lexical sophistication is an important construct of writing proficiency. Research shows that lexical profiles of L2 writing often differ from those of L1 writing, which may indicate different proficiency and trajectory of lexical development. However, less attention has been paid to whether L1 and L2 writing of comparable quality demonstrates systematic lexical profile differences in the language assessment context.
- This study examines whether the lexical features of advanced English writing performances by English and Spanish L1 test takers of CELPIP-General are different between the two groups.

## Methodology

### *Writing Samples*

- To investigate lexical sophistication in advanced L1 and L2 writing, we analyzed the lexical sophistication in advanced writing performances by self-reported English and Spanish L1 test takers of the CELPIP-General Test, a standardized English proficiency test calibrated to the Canadian Language Benchmark (CLB).
- The English writing samples consisted of two writing task types (one email, one essay explaining a choice) of 407 English L1 and 407 Spanish L1 test takers receiving advanced CELPIP writing scores (9-12, equivalent to CLB 9-12, stage III advanced proficiency; mean writing scores English L1 = 10.34, Spanish L1 = 10.31). The running total of the corpus is 315,779 words (English L1 = 156,571 words, Spanish L1 = 159,208 words).
- The English L1 test taker sample (male = 171, female = 236) represented 26 nationalities, with the most common being United States (n = 100), United Kingdom (n = 88), and Ireland (n = 64). The Spanish L1s (male = 195, female = 212) reported 33 nationalities, with the top three being Mexico (n = 111), Venezuela (n = 50), and Colombia (n = 46).

### *Analysis*

- Initially, 261 lexical features were measured using TAALES (Kyle & Crossley, 2015), including text length, use of academic vocabulary (normed count of words occurring in Academic Word Lists and Academic Formula Lists), age of acquisition (Incremental age of exposure and mean age of acquisition scores), contextual distinctiveness, association strength, frequency (word and n-gram frequencies in COCA and BNC), range (word and n-gram in COCA and BNC), psycholinguistic norms, word neighbor Information (orthographic, phonographic, and phonological neighbors), and word recognition norms (lexical decision, naming accuracy and reaction time).
- Subsequently, features showing no variance in our writing samples were removed. To control for the effect of writing task type, we then removed features differing significantly ($p<.001$) between the two writing task types.
- After the initial feature removal, the remaining 102 indices were analyzed using linear discriminant analysis (LDA). Using LDA allowed us to find combinations of the lexical features that best separate the two groups, English and Spanish L1s.
- The contribution of each lexical feature to the discriminant function was examined and features with high discriminatory power were identified.

## Results

- Results show that the high proficiency writing samples by English and Spanish L1 writers demonstrate largely similar lexical profiles. The lexical indices cannot effectively separate the advanced writing samples of English and Spanish L1s, as shown in Figure 1. The significant overlap between English and Spanish L1 distributions even after LDA transformation indicates that the two groups have overall similar lexical profiles as measured by the selected features.
- The importance of the features to the discriminant function can be ranked based on the absolute value of the coefficients. Figure 2 shows the importance of features, ranked from high to low.
- The top 10 features after LDA are listed in Table 1, with a comparison of the mean values between English and Spanish L1s. Word recognition norms and word range features are among the 10 most important lexical features that can distinguish between the performance of English and Spanish L1 test takers.
- Among the top 10 ranked LDA features, English L1 test takers demonstrate a slightly lower range value (COCA_spoken_Range_Log_AW, COCA_spoken_Trigram_Range_Log), indicating that they may be using vocabulary with a slightly narrower range, i.e., the English L1 writers used more specialized vocabulary with a restricted distribution.

## Discussion

- Research on L1 and L2 performance is crucial to the understanding of L2 proficiency and beneficial for tracking learner development. Overall, the findings indicate that advanced writing by English and Spanish L1s demonstrates similar lexical sophistication as measured by the selected indices.
- The results call for further investigation of the effect of task type on L1 and L2 writing, as tasks of varying complexity and demand may require different lexical repertoire and elicit distinctive lexical features. As well, other factors associated with English learning and using experience will likely have an effect on lexical sophistication.
- While NLP tools do not offer a full representation of language constructs, they enable the systematic analysis of linguistic features across users. Building on this study, future studies can compare L1 and L2 writing using a wider range of linguistic features. Qualitative analyses may also reveal discriminating features of advanced L1 and L2 writing judged to be of similar quality.
- The current observations are restricted to advanced English writing samples by English and Spanish L1s. Extending the research to other proficiency levels (e.g., upper-intermediate proficiency) and other L1 backgrounds may reveal a wider range of lexical profiles, strategies, and combination of features employed to achieve a similar communicative efficacy.

Figure 1. Distributions after LDA Transformation

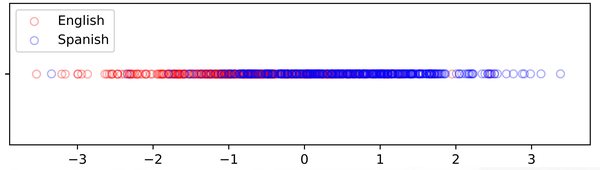
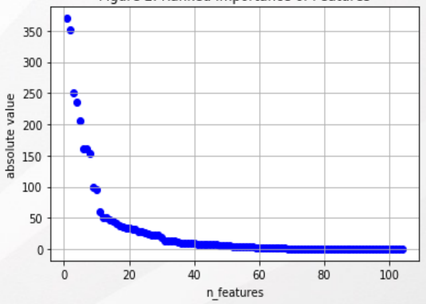



Table 1. Top 10. LDA Features

| Index | English | Spanish | Sig. | Diff. | Rank | Category |
|---|---|---|---|---|---|---|
| LD_Mean_RT_Zscore | -0.57 | -0.57 | | 0.00 | 1 | Word Recognition Norms |
| WN_Mean_Accuracy_FW | 0.99 | 0.99 | | 0.00 | 2 | Word Recognition Norms |
| LD_Mean_RT_Zscore_CW | -0.54 | -0.55 | | 0.01 | 3 | Word Recognition Norms |
| COCA_Academic_Trigram_Range | 0.02 | 0.02 | | 0.00 | 4 | Range |
| WN_Mean_Accuracy_CW | 0.99 | 0.99 | | 0.00 | 5 | Word Recognition Norms |
| LD_Mean_RT_Zscore_FW | -0.60 | -0.60 | | 0.00 | 6 | Word Recognition Norms |
| LD_Mean_Accuracy | 0.96 | 0.96 | * | 0.00 | 7 | Word Recognition Norms |
| WN_Mean_Accuracy | 0.99 | 0.99 | * | 0.00 | 8 | Word Recognition Norms |
| COCA_spoken_Range_Log_AW | -0.57 | -0.55 | * | -0.02 | 9 | Range |
| COCA_spoken_Trigram_Range_Log | -2.04 | -2.01 | * | -0.03 | 10 | Range |

Figure 2. Ranked Importance of Features




## References

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), pp. 757-786. https://doi.org/10.1002/tesq.194

**Contact Information** You-Min Lin: ylin@paragontesting.ca Michelle Chen: mchen@paragontesting.ca