**Supplementary Materials Online**

**Predicting Processing Effort During L1 and L2 Reading:**

**The Relationship Between Text Linguistic Features and Eye Movements**

Shingo Nahatame

**Contents of Supplementary Materials:**

|  |  |
| --- | --- |
| Appendix S1: Overview of the tools, linguistic features, and measures used in this study | …1 |
| Appendix S2: Detailed description of the indices included in the L1 and L2 reading models with complex features | …5 |
| Appendix S3: Detailed results of relative importance analysis | …7 |
| Appendix S4: Further discussion of the results | …14 |
| Appendix S5: Example texts with higher and lower bigram/trigram frequency scores | …17 |

**Appendix S1: Overview of the tools, linguistic features, and measures used in this study**

Table S1. *Overview of the tools used, linguistic features analyzed, and measures and example indices assessed in this study*

|  |  |  |
| --- | --- | --- |
| tools1 | features | Measures and examples of the indices |
| TAALES  2.2 | Lexical sophistication2 | *Frequency*:  The mean frequency score based on the SUBTLEXus, BNC (written, spoken), and COCA (spoken, fiction)  *Range*:  The mean range score based on the SUBTLEXus, BNC (written, spoken), and COCA (spoken, fiction)  *Age of acquisition* (*AoA*):  The mean AoA score based on human judgments of the age at which a particular word is learned  *Familiarity*:  The mean familiarity score based on human judgments of how familiar a particular word is  *Concreteness*:  The mean concreteness score based on human judgments of how concrete a particular word is  *Meaningfulness*:  The mean meaningfulness score based on human judgments of how meaningful a particular word is  *Word response norms*:  The mean lexical decision or naming reaction time across all participants for a particular word  *Word neighbor information*:  The number of orthographic and photogeological neighbors of a particular word  *Word association*:  The number of response types/tokens for a given stimulus in a free-association task  *Contextual distinctiveness*:  The co-occurrence probability of a particular word based on corpora  *Polysemy*:  The average number of senses for content words  *Hypernymy*:  The average hypernymy score for nouns and verbs  *N-gram features*:  The mean frequency, range, and association score of bigrams and trigrams in the text based on the BNC (written, spoken) and COCA (spoken, fiction) |
| TAASSC 1.3.8 | Syntactic complexity | *Traditional indices of syntactic complexity*:  The mean length of T-units, mean length of clauses, average number of clauses per sentence, average number of dependent clauses per T-unit, average number of complex nominals per clause, and average number of coordinate phrases per T-unit  *Fine-grained indices of clausal complexity:*  The average number of specific structures (e.g., nominal subject, direct object, passive agents) per clause, average number of dependents per clause, and standard deviation of the number of dependents per clause  *Fine-grained indices of phrasal complexity*:  The average number of dependents per specific phrase type (e.g., nominal, direct objects, agent), incidence of specific dependent types (e.g., adjective modifiers, determiners, prepositions) per nominal, and incidence of specific dependent types in specific types of noun phrases (e.g., adjective modifiers occurring in direct object phrases) |
| TAACO 2.0.4 | Cohesion | *Lexical overlap*:  All lemma overlap, content word lemma overlap, and lemma overlap for POS tags between sentences or paragraphs  *Semantic* *overlap*:  The average latent semantic analysis (LSA) cosine, latent dirichlet allocation (LDA) divergence score, and Word2Vec similarity score between sentences and paragraphs  *Connectives*:  The frequency of positive, negative, causal, additive, and temporal connectives in the text  *Type-token ratio (TTR)*:  Simple TTR, content word TTR, lemma TTR, and content lemma TTR |
| SiNLP | Simple | *Word count*:  The total number of words in the text  *Average word length*:  The average number of letters per word  *Average sentence length*:  The average number of words per sentence |

**Note**

1. TAALES: the Tool for the Automatic Analysis of Lexical Sophistication (Kyle et al., 2018); TASSC: the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (Kyle & Crossley, 2018); TAACO: the Tool for the Automatic Analysis of Text Cohesion (Crossley et al., 2019); SiNLP: a simple NLP tool (Crossley et al., 2014).
2. L1 AoA indices were based on Kuperman, Stadthagen-Gonzales, and Brysbaert’s (2012) AoA list. Psycholinguistic word information indices were taken from the Medical Research Council psycholinguistic database (Coltheart, 1981) and Brysbaert, Warriner, and Kuperman’s (2014) concreteness norms. Word response norms were obtained from the English Lexicon Project (Balota et al., 2007). Word neighbor information included the number of orthographic and phonological neighbors that a word has (i.e., Coltheart’s N; Coltheart et al., 1977; Balota et al., 2007), and orthographic Levenshtein distance (OLD) 20 and phonological Levenshtein distance (PLD) 20 values (Yarkoni et al., 2008; Yap & Balota, 2009). Word association indices were taken from the Edinburgh Associative Thesaurus (EAT; Kiss et al., 1973) and the University of South Florida (USF) stimuli count index (Nelson et al., 1998). Contextual distinctiveness indices included semantic distinctiveness (Hoffman et al., 2013) and the McDonald co-occurrence probability (McDonald & Schillcock, 2001). Semantic lexical relations were the polysemy and hypernymy scores obtained from WordNet (Fellbaum, 1998). Strength-of-association norms for n-grams included mutual information (MI), MI2, and T scores (Evert, 2005) as well as DeltaP and approximate collexeme strength scores (Gries, Hampe & Schönefeld, 2005).

**References**

Balota, DA, Yap, MJ, Hutchison, KA, Cortese, MJ, Kessler, B, Loftis, B, Neely, NH, Nelson, DL, Simpson, GB and Treiman, R (2007) The English lexicon project. *Behavior Research Methods* 39(3), 445–459. <https://doi.org/10.3758/BF03193014>

Brysbaert, M, Warriner, AB, and Kuperman, V (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 904–911. <https://doi.org/10.3758/s13428-013-0403-5>

Coltheart, M (1981) The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A* 33, 497–505. <https://doi.org/10.1080/14640748108400805>

Coltheart, M, Davelaar, E, Jonasson, T and Besner, D (1977) Access to the internal lexicon. In Dornic, S (ed.), *Attention and performance IV*. Hillsdale, NJ: Erlbaum, pp. 535–555.

Crossley, SA, Allen, LK, Kyle, K and McNamara, DS (2014) Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes* 51(5-6), 511–534. <https://doi.org/10.1080/0163853X.2014.910723>

Crossley, SA., Kyle, K and Dascalu, M (2019) The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavioral Research Methods* 50, 1030–1046. <https://doi.org/10.3758/s13428-018-1142-4>

Evert, S (2005) *The statistics of word cooccurrences: Words pairs and collocations (Doctoral dissertation).* Germany: Universität Stuttgart.

Fellbaum, C (ed.) (1998) *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gries, ST, Hampe, B and Schönefeld, D (2005) Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16, 635–676. <https://doi.org/10.1515/cogl.2005.16.4.635>

Hoffman, P, Lambon Ralph, MA, and Rogers, TT (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods* 45, 718–730. <https://doi.org/10.3758/s13428-012-0278-x>

McDonald, SA and Shillcock, RC (2001) Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(3), 295–322. [https://doi.org/10.1177/00238309010440030101](https://doi.org/10.1177%2F00238309010440030101)

Nelson, DL, McEvoy, CL and Schreiber, TA (2004) The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36, 402–407. [doi.org/10.3758/BF03195588](https://doi.org/10.3758/BF03195588)

Kiss, GR, Armstrong, C, Milroy, R and Piper, J (1973) An associative thesaurus of English and its computer analysis. *The Computer and Literary Studies*, 153–165.

Kuperman, V, Stadthagen-Gonzales, H, and Brysbaert, M (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods* 44, 978–990. <https://doi.org/10.3758/s13428-012-0210-4>

Kyle, K, Crossley, S and Berger, C (2018) The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50, 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>

Kyle, K and Crossley, SA (2018) Measuring syntactic complexity in L2 writing using fine‐grained clausal and phrasal indices. *The Modern Language Journal* 102, 333–349. <https://doi.org/10.1111/modl.12468>

Yap, MJ, and Balota, DA (2009) Visual word recognition of multisyllabic words. *Journal of Memory and Language* 60(4), 502-529. <https://doi.org/10.1016/j.jml.2009.02.001>

Yarkoni, T, Balota, D and Yap, M (2008) Moving beyond Coltheart’s N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review* 15(5), 971–979. <https://doi.org/10.3758/PBR.15.5.971>

**Appendix S2: Detailed description of the indices included in the L1 and L2 reading models with complex features**

Table S2. *Detailed description of the indices included in the L1 reading model with complex linguistic features*

|  |  |
| --- | --- |
| Index | Description |
| MRC\_Familiarity\_CW (familiarity of content words) | the mean familiarity scores for content words based on the MRC Database |
| SUBTLEXus\_Freq\_FW  (frequency of function words) | the mean frequency scores for function words based on the SUBTLEXus corpus |
| COCA\_fiction\_bi\_prop\_70k  (bigram frequency) | proportion of bigrams in text that are among the 70,000 most frequent bigrams in the COCA corpus (fiction sub-corpus) |
| COCA\_fiction\_tri\_prop\_40k  (trigram frequency) | proportion of trigrams in text that are among the 40,000 most frequent trigrams in the COCA corpus (fiction sub-corpus) |
| eat\_types\_FW  (association for function words) | the mean association scores for function words (number of word types that come to mind in response to function word as stimuli in free association task) |
| eat\_tokens\_FW  (association for function words) | the mean association scores for function words (number of word tokens that come to mind in response to function word as stimuli in free association task) |
| PLD\_CW  (phonological neighborhood of content words) | mean phonological Levenshtein distances from a content word to its 20 closest neighbors |
| dobj\_NN\_stdev  (noun phrase variety) | the standard deviation of the number of dependents per direct object (no pronouns) |
| prep\_per\_cl (clause complexity) | the mean number of prepositions per clause |

*Note*. Based on the description in the index guide for each tool (available from the NLP for Social Science at <https://www.linguisticanalysistools.org/>)

Table S3. *Detailed description of the indices included in the L2 reading model with complex linguistic features*

|  |  |
| --- | --- |
| Index | Description |
| MRC\_Familiarity\_CW  (familiarity of content words) | the mean familiarity scores for content words based on the MRC Database |
| MRC\_Concreteness\_CW  (concreteness of content words) | the mean concreteness scores for content words based on the MRC Database |
| SUBTLEXus\_Freq\_CW  (frequency of content words) | the mean frequency scores for content words based on the SUBTLEXus corpus |
| BNC\_Spoken\_Freq\_FW\_Log  (frequency of function words) | the mean frequency scores (log transformed) for function words based on the BNC spoken corpus |
| COCA\_fiction\_bi\_DP  (bigram association strength) | the mean Delta P association score (left to right) for bigrams based on the COCA corpus (fiction sub-corpus) |
| COCA\_fiction\_bi\_prop\_70k  (bigram frequency) | proportion of bigrams in text that are among the 70,000 most frequent bigrams in the COCA corpus (fiction sub-corpus) |
| COCA\_fiction\_tri\_prop\_40k  (trigram frequency) | proportion of trigrams in text that are among the 40,000 most frequent trigrams in the COCA corpus (fiction sub-corpus) |
| eat\_types\_FW  (association of function words) | the mean association scores for function words (number of word types that come to mind in response to function word as stimuli in free association task) |
| eat\_tokens\_FW  (association of function words) | the mean association scores for function words (number of word tokens that come to mind in response to function word as stimuli in free association task) |
| McD\_CD\_FW  (contextual distinctiveness of function words) | the co-occurrence probability of function word with 500 highly frequent context lemmas (within 5 unigrams to the left and right of the target lemma) |
| DC\_C  (proportion of dependent clauses) | the number of dependent clauses in text divided by the number of clauses in text |

*Note*. Based on the description in the index guide for each tool (available from the NLP for Social Science at <https://www.linguisticanalysistools.org/>)

**Appendix S3: Detailed results of relative important analysis**

*S3-1 Relative importance analysis*

Although the standardized regression coefficients have been often used to determine the importance of predictors in the regression model, they can be flawed measures of relative importance due to the intercorrelations among the predictors. Relative weight analysis is widely accepted as one of the approaches to address this issue and assess relative importance of predictors in the regression model more accurately. This analysis uses a variable transformation approach to derive a new set of predictors that are orthogonal to each other, regressing the criterion on these new orthogonal predictors and then converting the resulting standardized regression coefficients to the metric of the original predictors (see Tonidandel and LeBreton, 2011 for more details). The current study mainly reported the results of relative weight analysis in order to discuss the particular role played by predictors in a regression model..

Nevertheless, there are other approaches to measure the relative importance of predictors in the regression model. Among them, Mizumoto (2022) recommended the use of dominance analysis instead of relative weight analysis accompanied by random forest, a machine learning method (see Mizumoto, 2022 for more details). Thus, this appendix reports the detailed results of relative weight analysis for L1 and L2 reading models along with the results of dominance analysis and random forest. All of the results were obtained by employing Mizumoto’s (2022)R-based web application from the langtest.jp, available at <http://langtest.jp/shiny/relimp/>.

References

Mizumoto, A (2022). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning* (Early View). <https://doi.org/10.1111/lang.12518>

Tonidandel, S and LeBreton, JM (2011) Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology* 26(1), 1–9. <https://doi.org/10.1007/s10869-010-9204-3>

*S3-2 Results of L1 reading*

Table S4. *Results of relative weight analysis for the L1 reading model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Raw rel.weight | Rescaled  rel.weight | *95%CI*  *(raw weights)* | *95%CI (test of significance)* |
| Word.Count | .724 | 87.70 | [.685, .775] | [.687, .778] |
| MRC\_Familiarity\_CW (familiarity of content words) | .003 | 0.36 | [.001, .008] | [−.004, .013] |
| SUBTLEXus\_Freq\_FW  (frequency of function words) | .012 | 1.39 | [.003, .028] | [.001, .030] |
| COCA\_fiction\_bi\_prop\_70k  (bigram frequency) | .023 | 2.84 | [.015, .035] | [.014, .038] |
| COCA\_fiction\_tri\_prop\_40k  (trigram frequency) | .016 | 1.93 | [.009, .027] | [.007, .031] |
| eat\_types\_FW  (association for function words) | .008 | 0.93 | [.002, .022] | [−.001, .024] |
| eat\_tokens\_FW  (association for function words) | .005 | 0.64 | [.001, .020] | [−.003, .024] |
| PLD\_CW  (phonological neighborhood of content words) | .008 | 0.96 | [.003, .018] | [.000, .021] |
| dobj\_NN\_stdev  (*SD* of the number of dependents per direct object) | .012 | 1.49 | [.002, .032] | [.000, .035] |
| prep\_per\_cl  (number of prepositions per clause) | .015 | 1.76 | [.003, .036] | [.002, .037] |

グラフ

自動的に生成された説明

Figure S1. Plot of raw relative weights for the L1 reading model.

Table S5. *Results of dominance analysis for the L1 reading model*

|  |  |
| --- | --- |
|  | Weight |
| Word.Count | .728 |
| MRC\_Familiarity\_CW (familiarity of content words) | .003 |
| SUBTLEXus\_Freq\_FW (frequency of function words) | .013 |
| COCA\_fiction\_bi\_prop\_70k (bigram frequency) | .021 |
| COCA\_fiction\_tri\_prop\_40k (trigram frequency) | .014 |
| eat\_types\_FW (association for function words) | .008 |
| eat\_tokens\_FW (association for function words) | .005 |
| PLD\_CW (phonological neighborhood of content words) | .007 |
| dobj\_NN\_stdev (*SD* of the number of dependents per direct object) | .012 |
| prep\_per\_cl (number of prepositions per clause) | .015 |

グラフ, ヒストグラム

自動的に生成された説明

Figure S2. Plot of variable importance assessed by random forest for the L1 reading model. See Table S6 below for the numbering of the variables.

Table S6. *Results of random forest for the L1 reading model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | meanZ | medianZ | min  Z | maxZ | normHits | decision |
| 1. Word.Count | 81.38 | 81.66 | 71.00 | 93.15 | 1.00 | Confirmed |
| 2. MRC\_Familiarity\_CW (familiarity of content words) | 1.05 | 0.87 | -0.99 | 3.62 | 0.06 | Rejected |
| 3. SUBTLEXus\_Freq\_FW (frequency of function words) | 6.41 | 6.43 | 3.50 | 8.91 | 1.00 | Confirmed |
| 4. COCA\_fiction\_bi\_prop\_70k (bigram frequency) | 13.99 | 14.05 | 10.75 | 16.89 | 1.00 | Confirmed |
| 5. COCA\_fiction\_tri\_prop\_40k (trigram frequency) | 9.19 | 9.25 | 6.46 | 12.04 | 1.00 | Confirmed |
| 6. eat\_types\_FW (association for function words) | 4.77 | 4.74 | 2.02 | 8.18 | 0.98 | Confirmed |
| 7. eat\_tokens\_FW (association for function words) | 2.56 | 2.62 | -0.08 | 4.91 | 0.71 | Confirmed |
| 8. PLD\_CW (phonological neighborhood of content words) | 4.34 | 4.40 | 1.85 | 7.41 | 0.97 | Confirmed |
| 9. dobj\_NN\_stdev (*SD* of the number of dependents per direct object) | 1.31 | 1.33 | -1.35 | 3.64 | 0.38 | Rejected |
| 10. prep\_per\_cl (number of prepositions per clause) | 2.82 | 2.89 | 0.33 | 5.71 | 0.77 | Confirmed |

*S3-3 L2 reading*

Table S7. *Results of relative weight analysis for the L2 reading model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Raw rel.weight | Rescaled  rel.weight | *95%CI*  *(raw weights)* | *95%CI (test of significance)* |
| Word.Count | .668 | 85.78 | [.620, .726] | [.614, .726] |
| MRC\_Familiarity\_CW  (familiarity of content words) | .006 | 0.74 | [.002, .015] | [−.015, .013] |
| MRC\_Concreteness\_CW  (concreteness of content words) | .004 | 0.48 | [.001, .012] | [−.018. .008] |
| SUBTLEXus\_Freq\_CW  (frequency of content words) | .009 | 1.12 | [.004, .019] | [−.012, .015] |
| BNC\_Spoken\_Freq\_FW\_Log  (frequency of function words) | .008 | 1.08 | [.001, .024] | [−.012, .021] |
| COCA\_fiction\_bi\_DP  (bigram association strength) | .010 | 1.26 | [.002, .026] | [−.011, .023] |
| COCA\_fiction\_bi\_prop\_70k  (bigram frequency) | .016 | 2.11 | [.009, .027] | [−.006, .023] |
| COCA\_fiction\_tri\_prop\_40k  (trigram frequency) | .021 | 2.75 | [.011, .038] | [.000, .035] |
| eat\_types\_FW  (association of function words) | .011 | 1.40 | [.002, .028] | [−.011, .025] |
| eat\_tokens\_FW  (association of function words) | .006 | 0.81 | [.001, .022] | [−.014, .020] |
| McD\_CD\_FW  (contextual distinctiveness of function words) | .009 | 1.22 | [.001, .026] | [−.011, .023] |
| DC\_C  (proportion of dependent clauses) | .010 | 1.24 | [.003, .036] | [−.012, .025] |

グラフ

中程度の精度で自動的に生成された説明

Figure S2. Plot of relative weights for the L2 reading model.

Table S8. *Results of dominance analysis for the L2 reading model*

|  |  |
| --- | --- |
|  | Weight |
| Word.Count | .671 |
| MRC\_Familiarity\_CW (familiarity of content words) | .005 |
| MRC\_Concreteness\_CW (concreteness of content words) | .004 |
| SUBTLEXus\_Freq\_CW (frequency of content words) | .009 |
| BNC\_Spoken\_Freq\_FW\_Log (frequency of function words) | .009 |
| COCA\_fiction\_bi\_DP (bigram association strength) | .010 |
| COCA\_fiction\_bi\_prop\_70k (bigram frequency) | .014 |
| COCA\_fiction\_tri\_prop\_40k (trigram frequency) | .020 |
| eat\_types\_FW (association of function words) | .012 |
| eat\_tokens\_FW (association of function words) | .007 |
| McD\_CD\_FW (contextual distinctiveness of function words) | .010 |
| DC\_C (proportion of dependent clauses) | .010 |

グラフ

自動的に生成された説明

Figure S4. Plot of variable importance assessed by random forest for the L2 reading model. See Table S9 below for the numbering of the variables.

Table S9. *Results of random forest for the L2 reading model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | meanZ | medianZ | min  Z | maxZ | normHits | decision |
| 1. Word.Count | 76.10 | 75.81 | 69.17 | 83.84 | 1.00 | Confirmed |
| 2. MRC\_Familiarity\_CW (familiarity of content words) | 1.75 | 1.78 | −1.28 | 4.53 | 0.41 | Tentative |
| 3. MRC\_Concreteness\_CW (concreteness of content words) | 2.94 | 2.90 | 0.55 | 5.70 | 0.75 | Confirmed |
| 4. SUBTLEXus\_Freq\_CW (frequency of content words) | 5.98 | 6.03 | 2.62 | 8.63 | 0.99 | Confirmed |
| 5. BNC\_Spoken\_Freq\_FW\_Log (frequency of function words) | 4.39 | 4.36 | 2.04 | 7.07 | 0.94 | Confirmed |
| 6. COCA\_fiction\_bi\_DP (bigram association strength) | 1.98 | 1.96 | −0.62 | 4.82 | 0.45 | Tentative |
| 7. COCA\_fiction\_bi\_prop\_70k (bigram frequency) | 9.29 | 9.30 | 6.84 | 11.70 | 1.00 | Confirmed |
| 8. COCA\_fiction\_tri\_prop\_40k (trigram frequency) | 10.36 | 10.37 | 6.99 | 12.95 | 1.00 | Confirmed |
| 9. eat\_types\_FW (association of function words) | 3.15 | 3.10 | 0.21 | 5.73 | 0.79 | Confirmed |
| 10. eat\_tokens\_FW (association of function words) | 3.02 | 2.98 | 0.83 | 6.24 | 0.75 | Confirmed |
| 11. McD\_CD\_FW (contextual distinctiveness of function words) | 3.08 | 2.95 | −0.11 | 5.39 | 0.80 | Confirmed |
| 12. DC\_C (proportion of dependent clauses) | 1.25 | 1.24 | −0.48 | 3.60 | 0.06 | Rejected |

**Appendix S4: Further discussion of the results**

*S4-1 The large variance explained by the models*

The variance explained by the current models may seem quite large, given that eye movement data tends to be noisy. This could be partly explained by the fact that the current study analyzed global reading measures (i.e., total reading time and fixation count) at the trial (text) level and included word count as a control variable. That is, the current eye movement measures provided a rough estimate of processing effort for an entire text rather than for detailed text regions (e.g., an individual word or a phrase). This likely alleviated the effects of some noisy factors, such as landing errors after a saccade to a new line, erroneous fixations, and blinks. Additionally, the inclusion of word count index explained the great majority of variance because it strongly correlated with total time and fixation count. The explained variance was further increased by the addition of other linguistic indices.

*S4-2 In comparison with E-Z reader and Graesser and McNamara’s (2011) model*

The current models indicated the importance of lexical features in modeling eye movements during reading. This concurs with E-Z reader (Reichle et al., 1999), which assumes a large effect of word length and frequency on eye movements. Nevertheless, the current models differ from E-Z reader in that they are data-driven models that were constructed using eye movement measures collected from L1 and L2 readers, respectively, and that they consider the effects of various linguistic features rather than word length and frequency. On the other hand, the current models do not provide as detailed a description of eye movement control as does E-Z reader.

The current models have implications for Graesser and McNamara’s (2011) model. Although this study analyzed text linguistic features at four of the six levels proposed by this model (i.e., the word, syntactic, textbase, and situation model), most of the predictive variables in the current models were indices at the word level. This suggests the importance of the lowest level (i.e., word) of the model compared to other levels when considering processing effort involved during reading. Thus, it may be beneficial to extend this model for processing by weighting the word level. In addition, given that the importance of the word level was similar for both L1 and L2 reading as shown in this study, such models may be applied to not only L1 but also L2 reading.

*S4-3 Bigram association strength in the L2 reading model*

The index of bigram association strength (based on the fiction sub-corpus) was only included and significant in the L2 reading model. However, the relationship between bigram strength and processing effort may be counterintuitive: processing effort increased for the text that contained more strongly associated bigrams (the bivariate correlation was also positive). Several studies found that the use of more strongly associated n-grams (as evaluated by Delta P scores as in the current study) in L2 writing led to high-quality essays as assessed by experts (e.g., Garner, Crossley & Kyle, 2019; Kyle et al., 2018; Zhang & Li, 2021). Thus, strongly associated n-grams can be indicative of sophisticated language. Given this, it is possible to assume that the smooth processing of strongly associated n-grams requires more sophisticated collocational knowledge on the part of readers; therefore, L2 readers, whose linguistic knowledge is limited relative to L1 readers, might have taken more time and effort to activate the association of such bigrams. Nevertheless, further investigation is required to confirm this notion by directly examining the effects of association strength on the processing of bigrams in L2 reading. Additionally, it should be noted that the results of random forest shown above did not confirm the index of bigram association strength to be an important variable in the L2 reading model.

*S4-4 Limitations and future directions of the use of eye movement measures*

The current study used the combination of total reading time and fixation count as the criterion variable. Total reading time is “a hybrid measure that conflates both early and late stages of processing” (Godfroid, 2019, p. 255). Similarly, fixation count is an aggregate measure and highly correlates with total reading time (Conklin et al., 2019). Given this, the current models should be interpreted as estimating cognitive effort at both early (or lower-level) and late (or higher-level) processing.

Thus, in future studies, the use of early and late processing measures, separately, will allow us to obtain more nuanced insight into the processing effort and text linguistic features. For instance, the model could estimate the cognitive effort specifically required for lower-level processing (i.e., word recognition) if a measure of early processing, such as first fixation time, is used. On the other hand, it could estimate the effort for higher-level or strategic processing if a measure of late processing, such as rereading time and lookbacks to previous sentences, is used.

References

Conklin, K, Pellicer-Sánchez, A and Carrol, G (2018) *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Garner, J, Crossley, S and Kyle, K (2019) N-gram measures and L2 writing proficiency. *System* 80, 176–187. <https://doi.org/10.1016/j.system.2018.12.001>

Godfroid, A (2019) *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York, NY: Routledge.

Graesser, AC and McNamara, DS (2011) Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 3(2), 371–398. <https://doi.org/10.1111/j.1756-8765.2010.01081.x>

Kyle, K, Crossley, S and Berger, C (2018) The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50, 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>

Reichle, ED, Rayner, K and Pollatsek, A (1999) Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ Reader model. *Vision Research* 39(26), 4403–4411. <https://doi.org/10.1016/S0042-6989(99)00152-2>

Zhang, X and Li, W (2021) Effects of n-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. *System* 97, 102437. <https://doi.org/10.1016/j.system.2020.102437>

**Appendix S5: Example texts with higher and lower bigram/trigram frequency scores**

Example text with higher bigram frequency score (0.83)

*I was quite honest with him. I told him, what was true, that I liked him very much, that I hoped to come to like him more, but that I was not in any way what the world calls 'in love' with him. He declared that that satisfied him, and so...we were married. She waited a long time, a little frown had gathered on her forehead. She seemed to be looking back earnestly into those past days. I think...I am sure...he cared for me at first. But I suppose we were not well matched. Almost at once, we drifted apart. He...it is not a pleasing thing for my pride, but it is the truth...tired of me very soon.*

Example text with higher trigram frequency score (0.34)

*I must have made some murmur of dissent, for she went on quickly: Oh, yes, he did! Not that it matters now...now that we've come to the parting of the ways. What do you mean? She answered quietly: I mean that I am not going to remain at Styles. You and John are not going to live here? John may live here, but I shall not. You are going to leave him? Yes. But why? She paused a long time, and said at last: Perhaps...because I want to be...free!*

Example text with lower bigram frequency score (0.44)

*Who is that? I asked sharply, for instinctively I distrusted the man. That's Dr. Bauerstein, said John shortly. And who is Dr Bauerstein? He's staying in the village doing a rest cure, after a bad nervous breakdown. He's a London specialist; a very clever man...one of the greatest living experts on poisons, I believe. And he's a great friend of Mary's, put in Cynthia, the irrepressible. John Cavendish frowned and changed the subject.*

Example text with lower trigram frequency score (0.05)

*Oh, this fellow! He turned up from nowhere, on the pretext of being a second cousin or something of Evie's, though she didn't seem particularly keen to acknowledge the relationship. The fellow is an absolute outsider, anyone can see that. He's got a great black beard, and wears patent leather boots in all weathers! But the mater cottoned to him at once, took him on as secretary...you know how she's always running a hundred societies?*