**Appendix 1: Indices of linguistic complexity, data source and data type of the eight readability formulas**

|  |  |  |  |
| --- | --- | --- | --- |
| Formulas | Linguistic indices involved  | Data Source | Data type |
| FRE | Number of words per sentence; Number of syllables per word  | English-L1 users | Scores of reading comprehension  |
| FKG  | Number of words per sentence; Number of syllables per word | English-L1 users | Cloze test scores |
| SMOG | Number of sentences; Number of words with ≥3 syllables | English-L1 users | Scores of reading comprehension |
| NDC | Percentage of difficult words; Number of words per sentence | English-L1 users | Reading grades |
| ARI | Number of letters per word; Number of words per sentence | English-L1 users | Reading grades |
| CML2RI | Word frequency; Sentence similarity value; Content word overlap value | English-L2 learners | Cloze test scores |
| CAREC | Age of acquisition scores for content words; Bigram range; Trigram proportion; Trigram type-token ratio; lexical overlap between paragraph; characters per word; temporal connectives; word imageability scores; positive adjectives; noun overlap in sentences; types of content lemma; character entropy; word frequency | English-L1 users  | Comparative judgement on text comprehensibility  |
| CARES | Number of content words; word naming time; complex nominals per T-unit; word concreteness; semantic distinctiveness of words; characters per word; number of function words; dependents per direct object nominal | English-L1 users  | Comparative judgement on text reading speed |
| *Note*. FRE = Flesch Reading Ease; FKG = Flesch-Kincaid Grade; ARI =Automated Readability Index; SMOG = Simple Measure of Gobbledygook Formula; NDC=New Dale-Chall Formula; CML2RI =Coh-Metrix L2 Reading Index; CAREC = Crowdsourced Algorithm of Reading Comprehension; CARES =Crowdsourced Algorithm of Reading Speed |

**Appendix 2: Correlations between lexical, syntactic, and discoursal complexity indices and reading rate of 272 snippets in GECO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with Reading rate |
| ***Lexical sophistication indices*** |  |  |  |  |  |
| Word length | 4.835 | 6.807 | 5.534 | 0.305 | -.204\*\* |
| MRC\_Meaningfulness\_AW | 308.850 | 401.895 | 357.410 | 15.867 | .303\*\* |
| MRC\_Concreteness\_CW | 275.217 | 437.950 | 340.731 | 32.274 | -.233\*\* |
| MRC\_Imageability\_CW | 312.720 | 458.063 | 373.458 | 29.359 | -.252\*\* |
| MRC\_Concreteness\_FW | 230.029 | 344.500 | 280.452 | 22.288 | .444\*\* |
| MRC\_Imageability\_FW | 244.667 | 356.909 | 292.928 | 22.074 | .423\*\* |
| MRC\_Meaningfulness\_FW | 276.472 | 399.483 | 330.610 | 24.058 | .429\*\* |
| Kuperman AoA\_AW | 4.319 | 5.881 | 5.108 | 0.305 | -.285\*\* |
| Kuperman AoA\_CW | 4.633 | 7.246 | 5.817 | 0.510 | -.294\*\* |
| COCA fiction\_Range\_AW | 0.491 | 0.726 | 0.633 | 0.044 | .184\*\* |
| COCA fiction\_Frequency\_AW | 3825.548 | 12078.653 | 7540.064 | 1434.051 | -.310\*\* |
| COCA fiction\_Range\_CW | 0.229 | 0.589 | 0.426 | 0.072 | .316\*\* |
| COCA fiction\_Frequency\_CW | 278.240 | 2485.485 | 1350.075 | 415.266 | .168\*\* |
| COCA fiction\_Frequency\_FW | 8680.097 | 24370.123 | 15993.874 | 3004.883 | -.304\*\* |
| COCA fiction\_bi\_DP | 0.014 | 0.076 | 0.040 | 0.013 | -.181\*\* |
| ***Fine-grained syntactic complexity indices*** |  |  |  |  |  |
| Dependents per nominal | 0.174 | 1.619 | 0.724 | 0.257 | -.402\*\* |
| Dependents per direct object | 0.000 | 3.000 | 0.939 | 0.534 | -.128\* |
| Dependents per object of the preposition | 0.000 | 2.400 | 1.009 | 0.382 | -.201\*\* |
| Dependents per direct object (no pronouns) | 0.000 | 3.000 | 1.125 | 0.597 | -0.108\* |
| Dependents per object of the preposition (no pronouns, standard deviation) | 0.000 | 1.730 | 0.756 | 0.268 | -.206\*\* |
| Determiners per nominal | 0.000 | 0.556 | 0.225 | 0.105 | -.314\*\* |
| Adjectival modifiers per nominal | 0.000 | 0.455 | 0.099 | 0.073 | -.198\*\* |
| Prepositions per nominal | 0.000 | 0.389 | 0.094 | 0.071 | -.261\*\* |
| Determiners per direct object | 0.000 | 1.000 | 0.299 | 0.239 | -.133\* |
| Prepositions per direct object | 0.000 | 1.000 | 0.154 | 0.188 | -0.112\* |
| Discourse markers per clause | 0.000 | 0.167 | 0.014 | 0.032 | .224\*\* |
| Clausal negations per clause | 0.000 | 0.400 | 0.075 | 0.075 | .175\*\* |
| Nominal subjects per clause | 0.333 | 1.154 | 0.755 | 0.135 | .179\*\* |
| ***Global syntactic complexity indices*** |  |  |  |  |  |
| Mean dependency distance  | 1.936 | 4.912 | 2.948 | 0.545 | -.418\*\* |
| Mean length of sentence  | 3.316 | 37.000 | 12.911 | 6.194 | -.592\*\* |
| Mean length of clause  | 4.118 | 21.500 | 7.762 | 2.381 | -.487\*\* |
| Complex nominal per T-unit | 0.182 | 5.667 | 1.123 | 0.783 | -.536\*\* |
| Complex nominal per clause  | 0.118 | 2.600 | 0.723 | 0.416 | -.505\*\* |
| ***Discoursal complexity indices*** |  |  |  |  |  |
| All lemma overlap between adjacent sentences | 0.000 | 8.000 | 1.636 | 1.388 | -.500\*\* |
| all lemmas overlap between binary adjacent sentence | 0.000 | 1.000 | 0.639 | 0.264 | -.474\*\* |
| Adjacent sentence overlap function lemmas | 0.000 | 0.500 | 0.197 | 0.104 | -.417\*\* |
| Function lemmas overlap between adjacent sentences | 0.000 | 5.667 | 1.214 | 1.057 | -.499\*\* |
| Function lemmas overlap between binary adjacent sentences | 0.000 | 1.000 | 0.570 | 0.280 | -.487\*\* |
| Noun lemmas overlap between adjacent sentences | 0.000 | 2.500 | 0.144 | 0.292 | -.368\*\* |
| Function lemmas overlap between binary adjacent sentences | 0.000 | 1.000 | 0.112 | 0.193 | -.366\*\* |
| Binary adjacent sentence overlap pronoun lemmas | 0.000 | 1.000 | 0.379 | 0.278 | -.267\*\* |
| Noun and pronoun lemmas overlap between adjacent sentences  | 0.000 | 2.667 | 0.552 | 0.464 | -.402\*\* |
| Binary adjacent sentence overlap noun and pronoun lemmas | 0.000 | 1.000 | 0.414 | 0.276 | -.371\*\* |
| determiners | 0.000 | 0.204 | 0.089 | 0.033 | -.362\*\* |
| Repeated content lemmas | 0.000 | 0.297 | 0.115 | 0.057 | .145\* |
| Repeated content and pronoun lemmas | 0.000 | 0.403 | 0.186 | 0.067 | .128\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Cumulative variance explained by two extracted components: Reading rate |
| Components | X Explained Variance | Y Explained Variance | R2 |
| 1 | 33.617  | 37.134  | 0.371 |
| 2 | 42.701 | 41.717 | 0.417 |

**Appendix 3: Correlations between lexical, syntactic, and discoursal complexity indices and mean fixation duration of 272 snippets in GECO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlation with MFD |
| ***Lexical sophistication indices*** |
| COCA fiction Range CW |  |  |  |  |  |
| COCA fiction\_Frequency\_Log\_CW | 1.503 | 2.695 | 2.217 | 0.235 | -.305\*\* |
| Kuperman\_AoA\_CW | 4.633 | 7.246 | 5.817 | 0.510 | .293\*\* |
| Kuperman\_AoA\_AW | 4.319 | 5.881 | 5.108 | 0.305 | .292\*\* |
| MRC Meaningfulness\_FW | 276.472 | 399.483 | 330.610 | 24.058 | -.280\*\* |
| MRC Concreteness\_FW | 230.029 | 344.500 | 280.452 | 22.288 | -.279\*\* |
| MRC Imageability\_FW | 244.667 | 356.909 | 292.928 | 22.074 | -.246\*\* |
| COCA fiction\_Range\_AW | 0.491 | 0.726 | 0.633 | 0.044 | -.241\*\* |
| MRC Imageability\_CW | 312.720 | 458.063 | 373.458 | 29.359 | .230\*\* |
| MRC Concreteness\_CW | 275.217 | 437.950 | 340.731 | 32.274 | .229\*\* |
| Aoe index above threshold\_40 | 0.513 | 3.893 | 1.843 | 0.615 | .209\*\* |
| COCA fiction Frequency\_AW | 3825.548 | 12078.653 | 7540.064 | 1434.051 | .203\*\* |
| MRC Meaningfulness\_AW | 308.850 | 401.895 | 357.410 | 15.867 | -.196\*\* |
| COCA fiction\_Frequency\_Log\_AW | 2.453 | 3.294 | 2.955 | 0.151 | -.193\*\* |
| MRC Familiarity\_AW | 577.188 | 605.535 | 593.732 | 4.745 | -.190\*\* |
| COCA fiction\_Frequency\_CW | 278.240 | 2485.485 | 1350.075 | 415.266 | -.188\*\* |
| MRC Familiarity\_CW | 543.303 | 598.941 | 576.662 | 9.023 | -.183\*\* |
| COCA fiction Bigram Range\_Log | -1.566 | -0.950 | -1.192 | 0.102 | -.149\* |
| COCA fiction Frequency\_Log\_FW | 3.639 | 4.185 | 3.967 | 0.097 | .149\* |
| MRC Familiarity\_FW | 595.333 | 622.476 | 609.990 | 3.854 | -.127\* |
| COCA fiction Bigram\_Frequency\_Log | 0.939 | 1.686 | 1.403 | 0.126 | -.122\* |
| ***Fine-grained syntactic complexity indices*** |  |  |  |  |  |
| Dependents per nominal | 0.174 | 1.619 | 0.724 | 0.257 | .251\*\* |
| Nominal subjects per clause | 0.333 | 1.154 | 0.755 | 0.135 | -.223\*\* |
| Determiners per nominal | 0.000 | 0.556 | 0.225 | 0.105 | .210\*\* |
| Dependents per object of the preposition | 0.000 | 2.400 | 1.009 | 0.382 | .204\*\* |
| Possessives per nominal | 0.000 | 0.243 | 0.086 | 0.059 | .198\*\* |
| Adjectival modifiers per object of the preposition | 0.000 | 0.667 | 0.134 | 0.131 | .173\*\* |
| Dependents per direct object | 0.000 | 3.000 | 0.939 | 0.534 | .167\*\* |
| Prepositions per direct object | 0.000 | 1.000 | 0.154 | 0.188 | .165\*\* |
| Prepositions per nominal | 0.000 | 0.389 | 0.094 | 0.071 | .161\*\* |
| Clausal negations per clause | 0.000 | 0.400 | 0.075 | 0.075 | -.158\*\* |
| Determiners per direct object | 0.000 | 1.000 | 0.299 | 0.239 | .153\* |
| Adjectival modifiers per object of the preposition (no pronouns) | 0.000 | 1.000 | 0.150 | 0.152 | .147\* |
| Possessives per object of the preposition | 0.000 | 0.667 | 0.140 | 0.138 | .142\* |
| Dependents per object of the preposition (no pronouns, standard deviation) | 0.000 | 1.730 | 0.756 | 0.268 | .140\* |
| Dependents per direct object (no pronouns) | 0.000 | 3.000 | 1.125 | 0.597 | .135\* |
| Discourse markers per clause | 0.000 | 0.167 | 0.014 | 0.032 | -.131\* |
| Determiners per direct object (no pronouns) | 0.000 | 1.000 | 0.395 | 0.295 | .131\* |
| Adjectival modifiers per nominal | 0.000 | 0.455 | 0.099 | 0.073 | .129\* |
| Direct objects per clause | 0.000 | 0.889 | 0.342 | 0.139 | .122\* |
| (non-clausal) adverbial modifiers per object of the preposition | 0.000 | 0.400 | 0.025 | 0.060 | .120\* |
| ***Global syntactic complexity indices*** |  |  |  |  |  |
| Complex nominals per clause  | 0.118 | 2.600 | 0.723 | 0.416 | .346\*\* |
| Mean length of clause  | 4.118 | 21.500 | 7.762 | 2.381 | .339\*\* |
| Mean length of sentences | 3.316 | 37.000 | 12.911 | 6.194 | .299\*\* |
| Complex nominals per T-unit | 0.182 | 5.667 | 1.123 | 0.783 | .290\*\* |
| Mean dependence distance  | 1.940 | 4.910 | 2.948 | 0.545 | .209\*\* |
| ***Discoursal complexity indices*** |  |  |  |  |  |
| Repeated content lemmas | 0.000 | 0.297 | 0.115 | 0.057 | -.242\*\* |
| Number of determiners | 0.000 | 0.204 | 0.089 | 0.033 | .225\*\* |
| Binary adjacent sentence overlap function lemmas | 0.000 | 1.000 | 0.570 | 0.280 | .222\*\* |
| Adjacent sentence overlap function lemmas (sentence normed) | 0.000 | 5.667 | 1.214 | 1.057 | .218\*\* |
| Adjacent sentence overlap all lemmas (sentence normed) | 0.000 | 8.000 | 1.636 | 1.388 | .195\*\* |
| Binary adjacent sentence overlap noun and pronoun lemmas | 0.000 | 1.000 | 0.414 | 0.276 | .190\*\* |
| Binary adjacent sentence overlap all lemmas | 0.000 | 1.000 | 0.639 | 0.264 | .175\*\* |
| Adjacent sentence overlap function lemmas | 0.000 | 0.500 | 0.197 | 0.104 | .168\*\* |
| Adjacent sentence overlap noun and pronoun lemmas (sentence normed) | 0.000 | 2.667 | 0.552 | 0.464 | .167\*\* |
| Binary adjacent sentence overlap noun lemmas | 0.000 | 1.000 | 0.112 | 0.193 | .165\*\* |
| Repeated content lemmas and pronouns | 0.000 | 0.403 | 0.186 | 0.067 | -.155\* |
| Adjacent sentence overlap noun lemmas (sentence normed) | 0.000 | 2.500 | 0.144 | 0.292 | .154\* |
| Adjacent sentence overlap verb lemmas | 0.000 | 0.333 | 0.056 | 0.074 | -.137\* |
| Binary adjacent sentence overlap pronoun lemmas | 0.000 | 1.000 | 0.379 | 0.278 | .134\* |
| Adjacent sentence overlap adverb lemmas | 0.000 | 0.500 | 0.041 | 0.090 | -.121\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Cumulative variance explained by two extracted components: mean fixation duration in GECO |
| Components | *X*-Explained Variance | *Y*-Explained Variance | R2 |
| 1 | 27.928 | 20.772  | 0.208 |
| 2 | 33.641 | 30.123 | 0.301 |

**Appendix 4: Correlations between lexical, syntactic, and discoursal complexity indices and mean regression rates of 272 snippets in GECO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with RR |
| ***Lexical sophistication indices*** |
| COCA fiction Range\_AW | 0.491 | 0.726 | 0.633 | 0.044 | .332\*\* |
| COCA fiction Frequency\_Log\_AW | 2.453 | 3.294 | 2.955 | 0.151 | .324\*\* |
| COCA fiction Range\_Log\_AW | -0.789 | -0.220 | -0.452 | 0.101 | .310\*\* |
| COCA fiction Range\_Log\_CW | -1.178 | -0.384 | -0.748 | 0.163 | .289\*\* |
| COCA fiction Frequency\_Log\_CW | 1.503 | 2.695 | 2.217 | 0.235 | .280\*\* |
| COCA fiction Range\_CW | 0.229 | 0.589 | 0.426 | 0.072 | .275\*\* |
| Age of Acquisition AW | 4.319 | 5.881 | 5.108 | 0.305 | -.274\*\* |
| Age of Acquisition FW | 3.763 | 4.835 | 4.273 | 0.182 | -.261\*\* |
| MRC Familiarity AW | 577.188 | 605.535 | 593.732 | 4.745 | .258\*\* |
| MRC Concreteness\_FW | 230.029 | 344.500 | 280.452 | 22.288 | .240\*\* |
| MRC Familiarity\_CW | 543.303 | 598.941 | 576.662 | 9.023 | .224\*\* |
| COCA fiction Frequency\_CW | 278.240 | 2485.485 | 1350.075 | 415.266 | .224\*\* |
| MRC Imageability\_FW | 244.667 | 356.909 | 292.928 | 22.074 | .207\*\* |
| MRC Meaningfulness\_FW | 276.472 | 399.483 | 330.610 | 24.058 | .199\*\* |
| COCA fiction Bigram Range\_Log | -1.566 | -0.950 | -1.192 | 0.102 | .193\*\* |
| COCA fiction Bigram Frequency\_Log | 0.939 | 1.686 | 1.403 | 0.126 | .175\*\* |
| Aoe index above threshold\_40 | 0.513 | 3.893 | 1.843 | 0.615 | -.170\*\* |
| MRC Concreteness\_CW | 275.217 | 437.950 | 340.731 | 32.274 | -.150\* |
| MRC Meaningfulness\_AW | 308.850 | 401.895 | 357.410 | 15.867 | .132\* |
| COCA fiction trigram Range\_Log | -2.326 | -1.473 | -1.874 | 0.133 | .131\* |
| COCA fiction trigram Frequency\_Log | 0.069 | 1.024 | 0.546 | 0.146 | .130\* |
| ***Fine-grained syntactic complexity indices*** |  |  |  |  |  |
| Nouns as a nominal dependent per nominal | 0.000 | 0.320 | 0.059 | 0.059 | -.235\*\* |
| Nouns as a nominal subject dependent per nominal subject | 0.000 | 0.714 | 0.057 | 0.091 | -.205\*\* |
| Dependents per nominal | 0.174 | 1.619 | 0.724 | 0.257 | -.181\*\* |
| Dependents per nominal subject | 0.000 | 1.200 | 0.313 | 0.240 | -.168\*\* |
| Prepositions per clause | 0.000 | 1.000 | 0.314 | 0.170 | -.165\*\* |
| Adjectival modifiers per nominal | 0.000 | 0.455 | 0.099 | 0.073 | -.164\*\* |
| Nouns as a nominal dependent per nominal (no pronouns) | 0.000 | 0.400 | 0.092 | 0.084 | -.163\*\* |
| Adjectival modifiers per direct object | 0.000 | 2.000 | 0.146 | 0.217 | -.157\*\* |
| Nouns as a nominal subject dependent per nominal subject (no pronouns) | 0.000 | 1.250 | 0.151 | 0.233 | -.148\* |
| Adjectival modifiers per direct object (no pronouns) | 0.000 | 2.000 | 0.193 | 0.272 | -.143\* |
| Nouns as a object of the preposition dependent per object of the preposition | 0.000 | 0.500 | 0.065 | 0.105 | -.142\* |
| Clausal negations per clause | 0.000 | 0.400 | 0.075 | 0.075 | .142\* |
| Nouns as a object of the preposition dependent per object of the preposition (no pronouns) | 0.000 | 0.500 | 0.070 | 0.112 | -.134\* |
| Passive nominal subjects per clause | 0.000 | 0.389 | 0.059 | 0.073 | -.132\* |
| Nominal subjects per clause | 0.333 | 1.154 | 0.755 | 0.135 | .132\* |
| Dependents per object of the preposition (no pronouns, standard deviation) | 0.000 | 1.730 | 0.756 | 0.268 | -.127\* |
| Conjunction “or” as a nominal dependent per nominal | 0.000 | 0.074 | 0.003 | 0.011 | .127\* |
| Passive auxiliary verbs per clause | 0.000 | 0.400 | 0.059 | 0.071 | -.121\* |
| Determiners per nominal subject | 0.000 | 0.500 | 0.081 | 0.100 | -.119\* |
| ***Global syntactic complexity indices*** |  |  |  |  |  |
| Mean length of clause  | 4.118 | 21.500 | 7.762 | 2.381 | -.179\*\* |
| Complex nominals per clause | 0.118 | 2.600 | 0.723 | 0.416 | -.176\*\* |
| Complex nominals per T-unit | 0.182 | 5.667 | 1.123 | 0.783 | -.146\* |
| Mean length of sentence | 3.316 | 37.000 | 12.911 | 6.194 | -.135\* |
| ***Discoursal complexity indices*** |  |  |  |  |  |
| Adjacent sentence overlap noun and pronoun lemmas | 0.000 | 0.364 | 0.140 | 0.085 | .132\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Cumulative variance explained by two extracted components: regression rates in GECO |
| Components | *X*-Explained Variance | *Y*-Explained Variance | R2 |
| 1 | 29.214 | 15.040 | 0.150 |
| 2 | 35.329 | 25.868 | 0.259  |

**Appendix 5: Correlations between lexical, syntactic, and discoursal complexity indices and skipping rates of 272 snippets in GECO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linguistic complexity indices | Min | Max | Mean | SD | Correlations with SR |
| ***Lexical sophistication indices*** |  |  |  |  |  |
| Character bigram Frequency | 2883.466 | 4546.783 | 3690.731 | 306.446 | -.125\* |
| Character bigram Frequency\_CW | 2653.585 | 4347.391 | 3558.401 | 295.055 | -.143\* |
| ***Fine-grained syntactic complexity indices*** |  |  |  |  |  |
| Prepositions per nominal | 0.000 | 0.389 | 0.094 | 0.071 | -.154\* |
| Adjectival modifiers per nominal | 0.000 | 0.455 | 0.099 | 0.073 | -.152\* |
| Determiners per nominal (no pronouns) | 0.000 | 0.667 | 0.340 | 0.130 | -.123\* |
| Adjectival modifiers per nominal (no pronouns) | 0.000 | 0.714 | 0.154 | 0.113 | -.144\* |
| Prepositions per nominal (no pronouns) | 0.000 | 0.462 | 0.129 | 0.097 | -.142\* |
| Nouns as a nominal dependent per nominal (no pronouns) | 0.000 | 0.400 | 0.092 | 0.084 | .130\* |
| Prepositions per nominal subject | 0.000 | 0.375 | 0.040 | 0.069 | -.165\*\* |
| Nouns as a nominal subject dependent per nominal subject | 0.000 | 0.714 | 0.057 | 0.091 | .156\*\* |
| Prepositions per nominal subject (no pronouns) | 0.000 | 1.000 | 0.110 | 0.199 | -.172\*\* |
| Nouns as a nominal subject dependent per nominal subject (no pronouns) | 0.000 | 1.250 | 0.151 | 0.233 | .161\*\* |
| Adjectival modifiers per object of the preposition | 0.000 | 0.667 | 0.134 | 0.131 | -.128\* |
| Existential "there" per clause | 0.000 | 0.235 | 0.025 | 0.044 | -.170\*\* |
| Instances of parataxis per clause | 0.000 | 0.250 | 0.029 | 0.046 | -.149\* |
| ***Discoursal complexity indices*** |  |  |  |  |  |
| Overlap of adjective lemmas across adjacent sentences (sentence normed) | 0.000 | 0.333 | 0.012 | 0.047 | .134\* |
| Overlap of adjective lemmas across adjacent sentences | 0.000 | 0.500 | 0.017 | 0.069 | .122\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Cumulative variance explained by two extracted components: skipping rate in GECO |
| Components | *X-*Explained Variance | *Y-*Explained Variance | R2 |
| 1 | 16.995  | 14.078  | 0.141 |
| 2 | 26.611  | 17.828  | 0.178 |

**Appendix 6: Correlations between syntactic, and discoursal complexity indices and mean saccade amplitude of 272 snippets in GECO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with MSA  |
| ***Lexical sophistication indices*** |  |  |  |  |  |
| COCA\_fiction\_Frequency\_AW | 3825.548 | 12078.653 | 7540.064 | 1434.051 | .210\*\* |
| COCA\_fiction\_Frequency\_FW | 8680.097 | 24370.123 | 15993.874 | 3004.883 | .200\*\* |
| COCA Fiction bigram Frequency | 52.218 | 618.356 | 209.143 | 84.278 | .245\*\* |
| COCA\_Fiction\_Bigram\_Range | 0.094 | 0.282 | 0.182 | 0.028 | .136\* |
| COCA\_fiction\_bigram \_T | -38.041 | 103.456 | 51.901 | 14.988 | .211\*\* |
| COCA\_fiction\_bigram \_DP | 0.014 | 0.076 | 0.040 | 0.013 | .131\* |
| COCA\_fiction\_trigram\_MI | 1.445 | 4.207 | 2.458 | 0.346 | -.133\* |
| ***Global syntactic complexity indices*** |  |  |  |  |  |
| Mean length of sentence | 3.316 | 37.000 | 12.911 | 6.194 | .156\* |
| Mean length of T-unit | 4.267 | 37.000 | 11.759 | 4.892 | .132\* |
| Mean length of clause | 4.118 | 21.500 | 7.762 | 2.381 | .142\* |
| Complex nominals per T-unit | 0.182 | 5.667 | 1.123 | 0.783 | .127\* |
| Complex nominals per clause  | 0.118 | 2.600 | 0.723 | 0.416 | .138\* |
| ***Fine-grained syntactic complexity indices*** |  |  |  |  |  |
| Dependents per nominal | 0.174 | 1.619 | 0.724 | 0.257 | .135\* |
| Dependents per object of the preposition (no pronouns) | 0.000 | 2.250 | 0.991 | 0.385 | .180\*\* |
| Dependents per object of the preposition (standard deviation) | 0.000 | 1.920 | 0.788 | 0.267 | .141\* |
| Dependents per object of the preposition (no pronouns, standard deviation) | 0.000 | 1.730 | 0.756 | 0.268 | .189\*\* |
| Determiners per nominal | 0.000 | 0.556 | 0.225 | 0.105 | .151\* |
| Prepositions per nominal | 0.000 | 0.389 | 0.094 | 0.071 | .120\* |
| Determiners per nominal (no pronouns) | 0.000 | 0.667 | 0.340 | 0.130 | .120\* |
| Determiners per object of the preposition | 0.000 | 1.000 | 0.349 | 0.187 | .183\*\* |
| Determiners per object of the preposition (no pronouns) | 0.000 | 1.000 | 0.387 | 0.196 | .161\*\* |
| Passive auxilliary verbs per clause | 0.000 | 0.400 | 0.059 | 0.071 | .126\* |
| ***Discoursal complexity indices*** |  |  |  |  |  |
| Adjacent sentence overlap all lemmas | 0.000 | 0.357 | 0.131 | 0.061 | .148\* |
| Adjacent sentence overlap all lemmas (sentence normed) | 0.000 | 8.000 | 1.636 | 1.388 | .146\* |
| Binary adjacent sentence overlap all lemmas | 0.000 | 1.000 | 0.639 | 0.264 | .184\*\* |
| adjacent sentence overlap content lemmas | 0.000 | 0.212 | 0.049 | 0.045 | .169\*\* |
| Adjacent sentence overlap content lemmas (sentence normed) | 0.000 | 2.500 | 0.322 | 0.398 | .185\*\* |
| Binary adjacent sentence overlap content lemmas | 0.000 | 1.000 | 0.229 | 0.233 | .194\*\* |
| Adjacent sentence overlap function lemmas (sentence normed) | 0.000 | 5.667 | 1.214 | 1.057 | .121\* |
| Binary adjacent sentence overlap function lemmas | 0.000 | 1.000 | 0.570 | 0.280 | .162\*\* |
| Adjacent sentence overlap noun and pronoun lemmas (sentence normed) | 0.000 | 2.667 | 0.552 | 0.464 | .150\* |
| Binary adjacent sentence overlap noun and pronoun lemmas | 0.000 | 1.000 | 0.414 | 0.276 | .139\* |
| Determiners | 0.000 | 0.204 | 0.089 | 0.033 | .178\*\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Cumulative variance explained by two extracted components: Mean saccade amplitude in GECO |
| Components | X Explained Variance | Y Explained Variance | R2 |
| 1 | 33.016 | 10.998  | 0.110 |
| 2 | 46.830  | 17.049 | 0.170  |

**Appendix 7: Correlations between L2 learners’ eye-movement measures and eight readability scores of 272 snippets in GECO**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Formulas | Mean | SD | Correlations with RRate | Correlations with MFD | Correlations with MSA | Correlations with RR | Correlations with SR |
| FRE | 76.920 | 11.686 | -0.450 | -0.284 | -0.052 | 0.282 | 0.099 |
| FKGL | 5.471 | 2.830 | 0.560 | 0.319 | 0.057 | -0.224 | -0.069 |
| ARI | 4.877 | 3.639 | 0.573 | 0.322 | 0.055 | -0.258 | -0.070 |
| SMOG | 8.474 | 2.034 | 0.482 | 0.297 | 0.034 | -0.252 | -0.047 |
| NDC | 7.256 | 0.971 | 0.323 | 0.220 | 0.051 | -0.317 | 0.031 |
| CAREC | 0.100 | 0.068 | 0.150 | 0.068 | 0.109 | -0.147 | 0.024 |
| CARES | 0.179 | 0.074 | 0.124 | 0.203 | 0.021 | -0.147 | 0.020 |
| CML2RI | 17.775 | 6.138 | -0.267 | -0.311 | 0.011 | 0.277 | 0.073 |
| *Note*. RRate= reading rate; FDM = mean fixation duration; MSA= Mean saccade amplitude; RR= regression rate; SR = skipping rate; FRE = Flesch Reading Ease formula; FKGL = Flesch-Kincaid Grade Level formula; ARI= Automated Readability Index; SMOG= Simple Measure of Gobbledygook; NDC= New-Dale Chall; CML2RI = Coh-Metrix L2 Reading Index; CAREC = Crowdsourced Algorithm of Reading Comprehension; CARES = Crowdsourced Algorithm of Reading Speed |

**Appendix 8: Readability scores of 12 paragraphs in MECO-L2, mean and standard deviation of the variables in the PLS-R variables for the five eye-movement measures**

|  |
| --- |
| Correlations between L2 learners’ eye-movement measures and eight readability scores of 12 paragraphs in MECO-L2 (for RQ3) |
| Readability formulas | Mean | SD | Correlations with RRate | Correlations with MFD | Correlations with MSA | Correlations with RR | Correlations with SR |
| FRE | 52.635  | 14.303  | 0.103 | -.265 | -0.018 | .167 | .129 |
| FKGL | 10.647  | 2.682  | -0.082 | .256 | -0.029 | -.155 | -0.055 |
| ARI | 11.344  | 2.968  | -0.103 | .248 | -0.053 | -.166 | -0.075 |
| SMOG | 11.833  | 2.329  | -0.031 | .240 | -0.012 | -.105 | -0.034 |
| NDC | 8.627  | 1.107  | -.154 | .279 | -0.011 | -.150 | -0.079 |
| CAREC | 0.167  | 0.070  | -.146 | .142 | -0.016 | 0.007 | -.108 |
| CARES | 0.410  | 0.103  | 0.026 | .181 | .120 | -0.097 | -.164 |
| CML2RI | 15.026  | 4.536  | 0.007 | -0.052 | 0.04 | .165 | 0.036 |
| *Note*. RRate= reading rate; FDM = mean fixation duration; MSA= Mean saccade amplitude; RR= regression rate; SR = skipping rate; FRE = Flesch Reading Ease formula; FKGL = Flesch-Kincaid Grade Level formula; ARI= Automated Readability Index; SMOG= Simple Measure of Gobbledygook; NDC= New-Dale Chall; CML2RI = Coh-Metrix L2 Reading Index; CAREC = Crowdsourced Algorithm of Reading Comprehension; CARES = Crowdsourced Algorithm of Reading Speed |

|  |
| --- |
| Mean and standard deviation of variables in the PLS-R model for mean fixation duration in MECO-L2 (for RQ3) |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with mean fixation duration |
| Kuperman AoA\_AW | 5.237 | 6.324 | 5.679 | 0.304 | .163\*\* |
| Kuperman AoA\_CW | 6.004 | 7.356 | 6.393 | 0.399 | .239\*\* |
| COCA fiction Range\_CW | 0.234 | 0.409 | 0.339 | 0.051 | -.322\*\* |
| COCA fiction Frequency\_Log\_CW | 1.577 | 2.195 | 1.934 | 0.170 | -.283\*\* |
| AoE index above threshold\_40 | 1.792 | 3.449 | 2.688 | 0.506 | 0.124\* |
| Repeated content lemmas | 0.125 | 0.310 | 0.229 | 0.053 | -0.011 |
| Mean length of sentence | 12.909 | 29.400 | 18.968 | 4.822 | .106\* |
| Mean length of clause | 7.474 | 16.333 | 10.172 | 2.481 | 0.078 |
| Complex nominals per clause | 0.824 | 3.222 | 1.275 | 0.591 | .135\* |
| MRC Concreteness\_CW | 323.038 | 398.393 | 363.007 | 24.790 | 0.066 |
| Nominal subjects per clause | 0.235 | 0.850 | 0.598 | 0.170 | .147\*\* |
| Lemma overlap between adjacent sentences  | 1.857 | 5.200 | 3.200 | 1.141 | .139\*\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Mean and standard deviation of the variables in the PLS-R model for regression rate in MECO-L2 (for RQ3) |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with regression rate |
| Word length  | 4.452 | 5.852 | 4.942 | 0.408 | -.166\*\* |
| Kuperman AoA\_FW | 4.245 | 4.664 | 4.457 | 0.118 | -0.096 |
| COCA fiction Range\_AW | 0.468 | 0.605 | 0.543 | 0.038 | .198\*\* |
| COCA fiction Frequency\_Log\_AW | 2.393 | 2.833 | 2.642 | 0.134 | .158\*\* |
| COCA fiction Range\_CW | 0.234 | 0.409 | 0.339 | 0.051 | .221\*\* |
| COCA fiction Frequency\_Log\_CW | 1.577 | 2.195 | 1.934 | 0.170 | .178\*\* |
| COCA fiction bigram Frequency\_Log | 60.230 | 482.666 | 187.045 | 121.547 | 0.006 |
| Nouns as a nominal dependent per nominal | 0.000 | 0.147 | 0.067 | 0.040 | -.186\*\* |

\*\**p* < .01

|  |
| --- |
| Mean and standard deviation of the variables in the PLS-R model for reading rate in MECO-L2 (for RQ3) |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with reading rate |
| Mean length of sentence | 12.909  | 29.400  | 18.968  | 4.822  | -0.004 |
| Mean length of clause  | 7.474  | 16.333  | 10.172  | 2.481  | -0.033 |
| Complex nominals per clause  | 0.824  | 3.222  | 1.275  | 0.591  | -0.140\* |
| Complex nominals per T-unit  | 1.300  | 4.143  | 1.989  | 0.790  | -0.128\* |
| MRC Meaningfulness\_FW | 263.787  | 305.632  | 284.830  | 15.292  | .113\* |
| All lemma overlap between adjacent sentences  | 1.857  | 5.200  | 3.205  | 1.143  | -0.046 |
| All lemmas overlap between binary adjacent sentence | 0.750  | 1.000  | 0.940  | 0.080  | -.162\*\* |
| Function lemmas overlap between adjacent sentences  | 0.714  | 4.400  | 1.795  | 1.069  | 0.032 |
| Function lemmas overlap between binary adjacent sentences  | 0.400  | 1.000  | 0.753  | 0.205  | -0.028 |
| Noun and pronoun lemmas overlap between adjacent sentences | 0.333  | 1.400  | 0.873  | 0.394  | -.167\*\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Mean and standard deviation of the variables in the PLS-R model for skipping rate in MECO-L2 (for RQ3) |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with skipping rate |
| Word length | 4.452 | 5.852 | 4.942 | 0.408 | -.241\*\* |
| Adjectival modifiers per nominal | 0.143 | 0.449 | 0.268 | 0.100 | -.245\*\* |
| AOE inverse linear regression slope | 1.037  | 1.426  | 1.173  | 0.114  | 0.064 |
| Determiners per nominal (no pronouns) | 0.188 | 0.516 | 0.310 | 0.086 | 0.041 |
| Adjectival modifiers per nominal (no pronouns) | 0.163 | 0.536 | 0.291 | 0.108 | -.230\*\* |
| Existential “there” per clause | 0.000 | 0.133 | 0.026 | 0.040 | -.117\* |

\**p* < .05, \*\**p* < .01

|  |
| --- |
| Mean and standard deviation of variables in the PLS-R model for mean saccade amplitude in MECO-L2 (for RQ3) |
| Linguistic complexity indices | Minimum | Maximum | Mean | SD | Correlations with mean saccade amplitude |
| COCA Fiction bigram Frequency | 60.230  | 482.666  | 186.554  | 121.323  | -0.092 |
| Adjacent sentence overlap all lemmas | 0.137  | 0.236  | 0.191  | 0.036  | -.144\*\* |
| Adjacent sentence overlap content lemmas | 0.055  | 0.235  | 0.133  | 0.059  | -0.055 |
| Adjacent sentence overlap content lemmas (sentence normed) | 0.429  | 2.400  | 1.304  | 0.568  | -0.078 |
| Binary adjacent sentence overlap content lemmas | 0.286  | 1.000  | 0.703  | 0.222  | -0.027 |
| Adjacent sentence overlap noun and pronoun lemmas (sentence normed) | 0.333  | 1.400  | 0.872  | 0.394  | -.128\* |

\**p* < .05, \*\**p* < .01