# The Common Readability Formula & Five Adjusted Readability Formulas for Text Simplification, Medical Documents and Other General Uses

Anonymous ACL submission

#### Abstract

Traditional readability formulas, or equations, are inaccurate and measure highly limited linguistic properties. Despite the recent machine learning-based readability assessment models, many researchers insist on using the outdated formulas. To replace the linguistically-shallow inaccurate formulas, we: : **1.** introduce Common Readability Formula (CoRF), **2.** recalibrate outdated formulas (Flesch-Kincaid Grade Level, Fog Index, SMOG Index, Coleman-Liau Index, and Automated Readability Index), **3.** evaluate the formulas, and **4.** develop a Python library for the wide dispersal of our variations.

#### 1 Introduction

002

007

011

016

017

022

026

028

037

#### 1.1 Contrasting Improvements

Readability Assessment (RA) quantitatively measures the ease of understanding or comprehension of any written text (Feng et al., 2010; Klare, 2000). Understanding text readability, or difficulty, is essential for research on any originated, studied, or shared ideas (Collins-Thompson, 2014). Such inherent property leads to RA's close applications to various areas of healthcare (Wu et al., 2013), education (Dennis, 2018), communication (Zhou et al., 2017), and Natural Language Processing (NLP), such as text simplification (Aluisio et al., 2010). Therefore, the performances of state-of-theart (SOTA) models in RA much impact the related fields. At least, that is supposed to be the case.

But it is rare to find examples where a SOTA model in RA is actively applied out of paper. For some reason, researchers insist on the so-called "traditional readability formulas." This created a big performance deviation between RA and the other fields. The most recent papers – that need RA as a supporting tool – fall back to the traditional formulas, despite the high-performance neural models.

The lack of accuracy is not a reason. Machine learning (ML) or transformer-based methods are successful in RA. The RoBERTa-RF-T1 model

by Lee et al. (2021) achieves a 99% classification accuracy on OneStopEnglish dataset (Vajjala and Lučić, 2018) and a BERT-based ReadNet model from Meng et al. (2020) achieves about 92% accuracy on WeeBit dataset (Vajjala and Meurers, 2012); both datasets are well-regarded in the field. Moreover, comparable results have been published since 2010s (Xia et al., 2016; Feng et al., 2010). Compared to these recent RA models, the "traditional readability formulas" are broadly known to have lower accuracy. They have been criticized by academia (Feng et al., 2010) and plain audiences (Jarrett and Redish, 2019) for more than a decade. So we, as researchers working on RA, questioned, "Why do people insist on the outdated formulas?"

041

042

043

044

045

047

049

050

051

055

056

060

061

062

063

064

065

066

067

069

070

071

073

074

075

076

077

078

079

### 1.2 Traditional Readability Formula

The earliest attempt to "calculate" text readability was by Lively and Pressey (1923), in response to their practical problem of selecting science textbooks for high school students (DuBay, 2004). In the consecutive years, many well-known readability formulas were developed, including Flesch Kincaid Grade Level (Kincaid et al., 1975), Gunning Fog Count (or Index) (Gunning et al., 1952), SMOG Index (Mc Laughlin, 1969), Coleman-Liau Index (Coleman and Liau, 1975), and Automated Readability Index (Smith and Senter, 1967). These formulas are mostly linear models with two or three variables, largely based on superficial properties concerning words or sentences (Feng et al., 2010). Though such formula-based methods are unreliable in terms of accuracy, they bring critical benefits.

Most importantly, traditional readability formulas provide generalized definitions of readability, like a classical physics equation. Their value is not in the most precise modeling of a specific dataset. Thus, these formulas must not be underestimated by simple comparisons of accuracy. They are also intuitive. A user can exactly follow which properties accumulated to affect the score. In addition,

176

177

178

130

they are both qualitatively and quantitatively motivated. Thus, they utilize linguistically meaningful features (though limited coverage), not computer-generated sequences (Collins-Thompson, 2014). This makes them dataset-specific. Further, they can easily combine with other systems with less burden of a large trained model (Xu et al., 2016). In short, traditional readability formulas provide:
1. generalized definitions of readability
2. intuitive approaches ∴ "appear" trustworthy
3. easiness in dispersal and implementation

081

087

091

093

096

102

103

104

105

106

107

109

110

111

112

113

114

115

116

117

118

119

121

122

123

124

125

126

127

128

129

Another concern in RA is that it is difficult to tell if two different datasets, with readability labeled by two different human experts, have consistent measures of readability. This troubles the generalization of machine-trained RA models. Further, a linguistically rich dataset that covers all textual variations that affect readability (Collins-Thompson, 2014) is yet to be developed. Also, to make such a dataset suitable for RA training, multiple human experts should go through every item to ensure the agreeable labeling of each item's readability. Without a collective effort for such an ideal dataset, traditional readability formulas will prolong.

#### 1.3 Proposed Solution & Impact

We propose a SOTA-less approach of recalibrating five traditional readability formulas, mostly for the non-experts in readability. We utilize the appendix B (Text Exemplars and Sample Performance Tasks) dataset, provided by the U.S. Common Core State Standards <sup>1</sup>. The five traditional readability formulas – mainly developed upon 20th century's military or technical documents – are adjusted for the modern, standard U.S. education curriculum.

But traditional readability formulas lack wide linguistic coverage (Feng et al., 2010). Therefore, we create a *new formula* that is mainly motivated by lexico-semantic and syntactic linguistic branches, as identified by Collins-Thompson (2014). From each, we search for the representative features. The resulting formula is named the Common Readability Formula, or simply **CoRF**, since it aims to give the most generally and commonly acceptable approach to calculate readability.

The necessity for this study has been raised by the general research community. They need intuitive approaches to readability instead of complex models, mostly appreciated by RA experts. We respond with a linguistically-rich alternative to the traditional readability formulas. Regrettably, the inaccurate (Jiang et al., 2015) and linguisticallylacking formulas are already widespread in the community, and it requires immediate attention.

To mention a fragment of what we found, there are FKBLEU (Xu et al., 2016), recommendation quality assessment (Schwartz et al., 2017), simplification for specific audience (Scarton and Specia, 2018), clinical letters' text simplification (Shardlow and Nawaz, 2019), authorship attribution (Uchendu et al., 2020). These are all recent \*ACL main conference or TACL papers that **use outdated formulas as a part of their evaluation tool or as a sole measure of readability**. Existing formulas must be improved and a new variation must be developed.

Outside computational linguistics, the problem worsens. Some research is directly related to the public medical knowledge, like measuring the difficulty of a patient information material (Gaeta et al., 2021; van Ballegooie and Hoang, 2021; Bange et al., 2019; Haller et al., 2019; Hansberry et al., 2018; Kiwanuka et al., 2017).

#### 1.4 Contributions & Public Resources

To sum up, we make the contributions below. The related public resources are linked in appendix A. **1.** We recalibrate five traditional readability formulas to show higher prediction accuracy on modern texts in the U.S. curriculum.

**2.** We develop CoRF, a generalized and easy-to-approach solution to readability.

**3.** We evaluate the uses and limits of readability formulas on several datasets.

**4.** We develop <Anonymous>, a fast open-source readability assessment software based on Python.

#### 2 Datasets

### 2.1 Common Core - Appendix B (CCB)

We use the CCB corpus to calibrate formulas. The article excerpts included in CCB are divided into the categories of story, poetry, informational text, and drama. For the simplification of our approach, we limit our research to story-type texts. This left us with only 69 items to train with. But those are directly from the U.S. Common Core Standards. Hence, we assume with confidence that the item classification is generally agreeable in the U.S.

We also check the human-perceived difficulty of each item. We used Amazon Mechanical Turk to ask U.S. Bachelor's degree holders, "Which U.S. grade does this text belong to?" Every item was

<sup>&</sup>lt;sup>1</sup>corestandards.org

| Properties  | ССВ   | WBT   | CAM   | CKC   | OSE  | NSL  |
|-------------|-------|-------|-------|-------|------|------|
| audience    | Ntve  | Ntve  | ESL   | ESL   | ESL  | Ntve |
| grade       | K1-12 | K2-10 | A2-C2 | S7-12 | N/A  | N/A  |
| curriculum? | Yes   | No    | Yes   | Yes   | No   | No   |
| balanced?   | No    | Yes   | Yes   | No    | Yes  | No   |
| #class      | 6     | 5     | 5     | 6     | 3    | 5    |
| #item/class | 11.5  | 625   | 60.0  | 554   | 189  | 2125 |
| #word/item  | 362   | 213   | 508   | 117   | 669  | 752  |
| #sent/item  | 25.8  | 17.0  | 28.4  | 54.0  | 35.6 | 50.9 |

Table 1: Modified data. These stats are based on respective original versions. S: S.Korea Grade, Ntve: Native

answered by 10 different workers to ensure breadth. Details on survey & datasets are in appendix B, C.

CCB is the only dataset that we use in the calibration of our formulas. All below datasets are mainly for feature selection purposes only. Also, "independent" refers to the wide form data where each row consists of an item, class, and other properties.

#### 2.2 WeeBit (WBT) - independent

179

181

182

184

185

186

188

189

190

191

193

194

195

196

197

198

199

200

204

205

WBT, the largest native dataset available in RA, contains articles targeted at readers of different age groups from the Weekly Reader magazine and the BBC-Bitesize website. In table 1, we translate those age groups into U.S. schools' K-\* format. We downsample to  $625 \frac{item}{class}$  as per common practice.

## 2.3 Cambridge English (CAM) - independent

CAM (Xia et al., 2016) classifies 300 items in Common European Framework of Reference (CEFR) (Verhelst et al., 2001). The passages are from the past reading tasks in the five main suite Cambridge English Exams (KET, PET, FCE, CAE, CPE), targeted at learners at A2–C2 levels of CEFR.

#### 2.4 Corpus of the Korean ELT (English Lang. Train.) Curriculum (CKC) - independent

CKC (Lee and Lee, 2020b,a) is less-explored. It developed upon the reading passages appearing in the Korean English education curriculum. These passages' classifications are from the official sources from the Korean Ministry. CKC represents a nonnative country's official ESL education curriculum.

### 2.5 OneStopEnglish (OSE) - indep. & paired

209OSE is another well-regarded dataset in RA. It aims210at ESL (English as Second Language) learners and211consists of three paraphrased versions of an article212from The Guardian Newspaper. Along with the213original OSE dataset, we created a paired version214(OSE-Pair). This variation has 189 items and each215item has advanced-intermediate-elementary pairs.

In addition, OSE-Sent is a sentence-paired version of OSE. The dataset consists of three parts: adv-ele (1674 pairs), adv-int (2166), int-ele (2154). 216

217

218

219

220

221

222

224

225

228

229

230

232

233

234

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

256

257

#### 2.6 Newsela (NSL) - indep. & paired

NSL (Xu et al., 2015) is a dataset particularly developed for text simplification studies. The dataset consists of 1,130 articles, with each item re-written 4 times for children at different grade levels. We create a paired version (NSL-Pair) (2125 pairs).

#### 2.7 ASSET - paired

ASSET (Alva-Manchego et al., 2020) is a paired sentence dataset. The dataset consists of 360 sentences, with each item simplified 10 times.

#### **3** Recalibration for Modern Usage

### 3.1 Choosing Traditional Read. Formulas

We start by recalibrating five readability formulas. We considered Zhou et al. (2017) and section 1.3 to sort out the most popular traditional readability formulas. Further, to make a fair comparison with our adjusted variations, we choose the formulas originally intended to output U.S. school grades.

Flesh-Kincaid Grade Level (FKGL) is primarily developed for U.S. Navy personnel. The readability level of 18 passages, from Navy technical training manuals, were calculated based on the criterion that 50% of subjects with a reading ability at the specific level had to score  $\geq 35\%$  on a cloze test. Responses from 531 Navy personnel were used.

$$FKGL = a \cdot \frac{\#word}{\#sent} + b \cdot \frac{\#syllable}{\#word} + c$$

where sent is sentence, and # refers to "count of."

The genius of Gunning Fog Index (FOGI) is the idea that word difficulty highly correlates with the number of syllables. Such a conclusion was deduced upon the inspection of Dale's list of easy words (Zhou et al., 2017; Dale and Chall, 1948). However, the shortcoming of FOGI is the overgeneralization that "all" words with more than two syllables are difficult. Indeed, "banana" is quite an easy word.

$$FOGI = a \cdot \left(\frac{\#word}{\#sent} + b \cdot \frac{\#difficult \ word}{\#word}\right) + c$$

Simple Measure of Gobbledygook (SMOG) Index, known for its simplicity, resembles FOGI in that both use the number of syllables to classify a word's difficulty. But SMOG sets its criterion a

262

263

265

269

270

271

272

273

274

275

276

277

278

281

282

283

286

287

291

296

297

301

little high to more than three syllables per word. Additionally, SMOG incorporates a square root approach instead of a linear regression model.

$$SMOG = a \cdot \sqrt{b \cdot \frac{\text{\#polysyllable word}}{\text{\#sent}}} + c$$

Coleman-Liau Index (COLE) is more of a lesserused variation among the five. But we could still find multiple studies outside computational linguistics that still partly depend on COLE (Kue et al., 2021; Szmuda et al., 2020; Joseph et al., 2020; Powell et al., 2020). The novelty of COLE is that it calculates readability without counting syllables, which was viewed as a time-consuming approach.

$$COLE = a \cdot 100 \cdot \frac{\text{#letter}}{\text{#word}} + b \cdot 100 \cdot \frac{\text{#sent}}{\text{#word}} + c$$

Automated Readability Index (AUTO) is developed for U.S. Air Force to handle more technical documents than textbooks. Like COLE, AUTO relies on the number of letters per word, instead of the more commonly-used syllables per word. Another quirk is that non-integer scores are all rounded up.

$$AUTO = a \cdot \frac{\text{#letter}}{\text{#word}} + b \cdot \frac{\text{#word}}{\text{#sent}} + c$$

#### 3.2 Recalibration & Performance

3.2.1 Traditional Formulas, Other Text Types

We only recalibrate formulas on the CCB dataset. As stated in section 2.1, we limit to CCB's storytype items. In a preliminary investigation, we obtained low r2 scores (< 0.3, before & after recalibration) between the traditional readability formulas and poetry, informational text, and drama.

### 3.2.2 Details on Recalibration

We started from a large research software, LingFeat (Lee et al., 2021), and expanded it to include more necessary features. Then, we extracted the surface-level features in traditional readability formulas (i.e.  $\frac{\text{#letter}}{\text{#word}}, \frac{\text{#syllable}}{\text{#synd}}$ ) and put them in a dataframe.

CCB has 6 readability classes but they are in the forms of range: K1, K2-3, K4-5, K6-8, K9-10, K11, and CCR (college and above). Since our objective is not the most accurate modeling of the CCB dataset, we sacrifice considerable accuracy by simply estimating items to K1, K2.5, K4.5, K7, K9.5, or K12. We model the general trend of CCB.

Using the class estimations as true labels and the created dataframe as features, we ran an optimization function to calculate the best coefficients (a, b,

| a) Coef.s  | FKGL     | FOGI    | SMOG   | COLE    | AUTO    |  |
|------------|----------|---------|--------|---------|---------|--|
| original-a | 0.390    | 0.4000  | 1.043  | 0.05880 | 4.710   |  |
| adjusted-a | 0.1014   | 0.1229  | 2.694  | 0.03993 | 6.000   |  |
| original-b | 11.80    | 100.0   | 30.00  | -0.2960 | 0.5000  |  |
| adjusted-b | 20.89    | 415.7   | 8.815  | -0.4976 | 0.1035  |  |
| original-c | -15.59   | 0.0000  | 3.129  | -15.80  | -21.43  |  |
| adjusted-c | -21.94   | 1.866   | 3.367  | -5.747  | -19.61  |  |
| b) Perf.   | FKGL     | FOGI    | SMOG   | COLE    | AUTO    |  |
| r2 score   | -0.03835 | -0.3905 | 0.1613 | 0.4341  | -0.5283 |  |
| r2 score   | 0.4423   | 0.4072  | 0.3192 | 0.4830  | 0.4263  |  |
| Pearson r  | 0.5698   | 0.5757  | 0.5649 | 0.6800  | 0.5684  |  |
| Pearson r  | 0.6651   | 0.6381  | 0.5649 | 0.6949  | 0.6529  |  |

Table 2: a) Original & adjusted coefficients. b) Performance on CCB. Measured on U.S. Standard Curriculum's K-\* Output. Bold refer to our new, adjusted versions.

c). The calculation was done by using non-linear least squares in fitting functions (Virtanen et al., 2020). Additional details available in appendix B.

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

329

331

333

334

335

337

#### 3.2.3 Coefficients & Performances

Table 2-a shows the original coefficients and the adjusted variations, rounded up to match significant figures. The adjusted traditional readability formulas can be obtained by simply plugging in these values to the formulas in section 3.1.

#### 4 The Common Readability Formula

#### 4.1 Criteria

Considering the value of traditional readability formulas as essentially the generalized definition of readability for the non-experts (section 1), what really matters is the included features. The coefficients (or weights) can be recalibrated anytime to fit a specific use. Therefore, it is important to first identify handcrafted linguistic features that universally affect readability. Additionally, to ensure breadth and usability, we set the following guides: **1.** We avoid surface-level features that lack linguistic value (Feng et al., 2010). They include  $\frac{\#letter}{\#word}$ . **2.** We include at most one linguistic feature from each linguistic subgroup. We use the classifications from Lee et al. (2021); Collins-Thompson (2014). **3.** We stick to a simplistic linear equation format.

#### 4.2 Feature Extraction & Ranking

We utilize LingFeat for feature extraction. It is a public software that supports 255 handcrafted linguistic features in the branches of advanced semantic, discourse, syntactic, lexico-semantic, and shallow traditional. They further classify into 14 subgroups. We study the linguistically-meaningful branches: discourse (entity density, entity grid),

|       |                        | Feature       |                                  | ССВ      | WBT       | CAM      | СКС      | OSE      |
|-------|------------------------|---------------|----------------------------------|----------|-----------|----------|----------|----------|
| Score | Branch Subgroup        | LingFeat Code | Brief Explanation                | r rk     | r rk      | r rk     | r rk     | r rk     |
| 43    | LxSem Psycholinguistic | as_AAKuL_C    | Kuperman Lemma AoA per Sent      | 0.540 25 | 0.505 1   | 0.722 42 | 0.711 4  | 0.601 25 |
| 43    | LxSem Psycholinguistic | as_AAKuW_C    | Kuperman Word AoA per Sent       | 0.537 28 | 0.503 2   | 0.722 43 | 0.711 6  | 0.602 24 |
| 40    | LxSem Psycholinguistic | at_AAKuW_C    | Kuperman Word AoA per Word       | 0.703 5  | 0.308 36  | 0.784 20 | 0.643 21 | 0.455 66 |
| 40    | Synta Tree Structure   | as_TreeH_C    | Tree Height per Sent             | 0.550 21 | 0.341 30  | 0.686 51 | 0.699 9  | 0.541 44 |
| 40    | Synta Part-of-Speech   | as_ContW_C    | # Content Words per Sent         | 0.534 29 | 0.453 13  | 0.667 56 | 0.688 14 | 0.544 43 |
| 39    | LxSem Psycholinguistic | at_AAKuL_C    | Kuperman Lemma AoA per Word      | 0.723 4  | 0.323 35  | 0.785 19 | 0.650 20 | 0.453 67 |
| 39    | Synta Phrasal          | as_NoPhr_C    | # Noun Phrases per Sent          | 0.550 20 | 0.406 25  | 0.660 58 | 0.673 18 | 0.582 35 |
| 39    | Synta Phrasal          | to_PrPhr_C    | Total # Prepositional Phrases    | 0.470 47 | 0.189 58  | 0.808 11 | 0.580 36 | 0.729 3  |
| 39    | Synta Part-of-Speech   | as_FuncW_C    | # Function Words per Sent        | 0.468 48 | 0.471 8   | 0.662 57 | 0.673 17 | 0.614 19 |
| 38    | LxSem Psycholinguistic | to_AAKuL_C    | Total Sum Kuperman Lemma AoA     | 0.428 71 | 0.189 59  | 0.835 3  | 0.627 22 | 0.716 5  |
| 38    | LxSem Psycholinguistic | to_AAKuW_C    | Total Sum Kuperman Word AoA      | 0.427 72 | 0.189 60  | 0.835 4  | 0.625 23 | 0.715 6  |
| 36    | Synta Phrasal          | as_PrPhr_C    | # Prepositional Phrases per Sent | 0.513 35 | 0.417 23  | 0.607 70 | 0.608 28 | 0.590 34 |
| 36    | LxSem Word Familiarity | as_SbL1C_C    | SubtlexUS Lg10CD Value per Sent  | 0.467 49 | 0.430 20  | 0.612 69 | 0.699 10 | 0.533 45 |
| 35    | LxSem Type Token Ratio | OCOTTTR_S     | Corrected Type Token Ratio       | 0.745 1  | 0.006 228 | 0.846 1  | 0.445 65 | 0.692 7  |
| 35    | LxSem Word Familiarity | as_SbL1W_C    | SubtlexUS Lg10WF Value per Sent  | 0.462 52 | 0.437 19  | 0.605 71 | 0.693 12 | 0.523 48 |

Table 3: Top 15 (score  $\geq$  35) handcrafted linguistic features under Approach A. r: Pearson's correlation between the feature and the dataset. rk: the feature's correlation ranking on the specific dataset. Full version in appendix D.

|       | Feature |                  |               |                                  |       |    | WB     | Г   | CAN   | A  | СКС   | 2  | OSE   |    |
|-------|---------|------------------|---------------|----------------------------------|-------|----|--------|-----|-------|----|-------|----|-------|----|
| Score | Branch  | Subgroup         | LingFeat Code | Brief Explanation                | r     | rk | r      | rk  | r     | rk | r     | rk | r     | rk |
| 35    | LxSem   | Psycholinguistic | as_AAKuL_C    | Kuperman Lemma AoA per Sent      | 0.540 | 25 | 0.505  | 1   | 0.722 | 42 | 0.711 | 4  | 0.601 | 25 |
| 35    | LxSem   | Psycholinguistic | as_AAKuW_C    | Kuperman Word AoA per Sent       | 0.537 | 28 | 0.503  | 2   | 0.722 | 43 | 0.711 | 6  | 0.602 | 24 |
| 32    | LxSem   | Psycholinguistic | at_AAKuL_C    | Kuperman Lemma AoA per Word      | 0.723 | 2  | 0.323  | 35  | 0.785 | 42 | 0.650 | 22 | 0.453 | 67 |
| 32    | LxSem   | Psycholinguistic | at_AAKuW_C    | Kuperman Word AoA per Word       | 0.703 | 5  | 0.308  | 36  | 0.784 | 20 | 0.643 | 21 | 0.455 | 66 |
| 31    | Synta   | Phrasal          | as_NoPhr_C    | # Noun Phrases per Sent          | 0.550 | 20 | 0.406  | 25  | 0.660 | 58 | 0.673 | 18 | 0.582 | 35 |
| 31    | Synta   | Part-of-Speech   | as_ContW_C    | # Content Words per Sent         | 0.534 | 29 | 0.453  | 13  | 0.667 | 56 | 0.688 | 14 | 0.544 | 43 |
| 31    | Synta   | Phrasal          | as_PrPhr_C    | # Prepositional Phrases per Sent | 0.513 | 35 | 0.417  | 23  | 0.607 | 70 | 0.608 | 28 | 0.590 | 34 |
| 31    | Synta   | Part-of-Speech   | as_FuncW_C    | # Function Words per Sent        | 0.468 | 48 | 0.471  | 8   | 0.662 | 57 | 0.673 | 17 | 0.614 | 19 |
| 31    | LxSem   | Psycholinguistic | to_AAKuL_C    | Total Sum Kuperman Lemma AoA     | 0.428 | 71 | 0.189  | 59  | 0.835 | 3  | 0.627 | 22 | 0.716 | 5  |
| 31    | LxSem   | Psycholinguistic | to_AAKuW_C    | Total Sum Kuperman Word AoA      | 0.427 | 72 | 0.189  | 60  | 0.835 | 4  | 0.625 | 23 | 0.715 | 6  |
| 30    | LxSem   | Type Token Ratio | CorrTTR_S     | Corrected Type Token Ratio       | 0.745 | 1  | 0.006  | 228 | 0.846 | 1  | 0.445 | 65 | 0.692 | 7  |
| 30    | LxSem   | Variation Ratio  | CorrNoV_S     | Corrected Noun Variation-1       | 0.717 | 3  | 0.0858 | 131 | 0.842 | 2  | 0.406 | 78 | 0.612 | 21 |
| 30    | Synta   | Tree Structure   | as_TreeH_C    | Tree Height per Sent             | 0.550 | 21 | 0.341  | 30  | 0.686 | 51 | 0.699 | 9  | 0.541 | 44 |
| 30    | Synta   | Phrasal          | to_PrPhr_C    | Total # Prepositional Phrases    | 0.470 | 47 | 0.189  | 58  | 0.808 | 11 | 0.580 | 36 | 0.729 | 3  |
| 30    | LxSem   | Word Familiarity | as_SbL1C_C    | SubtlexUS Lg10CD Value per Sent  | 0.467 | 49 | 0.430  | 20  | 0.612 | 69 | 0.699 | 10 | 0.533 | 45 |

Table 4: Top 15 (score  $\geq$  30) handcrafted linguistic features under Approach B. Italic for the feature not in Table 3.

syntax (phrasal, tree structure, part-of-speech), and lexico-semantics (variation ratio, type token ratio, psycholinguistics, word familiarity).

339

340

341

343

345

348

351

352

354

355

After extracting the features from CCB, WBT, CAM, CKC, and OSE, we first create feature performance ranking by Pearson's correlation. We used Sci-Kit Learn (Pedregosa et al., 2011). We take extra measures (Approach A & B) to model the features' general performances across datasets. Each approach runs under differing premises:

**Premise A:** "Human experts' dataset creation and labeling are partially faulty. The weak performance of a feature in a dataset does not necessarily indicate its weak performance in other data settings".

**Premise B**: "All datasets are perfect. The weak performance of a feature in a dataset indicates the feature's weakness to be used universally."

After 78 hours of running, we decided not to extract features from NSL. Computing details are

in appendix E. Among the features included in LingFeat, there are traditional readability formulas, like FKGL and COLE. These formulas performed generally well but a single killer feature, like type token ratio (TTR), often outperformed formulas. Traditional readability formulas and shallow traditional features are excluded from the rankings.

#### 4.3 Approach A - Comparative Ranking

Under premise A, each dataset poses a different linguistic environment to feature performance. Further, premise A takes human error into consideration and agrees that data labeling is most likely inconsistent in some way. The literal correlation value itself is not too important under premise A.

Rather, we look for features that perform better than the others, under the same test settings. Thus, approach A's rewarding system is rank-dependent. In a dataset, features that rank 1-10 are rewarded 10 points, rank 11-20 get 9 points, ... and rank 91-100 357

426

427

428

429

430

431

432

433

434

435

436

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

(variation) only appeared under approach B, and  
we did not include it. Putting together gives: 445  
$$a = \sum A a A + b \sum familiarity$$

$$\mathbf{CoRF} = \frac{a \cdot \sum AoA + b \cdot \sum familiarity}{\text{#sentence}}$$
446

$$+ \frac{c \cdot \#CW + d \cdot \#NP + e \cdot \sum TH}{\#sentence}$$
447

$$+ f \cdot \frac{\text{#unique word}}{\sqrt{\text{#word}}} + g$$
 448

where  $\sum$  is total sum, AoA is age-of-acquisition (Kuperman), familiarity is the measure of word frequency (SubtlexUS), CW is content word, NP is noun phrase and TH is tree height.

direct classification of word difficulties always out-

performed vague approaches like SubtlexUS word

below. We hoped to ensure linguistic breadth and

1. From top to bottom, go through ranking (table

3 & 4) to sort out the features that performed the

2. Conduct step 1 to both datasets and compare the

results to each other. Though this process, we only

leave the features that duplicate in both rankings.

both approach A and B. The final selected features

are as\_AAKuL\_C (psycholinguistic), as\_TreeH\_C

(tree structure), as\_ContW\_C (part-of-speech),

as\_NoPhr\_C (phrasal), as\_SbL1C\_C (word famil-

iarity), CorrTTR\_S (type token ratio). CorrNov\_S

The steps above produce the same results for

Back to feature selection, we follow the steps

frequency (Brysbaert and New, 2009).

meaningful correlation performance.

best in each linguistic subgroup.

we did not include it.

#### 4.6 More on CoRF & Calibration

The final CoRF (section 4.5) is brought in three parts. The first is lexico-semantics, which measures word-related difficulty. It adds the total sum of each word's age-of-acquisition (Kuperman's) and the sum of word familiarity scores (Lg10CD in SubtlexUS). The sum is divided by # sentences.

The second is syntax, which deals with how each sentence is structured. We look at the number of content words, noun phrases, and the total sum of sentence tree height. Here, content words (CW) are words that possess semantic content and contribute to the meaning of the specific sentence. Following LingFeat, we consider a word to be a content word if it has "NOUN", "VERB", "NUM", "ADJ", "ADV" as a POS tag. Also, a sentence's tree height (TH) is calculated from a constituency-parsed tree, which we used the CRF parser (Zhang et al., 2020)

get 1 point. Since there are five feature correlation 376 rankings (one per dataset), the maximum score is 377 50. The results are in Table 3, in the order of score.

### 4.4 Approach B - Absolute Correlation

381

391

393

396

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

Under premise B, the weak correlation of a feature in a dataset is solely due to the feature's weakness to generalize. This is because all datasets are supposedly perfect. Hence, we only measure the feature's absolute correlation across datasets.

Approach B's rewarding system is correlationdependent. In a dataset, features that show correlation value between 0.9-10 are rewarded 10 points, value between 0.8-0.89 get 9 points, ... and value between 0.0-0.09 get 1 point. Like approach A, the maximum score is 50. The result is in Table 4.

#### 4.5 Analysis & Manual Feature Selection

First and the most noticeable, the top features under premise A & B are similar. In fact, the two results are almost replications of each other except for minor changes in order. We initially set two premises to introduce differing views (and hence the results) to feature rankings. Then, we would choose the features that perform well in both.

There seems to be an inseparable correlation between ranking-based (premise A) and correlationbased (premise B) approaches but it was unexpected for the two results to be so similar. CorrNoV\_S (Corrected Noun Variation) was the only new top feature introduced under premise B.

Second, discourse-based features (mostly entityrelated) performed poorly for use in our final CoRF. As an exception, ra\_NNToT\_C (noun-noun transitions : total) scored 28 under premise A and 26 under premise B. On the other hand, a majority of lexico-semantic and syntactic features performed well throughout. This strongly suggests that a possible discovery of universally-effective features for readability is in lexico-semantics or syntax.

Third, the difficulty of a document heavily depended on the difficulty of individual words. In detail, as\_AAKuL\_C, as\_AAKuW\_C, to\_AAKuL\_C, to AAKuW C showed consistently high correlations across the five datasets. As shown in Section 2, these five datasets have different authors, target audience, average length, labeling techniques, and the number of classes. Each dataset had at least one of these features among the top 5 performances.

The four features come from age-of-acquisition research by Kuperman et al. (2012), which now prove to be a very important resource for RA. Such

| а       | b       | c d    |        | e      | f      | g      |
|---------|---------|--------|--------|--------|--------|--------|
| 0.04876 | -0.1145 | 0.3091 | 0.1866 | 0.2645 | 1.1017 | -4.125 |

Table 5: Coefficients for CoRF, calibrated on CCB (foroutput based on U.S. Standard School Curriculum)

to obtain. The related algorithms from NLTK (Bird et al., 2009) were used in calculating tree height.The same CRF parser was also used to count the number of noun phrase (NP) occurrences.

The third is the type token ratio (TTR). This part is separated as it is the only section of CoRF that is averaged on the word count. TTR measures how many unique vocabularies appear with respect to the total word count. TTR is often used as a measure of lexical richness (Malvern and Richards, 2012) and ranked the best performance on two native datasets (CCB and CAM). Importantly, these two datasets represent US and UK school curriculums, and TTR seems a good evaluator. What was interesting is that out of the five TTR variations from Lee et al. (2021); Vajjala and Meurers (2012), corrected TTR generalized particularly well.

Like section 3, we use the non-linear least fitting method on CCB to calibrate CoRF. The results in Table 5 match what we expected. Coefficient b for word familiarity, which measures how frequently the word is used in American English, is negative since more common words often have faster lexical comprehension times (Brysbaert et al., 2011).

### 5 Evaluation

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

503

506

507

509

510

511

512

513

514

515

Table 6 gives a performance comparison of CoRF against other traditional readability formulas and human performances. The human predictions were made by the U.S. Bachelor's degree holders living in the U.S. Ten human predictions were averaged to obtain the final prediction for each item, for comparison against the original CCB classifications.

The calibrated formulas show a particularly great increase in r2 score. This likely means that the new recalibrated formulas can capture the variance of the original CCB classifications much better when compared to the original formulas. We believe that such an improvement stems from the change in datasets. The original formulas are mostly built on human tests of 20th century's military or technical documents, whereas the recalibration dataset (CCB) are from the student-targeted school curriculum. Further, CCB is classified by trained professionals. Hence, the standards for readability can differ. The new recalibrated versions are more suit-

| Metric    | Human   | CoRF   | FKGL    | FOGI   | SMOG   | COLE   | AUTO    |
|-----------|---------|--------|---------|--------|--------|--------|---------|
| MAE       | N.A.    | N.A.   | 2.844   | 3.413  | 3.114  | 2.537  | 3.377   |
| MAE       | 3.509   | 2.154  | 2.457   | 2.516  | 2.728  | 2.378  | 2.514   |
| r2 score  | N.A.    | N.A.   | -0.0381 | -0.389 | 0.1613 | 0.4341 | -0.5283 |
| r2 score  | -0.0312 | 0.5536 | 0.4720  | 0.4445 | 0.3192 | 0.4830 | 0.4263  |
| Pearson r | N.A.    | N.A.   | 0.5701  | 0.5762 | 0.5649 | 0.6800 | 0.5684  |
| Pearson r | 0.0838  | 0.7440 | 0.6667  | 0.6400 | 0.5650 | 0.6949 | 0.6530  |

Table 6: Scores on CCB. Measured on U.S. Standard Curriculum's K-\* Output. Bold for new or adjusted.

able for analyzing the modern general documents and giving K-\* output by modernized standards.

MAE (Mean Absolute Error), r2 score, and Pearson's r improve once more with CoRF. Even though the same dataset, same fitting function, and same evaluation techniques (no split, all train) were used, the critical difference was in the features. The shallow surface-level features from the traditional readability formulas also showed top rankings across all datasets but lacked linguistic coverage. Hence, CoRF could capture more textual properties that led to a difference in readability.

Lastly, we observe that it is highly difficult for the general human population to exactly guess the readability of a text. Out of 690 predictions, only 286 were correct. We carefully posit that this is because: 1. the concept of "readability" is vague and 2. everyone goes through varying education. It could be easier to choose which item is more readable, instead of guessing how readable an item is. Given the general population, it is always better to use some quantified models than trust human.

#### 6 Investigations on Application

#### 6.1 Text Simplification - Passage-based

All readability formulas, whether recalibrated or not, show near-perfect performances in ranking the simplicity of texts. On both OSE-Pair & NSL-Pair, we designed a simple task of ranking the simplicity of an item. Both paired datasets include multiple simplified versions of an original item. Each row consists of various simplifications. A correct prediction is the corresponding readability formula output matching simplification level (e.g. original: highest prediction, ..., simplest: lowest prediction).

In OSE-Pair, a correct prediction must properly rank three simplified items. CoRF showed a meaningfully improved performance than the other five traditional readability formulas before recalibration. CoRF correctly classified 98.7% pairs, while the others stayed  $\leq$ 95% (FKGL: 93.4%, FOGI: 92.6%, SMOG: 94.4%, COLE: 94.9%, AUTO: 92.6%).

552

553

554

555

| a) Adv-Ele           | CoRF | FKGL | FOGI | SMOG           | COLE | AUTO |
|----------------------|------|------|------|----------------|------|------|
| Accuracy<br>Accuracy |      |      |      | 11.4%<br>11.4% |      |      |
| b) Adv-Int           | CoRF | FKGL | FOGI | SMOG           | COLE | AUTO |
| Accuracy<br>Accuracy |      |      |      | 12.2%<br>12.2% |      |      |
| c) Int-Ele           | CoRF | FKGL | FOGI | SMOG           | COLE | AUTO |
| Accuracy<br>Accuracy |      |      |      | 9.02%<br>9.02% |      |      |

Table 7: Scores on OSE-Sent. Bold for new or adjusted.

Recalibration generally helped the traditional readability formulas but CoRF still showed better performance (FKGL: 97.8%, FOGI: 97.1%, SMOG: 94.4%, COLE: 89.9%, AUTO: 95.8%).

In NSL-Pair, a correct prediction must properly rank five simplified items, which is a more difficult task than the previous. Nonetheless, all six formulas achieved 100% accuracies. The same results were achieved before and after CCB-recalibration. This hints that NSL-Pair is thoroughly simplified.

Readability formulas seem to perform well in ranking several simplifications on a passage-level. But there certainly are limits. First, one must understand that calculating "how much simple" is a much difficult task (Table 6). It is completely different from simply choosing a difficult one out of two or more. Second, the good results could be because sufficient simplification was done. If more there were more simplifications available, with lesser differences from simplification to simplification, simplistic readability formulas could not be enough.

#### 6.2 Text Simplification - Sentence-based

We were surprised that some existing text simplification studies are directly using traditional readability formulas for sentence difficulty evaluation. Our results show that using a formula-based approach is particularly useless in evaluating a sentence.

We tested both CCB-recalibrated and original formulas on ASSET. Here, a correct prediction must properly rank eleven simplified items. Despite the task difficulty, we anticipated seeing some correct predictions as there were 360 pairs. SMOG guessed 37 (after recalibration) and 89 (before recalibration) correct out of 360. But all the other formulas failed to make any correct prediction.

OSE-Sent poses an easier task. Since the dataset is divided into adv-int, adv-ele, and int-ele, the readability formulas now had to guess which is more difficult, out of the given two. We do obtain

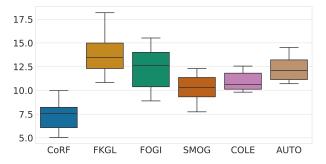


Figure 1: On medical texts. CoRF, against five others.

some positive results, showing that readability formulas can be useful in the cases where only two sentences are compared. On ranking two sentences, CoRF performs better by a large margin. 596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

Further research must be done as to how readability formulas can be applied to sentences. The results are inconsistent, especially with SMOG on ASSET and OSE-Sent. Also, recalibration seems to negatively affect sentence simplification ranking.

#### 6.3 Medical Documents

We argue that CoRF is effective in fixing the overinflated prediction of difficulty on medical texts. Such sudden inflation is widely-reported (Zheng and Yu, 2017) as the common weaknesses of traditional readability formulas on medical documents.

The U.S. National Institute of Health (NIH) guides that patient documents be  $\leq$ K-6 of difficulty. The most distinct characteristic of medical documents is the use of lengthy medical terms, like otolaryngology, urogynecology, and rheumatology. This makes traditional formulas, based on syllables, unreliable. But CoRF uses familiarity and age-of-acquisition to penalty and reward word difficulty.

A medical term not found in Kuperman's and SubtlexUS will have no effect. Instead, it will simply be labeled a content word. But in traditional formulas, the repetitive use of medical terms (which is likely the case) results in an insensible aggregation of text difficulty. In case various medical terms appear, CoRF rewards each as a unique word.

Among recent studies is Haller et al. (2019), which analyzed the readability of urogynecology patient education documents in FKGL, SMOG, and Fry Readability. We also analyze the same 18 documents from the American Urogynecologic Society (AUGS) by manual OCR-based scraping. As Figure 1 shows, it is evident that CoRF helps regulate the traditional readability formulas' tendencies to over-inflate on medical texts. An example of the collected resource is given in appendix B.

593

#### References

636

641

643

648

651

653

654 655

657

666

671

673

676

677

679

680

684

- Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.
- Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. arXiv preprint arXiv:2005.00481.
- Matthew Bange, Eric Huh, Sherwin A Novin, Ferdinand K Hui, and Paul H Yi. 2019. Readability of patient education materials from radiologyinfo. org: has there been progress over the past 5 years? *American Journal of Roentgenology*, 213(4):875–879.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".
- Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental psychology*.
- Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.
- Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.
- Murphy Odo Dennis. 2018. A comparison of readability and understandability in second language acquisition textbooks for pre-service eff teachers. *Journal of Asia TEFL*, 15(3):750–765.
- William H DuBay. 2004. The principles of readability. *Online Submission*.
- Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment.
- Laura Gaeta, Edward Garcia, and Valeria Gonzalez. 2021. Readability and suitability of spanish-language hearing aid user guides. *American Journal of Audiology*, 30(2):452–457.

Robert Gunning et al. 1952. Technique of clear writing.

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

724

725

726

728

729

730

731

732

733

734

735

737

739

- Jasmine Haller, Zachary Keller, Susan Barr, Kristie Hadden, and Sallie S Oliphant. 2019. Assessing readability: are urogynecologic patient education materials at an appropriate reading level? *Female pelvic medicine* & reconstructive surgery, 25(2):139–144.
- David R Hansberry, Michael D'Angelo, Michael D White, Arpan V Prabhu, Mougnyan Cox, Nitin Agarwal, and Sandeep Deshmukh. 2018. Quantitative analysis of the level of readability of online emergency radiology-based patient education resources. *Emergency radiology*, 25(2):147–152.
- Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Caroline Jarrett and Janice 'Ginny' Redish. 2019. Readability formulas: 7 reasons to avoid them and what to do instead.
- Zhiwei Jiang, Gang Sun, Qing Gu, Tao Bai, and Daoxu Chen. 2015. A graph-based readability assessment method using word coupling. In *Proceedings of the* 2015 Conference on Empirical Methods in Natural Language Processing, pages 411–420.
- Pradeep Joseph, Nicole A Silva, Anil Nanda, and Gaurav Gupta. 2020. Evaluating the readability of online patient education materials for trigeminal neuralgia. *World Neurosurgery*, 144:e934–e938.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
- Elizabeth Kiwanuka, Raman Mehrzad, Adnan Prsic, and Daniel Kwan. 2017. Online patient resources for gender affirmation surgery: an analysis of readability. *Annals of plastic surgery*, 79(4):329–333.
- George R Klare. 2000. The measurement of readability: useful information for communicators. *ACM Journal* of Computer Documentation (JCD), 24(3):107–121.
- Jennifer Kue, Dori L Klemanski, and Kristine K Browning. 2021. Evaluating readability scores of treatment summaries and cancer survivorship care plans. *JCO Oncology Practice*, pages OP–20.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

844

845

846

847

848

849

850

851

852

853

798

Bruce W Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669– 10686.

741

742

743 744

745

747

748

749

750

751

754

755

756

757

758

770

771

772

773

778

779

782

784

785

786

790

793

- Bruce W. Lee and Jason Lee. 2020a. LXPER index 2.0: Improving text readability assessment model for L2 English students in Korea. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 20– 24, Suzhou, China. Association for Computational Linguistics.
- Bruce W. Lee and Jason Hyung-Jong Lee. 2020b. Lxper index: A curriculum-specific text readability assessment model for eff students in korea. *International Journal of Advanced Computer Science and Applications*, 11(8).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7):389–398.
- David Malvern and Brian Richards. 2012. Measures of lexical richness. *The encyclopedia of applied linguistics*.
- G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.
- Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. *Advances in Information Retrieval*, 12035:33.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Lauren E Powell, Emily S Andersen, and Andrea L Pozez. 2020. Assessing readability of patient education materials on breast reconstruction by major us academic institutions. *Plastic and Reconstructive Surgery–Global Open*, 8(9S):127–128.

- Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.
- H. Andrew Schwartz, Masoud Rouhizadeh, Michael Bishop, Philip Tetlock, Barbara Mellers, and Lyle Ungar. 2017. Assessing objective recommendation quality through political forecasting. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2348–2357, Copenhagen, Denmark. Association for Computational Linguistics.
- Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389.
- Edgar A Smith and RJ Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.
- T Szmuda, C Özdemir, S Ali, A Singh, MT Syed, and P Słoniewski. 2020. Readability of online patient education material for the novel coronavirus disease (covid-19): a cross-sectional health literacy study. *Public Health*, 185:21–25.
- Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8384–8395, Online. Association for Computational Linguistics.
- Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.
- Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.
- Courtney van Ballegooie and Peter Hoang. 2021. Assessment of the readability of online patient education material from major geriatric associations. *Journal of the American Geriatrics Society*, 69(4):1051–1056.
- N Verhelst, Piet Van Avermaet, S Takala, N Figueras, and B North. 2001. *Common European Framework* of *Reference for Languages: learning, teaching, assessment*. Cambridge University Press.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay

Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

855

865

870

871

874

875

876

877

878

879

882

894

900

901

902

903

904

905

906

908

909

- Orion Weller, Jordan Hildebrandt, Ilya Reznik, Christopher Challis, E Shannon Tass, Quinn Snell, and Kevin Seppi. 2020. You don't have time to read this: An exploration of document reading time prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1789– 1794.
- Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.
- Danny TY Wu, David A Hanauer, Qiaozhu Mei, Patricia M Clark, Lawrence C An, Jianbo Lei, Joshua Proulx, Qing Zeng-Treitler, and Kai Zheng. 2013.
  Applying multiple methods to assess the readability of a large corpus of medical documents. *Studies in health technology and informatics*, 192:647.
- Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.
- Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural CRF constituency parsing. In *Proceedings of IJCAI*, pages 4046–4053.
- Jiaping Zheng and Hong Yu. 2017. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. *Journal of medical Internet research*, 19(3):e59.
- Shixiang Zhou, Heejin Jeong, and Paul A Green. 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 60(1):97–111.

# A Public Resources We Developed

#### A.1 <Anonymous> v.1.1.1

For the fast dissemination of our research results, we develop <Anonymous>. It is a python-based library that supports many useful features in analyzing a text. We distribute software under Attribution-NonCommercial-ShareAlike 4.0 International. 910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

957

### A.1.1 As a Readability Tool

<Anonymous> supports six readability formulas: CoRF, FKGL, FOGI, SMOG, COLE and AUTO. All formulas, other than CoRF, are also available in recalibrated variations. A particularly useful feature of this library is that all formulas are fitted to give the U.S. standard school grading system as output. Compared to some other traditional readability formulas where a user has to refer to a table understand output, K-\* based numbers are intuitive.

### A.1.2 As a General Tool

We have plans to expand <Anonymous> to support various menial tasks in text analysis. We are to focus on the tasks that can be better performed using simplistic approaches. One feature that we had already implemented is text reading time estimation. Weller et al. (2020) has previously shown in a large-scale study that a commonly used rule-ofthumb for online reading estimates, 240 words per minute (WPM), shows better RMSE and MAE results when compared to more modern approaches using XLNet (Yang et al., 2019), ELMo (Peters et al., 2018) and RoBERTa (Liu et al., 2019). We implement 175, 240 and 300 WPM. This fits more users with slow, average, or fast reading speed.

#### A.1.3 Basic Usage

For straightforward maintenance, we keep <Anonymous>'s architecture as simple as possible. There are not many steps for the user to take:

*slightly* faster to directly call in the full forms as:

import <Anonymous>

new\_object = <Anonymous>.request("string") 947 readability1 = new\_object.CoRF() 948 readability2 = new\_object.FKGL() 949 *readability3* = *new\_object.FOGI()* 950 readability4 = new object.SMOG()951 *readability5 = new\_object.COLE()* 952 readability6 = new object.AUTO()953 *time\_to\_read = new\_object.RT()* 954 CoRF(), FKGL(), FOGI(), SMOG(), COLE(), 955 AUTO(), RT() are shortcut functions. It can be 956

1022

1023

1024

1025

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

| 958 | common_readability_formula()             |
|-----|--|
| 959 | flesch_kincaid_grade_level()             |
| 960 | fog_index()                              |
| 961 | smog_index()                             |
| 962 | <pre>coleman_liau_index()</pre>          |
| 963 | <pre>automated_readability_index()</pre> |
| 964 | read_time()                              |

965

967

968

969

970

971

972

993

997

1001

1002

1003

1006

Further, all readability formula functions (except for CoRF) has option to choose the original or the adjusted variation. Default is set *adjusted* = *True*.

#### A.1.4 <Anonymous> Speed to Calculation

We care for the library's calculation speed so that it can be of practical use for research implementations. We chose the following items for evaluation. **ITEM A** 

In those times panics were common, and few days 973 passed without some city or other registering in its 974 archives an event of this kind. There were nobles, 975 who made war against each other; there was the 976 king, who made war against the cardinal; there was 977 Spain, which made war against the king. Then, in 978 addition to these concealed or public, secret or open 979 wars, there were robbers, mendicants, Huguenots, wolves, and scoundrels, who made war upon ev-981 982 erybody. The citizens always took up arms readily against thieves, wolves or scoundrels, often against nobles or Huguenots, sometimes against the king, 984 but never against the cardinal or Spain. It resulted, then, from this habit that on the said first Monday 987 of April, 1625, the citizens, on hearing the clamor, and seeing neither the red-and-yellow standard nor the livery of the Duc de Richelieu, rushed toward the hostel of the Jolly Miller. When arrived there, the cause of the hubbub was apparent to all.

The Three Musketeers, Alexandre Dumas 992 **ITEM B** 

The vaccine contains lipids (fats), salts, sugars and buffers. COVID-19 vaccines do not contain eggs, gelatin (pork), gluten, latex, preservatives, antibiotics, adjuvants or aluminum. The vaccines are safe, even if you have food, drug, or environmental allergies. Talk to a health care provider first before getting a vaccine if you have allergies to the following vaccine ingredients: polyethylene glycol (PEG), polysorbate 80 and/or tromethamine (trometamol or Tris).

COVID-19 Vaccine Information Sheet, Ministry of 1004 1005 Health, Ontario Canada

# **ITEM C**

BERT alleviates the previously mentioned unidirec-1007 tionality constraint by using a "masked language 1008

model"(MLM) pre-training objective, inspired by the Cloze task.

Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

| a) ITEM A | CoRF   | FKGL   | FOGI   | SMOG   | COLE   | AUTO   |
|-----------|--------|--------|--------|--------|--------|--------|
| item * 1  | 0.6371 | 0.0002 | 0.0001 | 0.0001 | 0.0000 | 0.0000 |
| item * 5  | 2.6450 | 0.0006 | 0.0005 | 0.0004 | 0.0001 | 0.0001 |
| item * 10 | 5.5175 | 0.0011 | 0.0010 | 0.0010 | 0.0004 | 0.0004 |
| item * 15 | 7.8088 | 0.0016 | 0.0016 | 0.0013 | 0.0003 | 0.0004 |
| item * 20 | 10.226 | 0.0021 | 0.0021 | 0.0018 | 0.0004 | 0.0004 |
| b) ITEM B | CoRF   | FKGL   | FOGI   | SMOG   | COLE   | AUTO   |
| item * 1  | 0.3531 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| item * 5  | 1.2842 | 0.0003 | 0.0003 | 0.0002 | 0.0000 | 0.0000 |
| item * 10 | 2.5178 | 0.0005 | 0.0005 | 0.0004 | 0.0001 | 0.0001 |
| item * 15 | 3.6545 | 0.0009 | 0.0007 | 0.0006 | 0.0002 | 0.0002 |
| item * 20 | 4.8308 | 0.0010 | 0.0010 | 0.0009 | 0.0002 | 0.0002 |
| c) ITEM C | CoRF   | FKGL   | FOGI   | SMOG   | COLE   | AUTO   |
| item * 1  | 0.1373 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| item * 5  | 0.1888 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| item * 10 | 0.2528 | 0.0002 | 0.0002 | 0.0002 | 0.0000 | 0.0000 |
| item * 15 | 0.3420 | 0.0003 | 0.0003 | 0.0002 | 0.0000 | 0.0000 |
| item * 20 | 0.3886 | 0.0004 | 0.0003 | 0.0003 | 0.0000 | 0.0000 |

Table 8: Speeds in seconds, on Items A, B and C.

First, it is very obvious that AUTO does a great 1014 job in keeping calculation speed short for longer 1015 texts as originally intended. Second, CoRF's calcu-1016 lation speed linearly increases in respect to the text 1017 length. Though, we believe that CoRF's speed is 1018 decent in its wide linguistic coverage, it seems true 1019 that the speed is weakness when compared to the 1020 other readability formulas. 1021

#### A.2 **Research Archive**

Our datasets, preprocessing codes and evaluation codes can be found in <Anonymous>. Copyrighted resources are given upon request to the first author.

#### В **External Resources**

#### Python Libraries **B.1**

# pandas v.1.3.4 (Wes McKinney, 2010)

Calculations for Kuperman's AoA CSV, SubtlexUS word familiarity CSV, manage and manipulate data. For feature study purposes, correlate and rank features in Tables 3 and 4. SuPar v.1.1.3 - CRF Parser Constituency parsing on input sentences -> calculate tree height and count noun phrases. spaCy v.3.2.0 (Honnibal and Johnson, 2015)

Sentence/dependency parsing on documents -> sent input into SuPar and count content words (POS).

| 1039 | Sci-Kit Learn v.1.0.1                                 |
|------|---|
| 1040 | Calculation, r2 score and MAE in Tables 2 and 6.      |
| 1041 | <b>SciPy</b> v.1.7.3                                  |
| 1042 | Calculation of Pearson's r for Tables 2 and 6. Fit-   |
| 1043 | ting function (scipy.optimize.curve_fit()) used to    |
| 1044 | recalibrate traditional readability formulas and give |
| 1045 | coefficients for CoRF in Tables 2 and 5.              |
| 1046 | <b>NLTK</b> v.3.6.5                                   |
| 1047 | Calculation of tree height for CoRF.                  |
| 1048 | LingFeat v.1.0.0-beta.19                              |
| 1049 | Extraction of handcrafted linguistic features.        |

#### B.2 Datasets

| New Class | ССВ               | WBT                |
|-----------|-------------------|--------------------|
| K1.0      | K1 (Age 6-7)      | N/A                |
| K2.0      | N/A               | Level 2 (Age 7-8)  |
| K2.5      | K2-3 (Age 7-9)    | N/A                |
| K3.0      | N/A               | Level 3 (Age 8-9)  |
| K4.0      | N/A               | Level 4 (Age 9-10) |
| K4.5      | K4-5 (Age 9-11)   | N/A                |
| K7.0      | K6-8 (Age 11-14)  | KS3 (Age 11-14)    |
| K9.5      | K9-10 (Age 14-16) | GCSE (Age 14-16)   |
| K12.0     | K11-CCR (Age 16+) | N/A                |

Table 9: Aged-based conversions for CCB and WBT.

We collected CCB by manually going through an official source<sup>2</sup>. WBT was obtained from the authors<sup>3</sup> in HTML format. We conducted basic preprocessing and manipulated WBT in CSV format. CAM was retrieved from an existing archive<sup>4</sup>. CKC was retrieved from a South Korean educational company<sup>5</sup>. OSE was retrieved from a public archive<sup>6</sup>. NSL was obtained from an American educational company<sup>7</sup>. AUGS medical texts (refer to Section 6.3) were manually scraped from the official website<sup>8</sup>. ASSET was obtained from a public repository<sup>9</sup>. Lastly, Table 9 shows how we converted WBT class labels to fit CCB and show in Table 1. All were consistent with intended use.

Further, to give more backgrounds to section 6.2, we give example pairs from ASSET and OSE-Sent. **ASSET** 

: Gable earned an Academy Award nomination for portraying Fletcher Christian in Mutiny on the Bounty.

<sup>2</sup>corestandards.org/assets/Appendix\_B.pdf

1: Gable also earned an Oscar nomination when he portrayed Fletcher Christian in 1935's Mutiny on the Bounty. : Gable won an Academy Award vote when he acted in 1935's Mutiny on the Bounty as Fletcher Christian.

: Gable also won an Academy Award nomination when he played Fletcher Christian in the 1935 film Mutiny on the Bounty.

: Gable was nominated for an Academy Award for portraying Fletcher Christian in 1935's Mutiny on the Bounty.

: Gable also earned an Academy Award nomination in 1935 for playing Fletcher Christian in "Mutiny on the Bounty.

: Gable also earned an Academy Award nomination when he played Fletcher Christian in 1935's Mutiny on the Bounty.

7: Gable recieved an Academy Award nomination for his role as Fletcher Christian. The film was Mutiny on the Bounty (1935).

: Gable earned an Academy Award nomination for his role as Fletcher Christian in the 1935 film Mutiny on the Bounty.

: Gable also got an Academy Award nomination when he played Fletcher Christian in 1935's movie, Mutiny on the Bounty.

: Gable also earned an Academy Award nomination when he portrayed Fletcher Christian in 1935's Mutiny on the Bounty.

#### **OSE-Sent** (ADV-ELE)

**ADV**: The Seattle-based company has applied for its brand to be a top-level domain name (currently .com), but the South American governments argue this would prevent the use of this internet address for environmental protection, the promotion of indigenous rights and other public interest uses.

**ELE**: Amazon has asked for its company name to be a top-level domain name (currently .com), but the South American governments say this would stop the use of this internet address for environmental protection, indigenous rights and other public interest uses.

### **OSE-Sent** (ADV-INT)

**ADV**: Brazils latest funk sensation, Anitta, has won millions of fans by taking the favela sound into the mainstream, but she is at the centre of a debate about skin colour.

**INT**: Brazils latest funk sensation, Anitta, has

<sup>&</sup>lt;sup>3</sup>Dr. Sowmya Vajjala, National Research Council, Canada <sup>4</sup>ilexir.co.uk/datasets/index.html

<sup>&</sup>lt;sup>5</sup>Bruce W. Lee, LAIR, LXPER Inc., South Korea

<sup>&</sup>lt;sup>6</sup>github.com/nishkalavallabhi/OneStopEnglishCorpus

<sup>&</sup>lt;sup>7</sup>Luke Orland, Newsela Inc., New York, U.S.A.

<sup>&</sup>lt;sup>8</sup>augs.org/patient-fact-sheets/

<sup>&</sup>lt;sup>9</sup>github.com/facebookresearch/asset

1172

1173

1174

1175

won millions of fans by making the favela sound 1122 popular, but she is at the centre of a debate about 1123 skin colour. 1124

### **OSE-Sent (INT-ELE)**

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136 1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1150

1151

1152

1153

1154

1156

1157

1158

1160

1161

1162

1163

1164

1165

1166

1167

1168

INT: Allowing private companies to register geographical names as gTLDs to strengthen their brand or to profit from the meaning of these names is not, in our view, in the public interest, the Brazilian Ministry of Science and Technology said.

ELE: Allowing private companies to register geographical names as gTLDs to profit from the meaning of these names is not, in our view, in the public interest, the Brazilian Ministry of Science and Technology said.

The following is an example of the AUGS medical documents used in Section 6.3 and Figure 1.

Interstitial Cystitis: Interstitial Cystitis/ Bladder Pain Syndrome Interstitial cystitis/bladder pain syndrome (IC/BPS) is a condition with symptoms including burning, pressure, and pain in the bladder along with urgency and frequency. About IC/BPS IC/BPS occurs in three to seven percent of women, and can affect men as well. Though usually diagnosed among women in their 40s, younger and older women have IC/BPS, too. It can feel like a constant bladder infection. Symptoms may be-1149 come severe (called a "flare") for hours, days or weeks, and then disappear. Or, they may linger at a very low level during other times. Individuals with IC/BPS may also have other health issues such as irritable bowel syndrome, fibromyalgia, chronic headaches, and vulvodynia. Depression 1155 and anxiety are also common among women with this condition. The cause of IC/BPS is unknown. It is likely due to a combination of factors. IC/BPS runs in families and so may have a genetic factor. 1159 On cystoscopy, the doctor may see damage to the wall of the bladder. This may allow toxins from the urine to seep into the delicate layers of the bladder lining, causing the pain of IC/BPS. Other research found that nerves in and around the bladder of people with IC/BPS are hypersensitive. This may also contribute to IC/BPS pain. There may also be an allergic component.

#### С **CCB Human Predictions**

In Section 2.1, we mention that human predictions 1169 were collected on Amazon Mechanical Turk. Then, 1170 we compared human performance to readability 1171

formulas in Table 6. Here, surveys are designed. Description: must choose which difficulty level does the text belong, "difficulty does not correlate with text length"

Qualification Requirement(s): Location is one of US, HIT Approval Rate (%) for all Requesters' HITs greater than 80, Number of HITs Approved greater than 50, US Bachelor's Degree equal to true, Masters has been granted

All 69 story-type items from CCB were given. Each item had to be completed by at least 10 different individuals, resulting in 690 responses in total. They were given 6 representative examples. Payments were adequately and they were informed that the responds shall be used for research.

#### D Handcrafted Linguistic Features and the Respective Generalizability

We give full generalizability rankings that we obtained through LingFeat. Considering that much work has to be done on the generalizability of RA, we believe that these rankings are particularly helpful. Table 10, Table 11, Table 12, Table 13, Table 14 and Table 15, Table 16 are expanded versions of Table 3 and Table 4. The features not shown in Table 10, Table 11 and Table 12 scored a 0.

From the full rankings, it is clear that shallow traditional (surface-level), lexico-semantic and syntactic features are effective throughout all datasets. Advanced semantics and discourse features show some what similar mid-low performances. However, it should be acknowledged that among the worst performing are lexico-semantic and syntactic features, too. This is perhaps because LingFeat itself has a very lexico-semantics and syntax-focused collection of handcrafted linguistic features. Thus, more study is needed.

Even if two features are from the same group (phrasal), they could show drastically varying performances (# Noun phrases per Sent - scored 39 in approach A v.s. # Verb phrases per Sent - scored 1 in approach A). Hence, thorough feature study must always be conducted during research. In a feature selection for a readability-related model, a cherry picking the most well performing feature from each feature group is recommended.

#### Е **Computing Power**

Single CPU chip. Architecture: x86\_64; CPU(s): 1218 16; Model name: Intel(R) Core(TM) i9-9900KF 1219 CPU @ 3.60GHz; CPU MHz: 800.024 1220

| Sourd Bandh Subgroup         LungFrat Code         Band Fughamizon         r         f.         f   |       |        |                  | Fe            | eature                           | ССВ       | WBT       | CAM      | СКС      | OSE       |
|---|-------|--------|------------------|---------------|----------------------------------|-----------|-----------|----------|----------|-----------|
| 43       Lasern Psycholingistis m_AAKuL_C lemma AoA of lemma per Sent       0.42       20       0.72       40       0.07       10       10       0.07       10       <  | Score | Branch | Subgroup         | LingFeat Code | Brief Explanation                | r rk      | r rk      | r rk     | r rk     | r rk      |
| 13       Same Ny-Line Name       a, Clara, C. # duranesse, present       0.39       21       0.897       21       0.907       24       0.917       1       60.02       24         13       Losen Pycholingistie a, AKRW, C. Aod erwales per Nent       0.357       7       0.847       2       0.912       1       6.002       24         14       Losen Pycholingistie a, AKRW, C. Aod erwales per Nent       0.345       5       0.318       60.075       20       0.847       20       0.914       20       0.841       20       0.841       20       0.841       20       0.841       20       0.841       20       0.841       20       0.841       20       0.841       20 </td <td></td> <td></td> <td></td> <td></td> <td>· ·</td> <td></td> <td></td> <td></td> <td></td> <td></td>   |       |        |                  |               | · ·                              |           |           |          |          |           |
| 41       LSAM       V.A.M.W.C. And V.S.M.P. E-SAM       0.573       28       0.72       1       0.77       50       0.75 <t< td=""><td></td><td></td><td></td><td></td><td>-</td><td></td><td></td><td></td><td></td><td></td></t<>   |       |        |                  |               | -                                |           |           |          |          |           |
| 12       Symp. True's Nume.'um       a, "There," C       length of datament Trees prestert       0.051       71       448       5       0.075       10       0.057       10       407       10       0.057   |       |        |                  |               |                                  |           |           |          |          |           |
| 40       Syan       Parto-Speech       a, Cawly, C. & Ventoris per Yond       0.701 5       0.128 30       0.768 25       0.128 14         40       Syant       Parto-Speech       a, Cawly, C. & Ventoris vock per Sent       0.551 40       0.128 14       0.468 14       0.468 14       0.468 14       0.464 15       0.068 14       0.468 14       0.464 15       0.068 14       0.468 14       0.464 15       0.068 14       0.468 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.463 14       0.464 14       0.463 14       0.464 14       0.463 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14       0.464 14 <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td> |       |        |                  |               | -                                |           |           |          |          |           |
| 40         Sinu Part-of-Speech         a.C.ouW/C. <i>a Countary</i> words per Sent         0.334         20         0.655         0.688         14         0.544         0.444         0.444         0.444         0.445         0.65         0.678         0.58         36           30         Synu Pirasal         a.NAPHC <i>V</i> komparison and per Sant         0.723         2         0.323         57         0.58         0.683         10         0.58         0.683         10         0.58         0.723         2         0.583         10.75         10         0.583         10         0.75         0.683         10         0.75         10         0.583         10         0.75         10         0.683         10         0.75         0.716         5         0.685         10         0.675         0.67         0.606         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717         0.50         0.717   |       | •      |                  |               |                                  |           |           |          |          |           |
| 40         Sam Dynalog         a., Takan, C. et alcans Ava d lemmas PWord         0.434         9         0.456         9         0.457         9         0.458         67         0.433         67           19         Syma Phrasal         a., NDFL, C. W Nom phrases per Sent         0.55         0.406         25         0.66         0.431         67           19         Syma Part-of-Speech         a., FineW, C. W Nom phrases per Sent         0.464         0.417         8         0.662         0.671         18         0.523         0.513         0.523         0.513         0.523         0.513         0.523         0.513         0.523         0.513         0.523         0.513         0.533         0.523         0.513         0.533  | 40    | Synta  | Tree Structure   | as_TreeH_C    | Tree height per Sent             |           |           |          |          |           |
| 19         Synn         Prinsal         as, Arku, C.         Rumus, Ark A (f. Rumus, per Word         0.723         0.733         0.743         0.65         0         0.65         0         0.65         0         0.65         0         0.65         0         0.65         0         0.65         0         0.65         0         0.75         0         0.65         0         0.75         0         0.65         0         0.75         0         0.65         0         0.75         0         0.67         0         0.85         0         0.75         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.69         0         0.53         0         0.67         0         0.69         0         0.53         0         0.67         0         0.67         0         0.67         0         0.67         0         0.69         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.67         0         0.6   |       | -      | -                |               | -                                |           |           |          |          |           |
| 95         Symp         Princed         a., NoPhr., C         104         Prince Normal         0.997         Normal         No   |       |        |                  |               | -                                |           |           |          |          |           |
| 39       Syna Parasal       total #pepositonal phraces       0.47       70       10.89       50.05       11       0.53       30       10.72       3         39       Syna Paroci Speech       Syna Parosal       0.428       71       0.189       50       0.535       4       0.662       51       0.662       51       0.553       4       0.662       51       0.553       4       0.662       51       0.554       0.662       51       0.554       0.673       30       0.622       20       0.663       8       0.97       22       0.673       1       0.064       1       0.445       61       61       61  |       |        |                  |               |                                  |           |           |          |          |           |
| 18       LSAEP Pycholinginstic       \ArkMaUC total AoN (Age of Aquiistino) of words       0427       72       1.0189       50       333       4       0.527       22       0.716       5         16       Syma Phicaal       ax,PPhir_C       # prepositional phrases per Sent       0.513       35       0.417       20       0.012       60.008       28       0.92       7.         15       LSAES mode framiliarity       Schull-CC       Subtect Dig Utel per Sent       0.475       1       0.060       28       0.060       1       0.031       1       0.067       1       0.060       1       0.061       0.061   |       | -      |                  |               |                                  |           |           |          |          |           |
| 18         LSAEm Psycholinginistic         to,AKRW/C total AoA (Age of Acquisition) of words         0.427         72         0.189         60         934         940         953         35         0.563         72         0.189         60         933         45         0.667         70         0.662         18         35         16         16  | 39    | Synta  | Part-of-Speech   | as_FuncW_C    |                                  | 0.468 48  | 0.471 8   | 0.662 57 | 0.673 17 | 0.614 19  |
| 36       Symm Pirsal       ac, Prfrb., C       #prepositional phrascip er Sent       0.513       35       0.417       0       0.607       10       0.609       10       0.533       45         35       LSASem Word Familiarius       SNLICC       Stutelicus US Ig 1007 value per Sent       0.475       1       0.067       10       0.0612       0.077       10       0.051       19       0.061       0.077       10       0.051       10       0.041       0.051       10       0.041       0.051       10       0.041       0.051       10       0.041       0.051       10       0.051 <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>   |       |        |                  |               |                                  |           |           |          |          |           |
| 15       LsCem Word Familiarly as, ShL1C_C       SubteVITE, S       Corected TR       0.474       0.477       0.047       20       0.842       0.145       0.047       0.045<   |       |        |                  |               |                                  |           |           |          |          |           |
| 35       LSAEm Type Token autor CortTR.S.       Corrected TR       0.745       1.       0.006 220       0.436       1       0.446       1       0.447       19       0.067       10.031       20.23       48         34       Syma       Part-of-Speech       as,NDTa_C       # Nom POS taps per Sent       0.551       19       0.034       20       0.049       18.         34       LSAEm Psycholinguisti       as,AABL_C       C and fermas, Bristol norm per Sent       0.542       31       0.349       21       0.552       18       0.049       62       0.573       18       0.691       0.571       1       0.049       85       0.582       73       0.653       19       0.441       0.584       0.691       0.571       2       0.571       2       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.571       0.531       0.511  |       |        |                  |               |                                  |           |           |          |          |           |
| 35         Sxacm         Word Familitary         as, ShLIW, C         Subtrop Probalinguistic         as, ACACL, C         Non POS tags per Sent         0.551         0         0.304         38         0.604         0.607         10         0.608         20         0.408         61           34         LSSEm Psycholinguistic         as, ACACL, C         C         AoA of lemmas, Dristol norm         0.451         50         0.339         0.609         20         0.673         10         0.851         0.532         0.339         0.609         6.637         10         0.861         10         0.863         10         0.863         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10         0.871         10 <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>   |       |        | -                |               |                                  |           |           |          |          |           |
| 34LxSem Psycholinguistic<br>at AABC_La. AACC_LNoA of lemmas, Bristol norm0.532300.330.640610.97320.4998734LxSem Psycholinguistic<br>to AABL_Lcontal lemmas AoA of lemmas, Bristol norm0.451550.438100.808100.51800.637120.4399735SyntaPrasalto ANBL_Lcontam SAOA of lemmas, Bristol norm0.451550.458110.5820.6530.651110.5820.651100.654110.5820.6510.654110.5820.6570.6760.672120.6570.1410.555550.551100.8080.651100.8080.651100.8080.651100.8080.651100.8080.651100.8080.651110.52552595910100.80890.671131210.808100.5717010100.80890.651110.5555259110.555525910100.80890.671131312101310141011101010101010 <td< td=""><td></td><td></td><td>••</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>  |       |        | ••               |               |                                  |           |           |          |          |           |
| 34       LxSem Psycholinguistic ax_AABL_C       intermas And Iemmas, Bristol norm per Sent       0.532       31       0.439       0.563       10       0.573       10       0.564       10       57       0.574       10       0.563       10       0.572       10       1.57       0.341       10       0.580       10       0.573       10       0.57       1.341       10       0.580       10       57       0.341       10       0.580       10       57       0.341       10       0.580       10       55       0.571       10       0.571       10       10       10       10       10       10       10       10       10       10       10       10       10       10       10   | 34    | Synta  | Part-of-Speech   | as_NoTag_C    | # Noun POS tags per Sent         | 0.551 19  | 0.304 38  | 0.624 65 | 0.608 29 | 0.48 61   |
| 14Lx8cmPsycholinguistic<br>to<br>AABLCor AABLCcontaines AoA of lemmas, Briel norm<br>per Sent0.451500.143000.88100.87100.838100.870.433033SyntaPrarsalto.NoPhr.Ctotal #Non phrases0.416760.1480.80980.27220.653935SyntaPrartof-Speechto.CortW.Ctotal #Content vords0.4010.71730.80890.651180.612110.818150.8110.8320.651180.6721222LxSemVariation RatioCorrNV-SCorrected Verb Variation-10.451170.14100.80890.651180.677120.578100.651180.677120.57810600.57110.418100.581130.575555553535100.5710.418100.5555335100.55510.57410.47210.45210.57510.57410.57510.57410.57510.57410.57510.57410.57510.57410.57510.57410.57510.57510.57510.57510.57510.57510.57510.5751<  |       |        |                  |               |                                  |           |           |          |          |           |
| 33       Synta       Partsal       C. Immas, SoA of Jemmas, Bind norm per Sent       0.49       95       0.587       0.603       19       0.43       69         33       Synta       Part-of-Speech       0.ComVC       ComPote-C       0.14       84       0.587       80.08       0.527       2       0.659       10       0.654       11       0.587       10.804       14       0.558       78       0.661       11       0.587       10       0.804       14       0.558       10       0.77       2       0.804       13       0.804       14       0.587       70       0.804       10       0.577       2       0.804       10       0.571       0.141       0.808       0.501       0.533       80       0.575       2       0.534       10       0.571       0.534       10       0.572       0.534       10       0.575       0.401       8.0       0.575       0.401       8.0       0.575       0.533       3       3       3       3       0.777       0.703       0.804       10       0.777       0.703       0.804       10       0.777       0.703       0.804       10       0.777       0.733       0.771       0.733       0.771   |       |        |                  |               | -                                |           |           |          |          |           |
| 33       Synta       Phresal       0. NoPR: C       total # None phrases       0.416       76       0.807       8       0.827       52       0.809       9         33       Synta       Part-of-Speech       ContWV-S       Corrected Veb Nation I       0.071       3       0.801       10.887       10.884       2       0.486       10.88       0.811       0.842       2       0.486       0.77       13         32       Lasem       Variation Ratio       Corrected Veb Nation-1       0.476       6       0.805       12       0.338       0.057       33       0.857       13       0.871       13       0.878       10.884       34       0.057       10.484       0.838       4       0.957       10.484       0.838       4       0.957       10.484       0.851       10.484       0.851       10.484       0.851       10.484       0.851       10.484       0.851       10.484       0.851       10.452       0.435       10.452       0.435       10.452       0.435       10.452       0.433       0.451       10.436       10.452       0.554       0.435       10.452       0.554       0.451       10.452       0.554       0.451       10.451       10.451       0.411  |       |        |                  |               | ,                                |           |           |          |          |           |
| 33       Synta       Part-of-Speech       0.ComV,C       total # Content words       0.71       3       0.808       13       0.804       1       0.874       1         32       LxSem       Variation Ratio       CorrNV-S.       Corrected Now Arraition-1       0.602       11       0.808       15       0.808       15       0.808       9       0.671       3       0.808       9       0.671       0.581       0.551       0.551       0.553       1       0.591       1       0.591       1       0.591       1  |       |        |                  |               |                                  |           |           |          |          |           |
| 12       LxSem Variation Ratio       CorrVVS S       Corrected Verb Variation-1       0.602       15       0.391       15       0.391       60       0.737       1         32       Synta       Part-of-Speech       as_VCB2_C       # Verb POS tags per Sent       0.428       7       0.146       6       0.578       14       0.583       4       0.651       9       0.805       12       0.584       4       0.583       4       0.655       9       0.805       12       0.584       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.583       4       0.573       5       0.441       6       0.219       12       0.584       0.563       6       0.573       5       0.533       5       0.533       5       0.533       5       0.533       5       0.533       5       0.533       5       0.533       5       0.  | 33    | -      |                  | to_ContW_C    | total # Content words            | 0.402 81  | 0.163 71  | 0.804 14 | 0.558 40 | 0.654 11  |
| 12       LxSem       Psycholinguistic       to,A.G.L.C.       total As of lemmas, Cortes and Khanna norm       0.431       70       0.476       0.588       30       0.575         23       Synta       Tree Structure       total length of flatened Trees       0.396       87       0.166       69       0.805       12       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.588       34       0.581       37       0.416       4       0.581       37       0.416       4       0.581       37       0.416       4       0.581       37       0.416       4       0.581       37       0.416       4       0.559       17       0.707       38       0.90       1.55       1.65       4       0.551       17       0.707       38       0.90       1.55       1.65       4       0.551       1.77       23       0.583       4       0.551       1.55       4.051       0.55       1.65       4       0.551       35       1.653       4.051       4.051       4.051 </td <td></td>  |       |        |                  |               |                                  |           |           |          |          |           |
| 32       Synta       Part-of-Speech       a.s_VEAg. C       4 Verb POS tags per Sent       0.428       70       0.476       6       0.578       74       0.588       40       0.505       52         31       LxSem       Variation Ratio       SquaNoV_S       Squared Noun Variation-1       0.641       9       0.121       109       0.815       7       0.401       8       0.583       40       0.583       40       0.583       40       0.583       40       0.583       40       0.581       40       0.777       21       0.402       40       0.584       40       0.581       40       0.579       777       21       0.541       40       0.551       71       0.551       10       0.777       21       0.541       40       0.551       10       0.777       21       0.541       40       0.551       10       0.777       21       0.554       40.55       12       0.579       72       0.586       35       0.555       40       0.621       12       0.586       30       100       10       0.532       80       0.532       80       0.532       80       0.532       80       0.523       80       0.523       80       0.523  |       |        |                  |               |                                  |           |           |          |          |           |
| 32       Synta Tree Sructure squaRvV_S       total length of flattend Trees       0.39       10.66       69       0.805       12       0.583       34         31       LSGem Variation Ratio       SquaRvV_S       Corrected Adjective Variation-1       0.591       12       0.076       13       0.777       23       0.504       54       0.523       44         31       LSGem Variation Ratio       SquaRvV_S       Squared Verb Variation-1       0.559       17       0.076       13       0.777       23       0.504       54       0.635       44       0.635       14       0.559       17       0.076       13       0.574       50.55       44       0.635       12       0.579       0.574       35       0.441       64       0.277       23       0.441       64       0.728       25       1.656       0.441       14       0.575       13       0.556       4       0.676       0.333       10.458       0.476       1.633       10.476       1.633       1.64       0.602       1.74       10.747       10.75       0.574       35       0.414       14       0.574       35       0.414       14       0.672       1.71       1.725       1.755       0.774       1.73   |       |        |                  |               | ,                                |           |           |          |          |           |
| 31       LxSem Variation Ratio       SquarVy_S       Square Non Variation-1       0.49       12       1078       14       108       0.815       7       0.401       84       0.884       34         31       Synta       Part-of-Speech       to_AJTa_C       total # Adjective POS tags       0.911       2       0.077       23       0.504       54       0.525       66         30       LxSem Variation Ratio       Square VerJ       Square VerJ       Square VerJ       0.441       61       0.177       20       0.844       90       0.777       2       0.844       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       90       0.777       2       0.843       80       0.777       2       0.843       80       0.777       2       0.843       80       0.777       2       0.843       80       0.777       2       0.843       80       0.777       2 <td></td> <td></td> <td></td> <td>- 0-</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>   |       |        |                  | - 0-          |                                  |           |           |          |          |           |
| 31       LxSem Variation Ratio       Corr Vy S       Corrected Adjective Variation-1       0.91       12       0.078       13       0.779       21       0.422       70       0.834       30         31       Synta       Part-of-Speech       to_ATTa_C       State       0.414       62       0.19       70       0.73       30       0.934       40       0.635       46         30       Synta       Part-of-Speech       to_NTB_C       total # Adjective POS tags per Sent       0.343       90       0.453       52       0.533       5       0.533       5       0.533       5       0.533       5       0.541       14       0.597       6       0.533       5       0.541       14       0.575       6       0.533       5       0.543       5       0.541       14       0.575       6       0.533       5       0.441       14       0.575       35       0.441       14       0.574       35       0.441       14       0.574       35       0.441       14       0.574       35       0.441       14       0.574       35       0.441       15       0.574       35       0.441       15       0.574       35       0.441       15       0.574<   |       | •      |                  |               | -                                |           |           |          |          |           |
| 30       Lissem Variation Ratio       Squared Verb Variation-I       0.075       17       0.077       128       0.077       22       0.384       90       0.716       4         30       Synta Phrasal       as, VePhr.C.       total # Noun POS tags       0.441       61       0.129       107       0.805       13       0.555       4       0.635       13       0.555       4       0.635       14       0.577       22       0.586       35       0.505       54         20       LxSem Word Familiarity       as, ATBG.C.       # Adjective POS tags per Sent       0.363       26       0.441       4       0.577       25       1.404       44         28       LxSem Word Familiarity       as, SbSDC.C.       SubtexUS CDP value per Sent       0.431       67       0.437       18       0.527       85       0.624       25       0.404       82         28       LxSem Word Familiarity       as, SbSBC.C.       SubtexUS SUBTLCD value per Sent       0.431       68       0.437       18       0.522       85       0.624       24       0.424       8.04       82       0.551       2.52       1.422       0.425       7.5       0.524       4.025       7.5       7.5       7.5   | 31    | LxSem  | Variation Ratio  | CorrAjV_S     | Corrected Adjective Variation-1  | 0.591 12  | 0.078 134 | 0.779 21 | 0.422 70 | 0.584 33  |
| 30       Synta       Part-of-Speech       iv       iv       Norho POS (ags)       0.441 (61)       0.129 (107)       0.805 13       0.55 44       0.635 54         30       Synta       Phrasal       as_VePhr_C       # Verb phrases per Sent       0.333 90       0.455 12       0.59 72       0.563 54       0.505 54         28       Synta       Part-of-Speech       as_Affag_C       # Adjective POS tags per Sent       0.426 66       0.411 14       0.527 82       0.563 76       0.533 51       0.403 81         28       Synta       Part-of-Speech       as_AFTCH_C       Tree height per Word       0.476 45       0.419 12       0.416 14       0.527 84       0.424 24       0.404 82         28       LxSem       Word Familiarity       as_SBSEC_C       SublexUS SUBTLCD value per Sent       0.431 68       0.437 18       0.525 84       0.624 25       0.404 83         28       LxSem       Word Familiarity       as_SBFL_C       SublexUS SUBTLCD value per Sent       0.441 60       0.441 15       0.509 91       0.524 24       0.404 83         28       LxSem       Word Familiarity       as_SBFL_C       SublexUS FREQ(w value per Sent       0.44 60       0.441 16       0.509 91       0.522 47       0.425 75       0.52       14       0.427 76 </td <td></td> <td>-</td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>                         |       | -      | -                |               |                                  |           |           |          |          |           |
| 30       Synta       Phrasel       as, Verhp C       # Verhp hrases per Sent       0.383       90       0.455       12       0.59       72       0.586       35       0.001       84         29       LxSem       Word Familiarity       as_SEDCL_C       SublexUS CDIow value per Sent       0.506       36       0.533       12       0.573       35       0.610       84         28       Synta       Part-of-Speect       as_TFeel, T       Tree heigh per Word       0.476       44       0.078       13       0.754       35       0.614       14       0.002       23       24       0.644       84         28       LxSem Word Familiarity       as_SEDC_C       SublexUS CD# value per Sent       0.431       67       0.437       18       0.427       16       0.477       70       0.525       85       0.624       25       0.602       24       0.404       82         28       LxSem Word Familiarity       as_SBBC_C       SublexUS SUBTLCD value per Sent       0.434       60       0.427       10       0.525       86       0.426       24       0.407       70       0.523       80       0.525       80       0.426       21       0.527       80       0.522   |       |        |                  | • –           |                                  |           |           |          |          |           |
| 29         L×Sem         Word Familiarity         as_AFIg_C         # Adjective POS tags per Sent         0.432         65         0.411         14         0.527         28         0.533         51         0.404         84           28         Disco         Entity Grid         r_NNTo_C         ratio of n transitions total         0.476         44         0.078         135         0.433         51         0.404         84           28         LySem Word Familiarity         as_SDSBC_C         StubtexUS SUBTLCD value per Sent         0.431         61         0.427         15         0.404         82           28         LySem Word Familiarity         as_SDSBC_C         SubtexUS SUBTLCD value per Sent         0.431         68         0.437         18         0.525         84         0.624         2         0.404         82           21         LySem Word Familiarity         as_SDSBC_C         SubtexUS LSUBTUCP value per Sent         0.441         63         0.441         16         0.507         13         0.422         10.425         42         0.422         16         1.435         10.427         10.427         10.427         10.427         10.427         10.427         10.427         10.427         10.427         10.427         10.427   |       |        | •                |               | -                                |           |           |          |          |           |
| 28         Synta         Part-of-Speech         as_ATBac_C         # Adjective POS tags per Sent         0.506 36         0.533 28         0.533 28         0.533 28         0.533 28         0.533 28         0.533 28         0.533 28         0.533 28         0.541 40         0.597 33         0.41 81           28         Synta         Tree Structure         0.476 45         0.417 65         0.417 0         0.525 84         0.624 24         0.404 83           28         LxSem         Word Familiarity         as_SbLCC_C         SubteUS SUBTLCD value per Sent         0.431 67         0.437 17         0.525 84         0.624 24         0.404 83           28         LxSem         Word Familiarity         as_SbLCC_C         SubteUS SUBTLCD value per Sent         0.441 63         0.424 16         0.552 42         0.422 57           7         LxSem Word Familiarity         as_SbLCC_C         SubteUS SUBTLW ralue per Sent         0.44 63         0.441 16         0.559 90         0.542 47         0.425 75           26         LxSem Word Familiarity         to_SbLCC_C         total SubteUS SUBTLWe alue per Sent         0.44 63         0.411 61         0.559 90         0.542 47         0.425 75           26         LxSem Word Familiarity         to_SbLCC_C         total SubteUS SUBTLOW alue per Sent         0.44  |       | -      |                  |               |                                  |           |           |          |          |           |
| 28         Synta         Tree Structure         at_TreeH_C         Tree height per Word         0.476         45         0.416         104         0.597         33         0.41         81           28         LxSem         Word Familiarity         as_SbCD_C         SubtlexUS SUBTLCD value per Sent         0.431         67         0.437         18         0.525         84         0.624         24         0.404         82           28         LxSem         Word Familiarity         as_SbSBC_C         SubtlexUS SUBTLCD value per Sent         0.431         60         0.42         21         0.525         84         0.677         0.573         37         0.77         125         LxSem         Word Familiarity         as_SbFL_C         SubtlexUS SUBTLWP value per Sent         0.443         60         0.414         15         0.509         01.542         48         0.425         76           26         LxSem         Word Familiarity         as_SbFL_C         SubtlexUS SUBTLWP value per Sent         0.444         60         0.414         16         0.509         01.542         48         0.425         76           25         LxSem         Word Familiarity         as_SbFL_C         Cotal SubtlexUS FREQW value         0.348         100  | 28    |        |                  |               | -                                |           |           |          |          | 0.404 84  |
| 28         LxSem         Word Familiarity         as_SbCDC_C         SubdteXUS CD# value per Sent         0.431         67         0.437         17         0.525         84         0.624         24         0.404         82           28         LxSem         Word Familiarity         as_SbSBC_C         SubdteXUS SUBTLCD value         0.37         93         0.14         94         0.754         34         0.407         77         0.573         37           27         LxSem         Word Familiarity         as_SbFLC_C         SubdteXUS SUBTLWF value per Sent         0.44         60         0.41         15         0.509         91         0.542         48         0.425         76           26         LxSem         Word Familiarity         as_SbFLC_C         SubdteXUS SUBTLWF value per Sent         0.44         64         0.441         16         0.509         91         0.542         47         0.425         75         0.43         61         12.5         Word Familiarity         as_SbFLC_C         SubdteXUS ELIOW         Value per Sent         0.44         64         0.441         16         0.50         91         0.426         21         0.525         80         0.55         92         1.44         0.45         0.417  |       |        |                  |               |                                  |           |           |          |          |           |
| 28       LXSem Word Familiarity       as_SbSBC_C       SublexUS SUBTLCD value per Sent       0.431       68       0.437       18       0.525       85       0.624       25       0.604       83         28       LxSem Word Familiarity       os_SbLC_C       total SublexUS Lg10CD value       0.37       93       0.14       94       0.757       43       0.407       77       0.573       37         27       LxSem Word Familiarity       as_SbFL_C       SublexUS FREQIov value per Sent       0.443       60       0.426       21       0.52       86       0.552       42       0.425       75         26       LxSem Word Familiarity       as_SbFL_C       SublexUS SUBTLWF value per Sent       0.446       60       444       16       0.509       90       0.524       47       0.425       75         26       LxSem Word Familiarity       to_SbFL_C       total SublexUS FREQNaulue       0.365       98       0.155       79       0.786       80       0.755       90       0.84       80       0.77       20       8.015       79       0.786       80       0.755       90       0.81       80       0.557       90       0.81       80       0.557       90       0.81       0.557   |       |        |                  |               |                                  |           |           |          |          |           |
| 28       LxSem Word Familiarity       to_SbL1C_C       total SubtlexUS Lg10CD value       0.37       93       0.14       95       0.77       16       0.491       56       0.621       17         27       LxSem Word Familiarity       scynaft       Squared Adjective Variation-1       0.531       32       0.141       94       0.754       34       0.407       77       0.573       37         26       LxSem Word Familiarity       as_SbFR_C       SubtlexUS SPREQW value per Sent       0.44       64       0.441       16       0.509       91       0.524       48       0.425       76         26       LxSem Word Familiarity       as_SbFR_C       SubtlexUS FREQW value per Sent       0.44       64       0.441       16       0.509       90       0.542       47       0.425       75         26       LxSem Word Familiarity       to_SbFrL_C       total SubtlexUS FREQW value       0.365       98       0.15       79       0.786       18       0.477       26       0.431       32       0.551       41         4       LxSem Word Familiarity       to_SbFrL_C       total SubtlexUS SUBTLWF value       0.34       116       0.206       47       0.77       26       0.433       39       0  |       |        | -                |               |                                  |           |           |          |          |           |
| 27       LxSem Variation Ratio       SquaAjV_S       Squared Adjective Variation-1       0.531       32       0.141       94       0.754       34       0.407       77       0.573       37         27       LxSem Word Familiarity as_SbFrL_C       SublexUS FREQIow value per Sent       0.444       60       0.426       21       0.524       42       0.425       76         26       LxSem Word Familiarity as_SbFrQ_C       SublexUS FREQ# value per Sent       0.44       64       0.441       16       0.509       90       0.542       48       0.425       75         26       LxSem Word Familiarity to_SbFrL_C       total SublexUS Lg10WF value       0.365       99       0.141       44       0.44       63       0.411       50       0.799       0.542       48       0.425       75         26       LxSem Word Familiarity to_SbFrL_C       total SublexUS FREQle value       0.365       99       0.141       94       0.757       37       77       26       0.425       75         21       LxSem Word Familiarity to_SbFrL_C       total SublexUS USTUF value       0.341       16       0.206       77       0.673       37       77       26       0.403       82       0.551       42       0.551       <   |       |        |                  |               | -                                |           |           |          |          |           |
| 26       LxSem Word Familiarity       as_SbFW_C       SubtlexUS SUBTLWF value per Sent       0.44       63       0.441       15       0.509       91       0.542       48       0.612       75         26       LxSem Word Familiarity       to_SbLIW_C       total SubtlexUS FREQ# value per Sent       0.44       64       0.441       16       0.509       90       0.542       47       0.425       75         26       LxSem Word Familiarity       to_SABI_C       total SubtlexUS FREQH value       0.365       98       0.155       79       0.476       88       0.611       22         25       LxSem Word Familiarity       to_SABI_C       total SubtlexUS FREQH value       0.34       116       0.206       48       0.77       26       0.418       80       51       11         24       LxSem Word Familiarity       to_SBPR_C       total SubtlexUS SUBTLWF value       0.34       116       0.202       50       0.727       0.403       81       0.551       42         23       ShaTr       Shalte       Misa       Sylla_C       # syllabordinate Clauses       0.367       96       0.202       50       0.71       43       0.462       61       0.419       78       0.406       30  |       |        |                  |               | -                                |           |           |          |          |           |
| 26       LxSem Word Familiarity as_SbFQ_C       SublexUS FREQ# value per Sent       0.44       64       0.44       16       0.509       90       0.542       47       0.425       75         26       LxSem Word Familiarity to_SbL1W_C       total SublexUS Lg10WF value       0.365       99       0.144       93       0.795       17       0.477       58       0.611       22         25       LxSem Word Familiarity to_SbFT_C       total SublexUS FREQ@ value       0.348       109       0.201       51       0.774       24       0.414       74       0.555       40         24       LxSem Word Familiarity to_SbFBW_C       total SublexUS FREQ# value       0.34       116       0.206       48       0.77       26       0.403       81       0.551       41         23       ShafT       Shala       a_Sylla_C       #syllables per Word       0.66       7       0.106       120       0.627       64       0.505       30       0.71       14       0.462       61       0.419       78         23       Synta       Phrasal       to_SuPhr_C       total # Subordinate Clauses       0.367       96       0.202       50       0.71       44       0.33       99       0.474       63  | 27    |        | -                |               | SubtlexUS FREQlow value per Sent | 0.443 60  |           |          |          |           |
| 26       LxSem Word Familiarity       to_SbLIW_C       total SubtlexUS Lg10WF value       0.365       99       0.144       93       0.795       17       0.477       58       0.611       22         25       LxSem Word Familiarity       to_SbFrLC       total SubtlexUS FREQlow value       0.365       98       0.155       79       0.786       18       0.473       59       0.555       40         24       LxSem Word Familiarity       to_SbFrLC       total SubtlexUS FREQW value       0.341       116       0.206       48       0.77       26       0.403       82       0.551       41         24       LxSem Word Familiarity       to_SbSBW_C       total SubtlexUS SUBTLWF value       0.341       116       0.206       48       0.77       27       0.403       82       0.551       42         23       Synta       Phrasal       to_SuPhr_C       total # Subordinate Clauses       0.367       96       0.202       50       0.721       43       0.406       7       7       8       0.406       7       7       0.403       82       57       37       8       0.406       7       0.402       83       0.417       28       0.406       80       0.417       80  |       |        | -                |               | -                                |           |           |          |          |           |
| 25       LxSem       Psycholinguistic       to_AABiL_C       total lemmas AoA of lemmas, Bird norm       0.365       98       0.155       79       0.786       18       0.473       59       0.555       40         24       LxSem       Word Familiarity       to_SbFrQ_C       total SubtexUS FREQ# value       0.34       116       0.206       48       0.77       26       0.403       82       0.551       41         23       ShaTr       Shallow       at_Sylla_C       # syllables per Word       0.34       115       0.206       47       0.77       26       0.403       82       0.551       41         23       Synta       Phrasal       to_SuPhr_C       total SubtexUS SPREQ# value       0.367       96       0.206       48       0.77       27       0.403       81       0.551       41         23       Synta       Phrasal       to_SuPhr_C       total WitexUS SUPtry       total Werb phrases       0.367       96       0.022       50       0.721       43       0.462       0.41       70.7       27       0.423       9.042       83       0.443       68       0.34       16       0.434       75       0.22       0.619       71       44       0.333  |       |        |                  |               |                                  |           |           |          |          |           |
| 25LxSem Word Familiarityto_SbFrL_Ctotal SubtlexUS FREQIow value $0.348$ 109 $0.201$ $51$ $0.774$ $24$ $0.414$ $74$ $0.555$ $40$ 24LxSem Word Familiarityto_SbFrQ_Ctotal SubtlexUS SREC# value $0.34$ $116$ $0.206$ $48$ $0.77$ $26$ $0.403$ $82$ $0.551$ $41$ 24LxSem Word Familiarityto_SbSBW_Ctotal SubtlexUS SUBTLWF value $0.34$ $116$ $0.206$ $48$ $0.77$ $26$ $0.403$ $81$ $0.551$ $42$ 23ShaTrShaInowat_Sylla_C# syllables per Word $0.66$ $7$ $0.106$ $0.66$ $61$ $0.416$ $72$ $0.57$ $38$ 23SyntaPhrasalto_VePhr_Ctotal # verb phrases $0.367$ $96$ $0.202$ $50$ $0.721$ $43$ $0.462$ $61$ $0.419$ $78$ 22AdSem Wiki KnowledgeWTopc15_SNumber of topics, 150 topics extracted from Wiki $0.542$ $23$ $0.007$ $27$ $0.645$ $63$ $0.406$ $80$ $0.434$ $72$ 23SyntaPart-of-Speechto_CorTay_Ctotal # Cordinating Conjunction POS tags $0.364$ $101$ $0.268$ $43$ $0.728$ $38$ $0.443$ $68$ 22SyntaPart-of-Speechto_FuncW_Ctotal # Verb POS tags $0.364$ $101$ $0.268$ $43$ $0.728$ $38$ $0.383$ $91$ $0.597$ $7$ 23SyntaPart-of-Speech <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td>  |       |        |                  |               | -                                |           |           |          |          |           |
| 24       LxSem Word Familiarity       to_SbSBW_C       total SubtlexUS SUBTLWF value       0.34       115       0.206       47       0.77       27       0.403       81       0.551       42         23       ShaTr       Shallow       at_Sylla_C       # syllables per Word       0.66       7       0.106       120       0.627       64       0.505       53       0.37       91         23       Synta       Phrasal       to_VePhr_C       total # Subordinate Clauses       0.367       96       0.202       50       0.721       43       0.462       61       0.419       78       38         23       Synta       Phrasal       to_VePhr_C       total # Verb phrases       0.324       127       0.166       60       0.605       30       0.416       72       0.57       38         24       LxSem Variation Ratio       CorrAvV_S       Corrected AdVerb Variation-1       0.542       23       0.059       154       0.71       44       0.333       99       0.474       63         22       Synta       Part-of-Speech       to_CoTag_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       43       0.728       88       0.838       91   |       |        |                  |               |                                  |           |           |          |          |           |
| 23ShaTrShallowa_Sylla_C# syllables per Word $0.66$ 7 $0.106$ $120$ $0.627$ $64$ $0.505$ $53$ $0.37$ $91$ 23SyntaPhrasalto_SuPhr_Ctotal # Subordinate Clauses $0.367$ $96$ $0.202$ $50$ $0.721$ $43$ $0.462$ $61$ $0.419$ $78$ 23SyntaPhrasalto_VePhr_Ctotal # Verb phrases $0.324$ $127$ $0.169$ $68$ $0.76$ $31$ $0.416$ $72$ $0.57$ $38$ 22AdSem Wiki KnowledgeWTopt15_SNumber of topics, 150 topics extracted from Wiki $0.58$ $15$ $0.007$ $227$ $0.645$ $63$ $0.605$ $30$ $0.191$ $122$ 22SyntaPart-of-Speechto_CoTag_Ccordinating Conjunction POS tags $0.344$ $101$ $0.268$ $43$ $0.728$ $39$ $0.406$ $80$ $0.434$ $72$ 22SyntaPart-of-Speechto_FuncW_Ctotal # Function words $0.331$ $126$ $0.159$ $77$ $0.773$ $25$ $0.385$ $89$ $0.636$ $14$ 22SyntaPart-of-Speechto_FuncW_Ctotal # Verb POS tags $0.381$ $81$ $0.173$ $63$ $0.783$ $38$ $0.383$ $91$ $0.597$ $27$ 21AdSem Wiki KnowledgeWTopc20_SNumber of topics, 200 topics extracted from Wiki $0.584$ $14$ $0.015$ $214$ $0.666$ $52$ $0.326$ $102$ $0.466$ $52$   | 24    | LxSem  | Word Familiarity | to_SbFrQ_C    | -                                | 0.34 116  |           | 0.77 26  |          | 0.551 41  |
| 23       Synta       Phrasal       to_SuPh_C       total # Subordinate Clauses       0.367       96       0.202       50       0.721       43       0.462       61       0.419       78         23       Synta       Phrasal       to_VePhr_C       total # Verb phrases       0.324       127       0.169       68       0.76       31       0.416       72       0.57       38         22       AdSem Wiki Knowledge       WTope15_S       Number of topics, 150 topics extracted from Wiki       0.581       15       0.007       227       0.645       63       0.605       30       0.191       122         23       Synta       Part-of-Speech       to_CoTag_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       3       0.728       39       0.406       80       0.434       72         22       Synta       Part-of-Speech       to_FuncW_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       3       0.738       38       0.436       80       0.434       72         23       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.33       126       0.159       77       0.773       25  |       |        |                  |               |                                  |           |           |          |          |           |
| 23       Synta       Phrasal       to_VePhr_C       total # Verb phrases       0.324       127       0.169       68       0.76       31       0.416       72       0.57       38         22       AdSem Wiki Knowledge       WTopc15_S       Number of topics, 150 topics extracted from Wiki       0.58       15       0.007       227       0.645       63       0.605       30       0.191       122         22       ShaTr       Shallow       at_Chara_C       # characters per Word       0.542       23       0.059       154       0.71       44       0.333       99       0.474       63         22       Synta       Part-of-Speech       to_CoTag_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       43       0.728       39       0.406       80       0.434       72         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.33       126       0.159       77       0.773       25       0.385       89       0.636       14         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.381       16       0.159       77       0.773       25       0.385 <t< td=""><td></td><td></td><td></td><td></td><td>5 1</td><td></td><td></td><td></td><td></td><td></td></t<>  |       |        |                  |               | 5 1                              |           |           |          |          |           |
| 22       AdSem Wiki Knowledge       WTopc15_S       Number of topics, 150 topics extracted from Wiki       0.58       15       0.007       227       0.645       63       0.605       30       0.191       122         22       LxSem Variation Ratio       CorrAvV_S       Corrected AdVerb Variation-1       0.542       23       0.059       154       0.71       44       0.333       99       0.474       63         22       ShaTr       Shallow       at_Chara_C       # characters per Word       0.443       59       0.26       52       0.619       67       0.402       83       0.443       68         22       Synta       Part-of-Speech       to_CoTag_C       total # Function words       0.33       126       0.159       77       0.732       25       0.385       89       0.636       14         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.288       138       0.173       63       0.738       38       0.383       91       0.597       27         21       AdSem Wiki Knowledge       WTopc20_S       Number of topics, 200 topics extracted from Wiki       0.515       34       0.093       128       0.686       20       0.326       102 </td <td></td>   |       |        |                  |               |                                  |           |           |          |          |           |
| 22       LxSem Variation Ratio       CorrAvV_S       Corrected AdVerb Variation-1       0.542       23       0.059       154       0.71       44       0.333       99       0.474       63         22       ShaTr       Shallow       at_Chara_C       # characters per Word       0.443       59       0.2       52       0.619       67       0.402       83       0.443       68         22       Synta       Part-of-Speech       to_CoTag_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       43       0.728       39       0.406       80       0.434       72         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.33       126       0.159       77       0.773       25       0.385       89       0.636       14         20       LxSem       Wriation Ratio       Squared AdVerb Variation-1       0.584       14       0.015       214       0.616       68       0.617       27       0.137       138         20       LxSem       Word Familiarity       to_SbCDL_C       total # Werb POS tags       0.387       89       0.357       26       0.532       80       0.495       5       0.265  |       |        |                  |               |                                  |           |           |          |          |           |
| 22       Synta       Part-of-Speech       to_CoTag_C       total # Coordinating Conjunction POS tags       0.364       101       0.268       43       0.728       39       0.406       80       0.434       72         22       Synta       Part-of-Speech       to_FuncW_C       total # Function words       0.33       126       0.159       77       0.773       25       0.385       89       0.636       14         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.288       138       0.173       63       0.738       38       0.383       91       0.597       27         21       AdSem Wiki Knowledge       WTopc20_S       Number of topics, 200 topics extracted from Wiki       0.515       34       0.093       128       0.686       52       0.326       102       0.46       65       137       63       0.738       88       0.405       5       0.666       52       0.646       52       0.326       102       0.46       65       52       0.626       112       19       137       43       0.728       89       0.357       26       0.532       80       0.495       5       0.265       112       19       13       1   | 22    | LxSem  | Variation Ratio  | CorrAvV_S     |                                  |           | 0.059 154 | 0.71 44  | 0.333 99 | 0.474 63  |
| 22       Synta       Part-of-Speech       to_FuncW_C       total # Function words       0.33       126       0.159       77       0.73       25       0.385       89       0.636       14         22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.288       138       0.173       63       0.738       38       0.383       91       0.597       27         21       AdSem Wiki Knowledge       WTopc20_S       Number of topics, 200 topics extracted from Wiki       0.584       14       0.015       214       0.616       68       0.617       27       0.137       138         20       LxSem Variation Ratio       SquaAvV_S       Squared AdVerb Variation-1       0.584       14       0.015       214       0.616       68       0.647       27       0.137       138         19       Synta       Phrasal       as_SuPhr_C       # subordinate Clauses per Sent       0.347       89       0.357       26       0.532       80       0.495       55       0.265       1.35       53         18       LxSem Type Token Ratio       UberTTR_S       Uber Index       0.646       8       0.041       174       0.369       122       0.052       23  |       |        |                  |               |                                  |           |           |          |          |           |
| 22       Synta       Part-of-Speech       to_VeTag_C       total # Verb POS tags       0.288       138       0.173       63       0.738       38       0.383       91       0.597       27         21       AdSem Wiki Knowledge       WTopc20_S       Number of topics, 200 topics extracted from Wiki       0.584       14       0.015       214       0.616       68       0.617       27       0.137       138         20       LxSem Variation Ratio       SquaAvU_S       Squared AdVerb Variation-1       0.515       34       0.093       128       0.686       52       0.326       102       0.46       65         19       Synta       Phrasal       as_SuPhr_C       # Subordinate Clauses per Sent       0.348       107       0.148       87       0.764       30       0.394       85       0.515       33         18       LxSem Type Token Ratio       Uber TTR_S       Uber Index       0.646       8       0.041       174       0.369       122       0.632       29       0.652       123       0.075       180         18       AdSem Wiki Knowledge       WTopc10_S       Number of topics, 100 topics extracted from Wiki       0.492       1.032       1.09       1.55       1.80       0.675  |       |        |                  | -             |                                  |           |           |          |          |           |
| 21AdSem Wiki KnowledgeWTopc20_SNumber of topics, 200 topics extracted from Wiki $0.584$ 14 $0.015$ $214$ $0.616$ $68$ $0.617$ $27$ $0.137$ $138$ 20LxSem Variation RatioSquaAvV_SSquared AdVerb Variation-1 $0.515$ $34$ $0.093$ $128$ $0.686$ $52$ $0.326$ $102$ $0.46$ $65$ 19SyntaPhrasalas_SuPhr_C# Subordinate Clauses per Sent $0.387$ $89$ $0.357$ $26$ $0.532$ $80$ $0.495$ $55$ $0.265$ $112$ 19LxSem Word Familiarityto_SbCDL_Ctotal SublexUS CDlow value $0.348$ $107$ $0.148$ $87$ $0.764$ $30$ $0.394$ $85$ $0.513$ $53$ 18LxSem Type Token RatioUber TTR_SUber Index $0.646$ $8$ $0.041$ $174$ $0.369$ $112$ $0.109$ $173$ $0.592$ $26$ 18AdSem Wiki KnowledgeWTopc10_SNumber of topics, 100 topics extracted from Wiki $0.52$ $33$ $0.004$ $229$ $0.552$ $79$ $0.552$ $43$ $0.075$ $80$ 18AdSem Wiki KnowledgeWNois20_SSemantic Noise, 200 topics extracted from Wiki $0.48$ $3$ $0.193$ $56$ $0.691$ $48$ $0.406$ $79$ $0.299$ $106$ 18LxSem Word Familiarityto_SbSDC_Ctotal SublexUS SUBTLCD value $0.347$ $111$ $0.146$ $90$ $0.764$ $28$ $0.392$ $88$ $0.515$ $52$   |       |        | •                |               |                                  |           |           |          |          |           |
| 20       LxSem Variation Ratio       SquaAvV_S       Squared AdVerb Variation-1       0.515       34       0.093       128       0.686       52       0.326       102       0.46       65         19       Synta       Phrasal       as_SuPhr_C       #Subordinate Clauses per Sent       0.387       89       0.357       26       0.532       80       0.495       55       0.265       112         19       LxSem       Word Familiarity       to_SbCDL_C       total SublexUS CDlow value       0.348       107       0.148       87       0.764       30       0.394       85       0.513       53         18       LxSem       Type Token Ratio       Uber TTR_S       Uber Index       0.646       0.041       174       0.369       112       0.091       13       0.592       26         18       AdSem       Wiki Knowledge       WTopc10_S       Number of topics, 100 topics extracted from Wiki       0.422       0.522       79       0.552       43       0.075       180         18       Ayna       Part-of-Speech       to_Surfag_C       total # Subordinating Conjunction POS tags       0.347       111       0.146       91       0.764       28       0.322       88       0.515       52  |       |        |                  | - 0-          | e e                              |           |           |          |          |           |
| 19       Synta       Phrasal       as_SuPhr_C       # Subordinate Clauses per Sent       0.387       89       0.357       26       0.532       80       0.495       55       0.265       112         19       LxSem       Word Familiarity       to_SbCDL_C       total SubtlexUS CDlow value       0.348       107       0.148       87       0.764       30       0.394       85       0.513       53         18       LxSem       Type Token Ratio       UberTTR_S       Uber Index       0.646       8       0.041       174       0.369       112       0.109       173       0.599       26         18       AdSem       Wiki Knowledge       WTopt10_S       Number of topics, 100 topics extracted from Wiki       0.52       33       0.004       229       0.532       79       0.552       43       0.075       180         18       AdSem       Wiki Knowledge       WNois2_S       Semantic Noise, 200 topics extracted from Wiki       0.48       0.492       41       0.042       10       0.566       75       0.572       37       0.022       10         18       Synta       Part-of-Speech       to_SUTAg_C       total # Subordinating Conjunction POS tags       0.347       111       0.146       <  |       |        | -                |               |                                  |           |           |          |          |           |
| 18       LxSem Type Token Ratio UberTTR_S       Uber Index       0.646       8       0.041       174       0.369       12       0.109       173       0.599       26         18       AdSem Wiki Knowledge       WTopc10_S       Number of topics, 100 topics extracted from Wiki       0.52       33       0.004       229       0.532       79       0.552       43       0.075       180         18       AdSem Wiki Knowledge       WNois20_S       Semantic Noise, 200 topics extracted from Wiki       0.492       10.322       109       0.567       0.572       37       0.025       221         18       Synta       Part-of-Speech       to_Subject/SESC_C       total # Subordinating Conjunction POS tags       0.44       83       0.193       56       0.691       48       0.407       9       0.299       106         18       LxSem Word Familiarity       to_SbSDC_C       total SubtlexUS CD# value       0.347       110       0.146       90       0.764       29       0.392       87       0.515       51         18       LxSem Word Familiarity       to_SbCDC_C       total SubtlexUS CD# value       0.347       110       0.146       90       0.764       29       0.392       87       0.515       51 <td></td> <td>Synta</td> <td>Phrasal</td> <td>as_SuPhr_C</td> <td>-</td> <td>0.387 89</td> <td>0.357 26</td> <td>0.532 80</td> <td>0.495 55</td> <td>0.265 112</td>                 |       | Synta  | Phrasal          | as_SuPhr_C    | -                                | 0.387 89  | 0.357 26  | 0.532 80 | 0.495 55 | 0.265 112 |
| 18       AdSem Wiki Knowledge       WTopc10_S       Number of topics, 100 topics extracted from Wiki       0.52       33       0.004       229       0.532       79       0.552       43       0.075       180         18       AdSem Wiki Knowledge       WNois20_S       Semantic Noise, 200 topics extracted from Wiki       0.492       41       0.032       190       0.566       75       0.572       37       0.025       221         18       Synta       Part-of-Speech       to_SubTag_C       total # Subordinating Conjunction POS tags       0.483       0.193       56       0.691       48       0.406       79       0.292       106         18       LxSem Word Familiarity       to_SbSBC_C       total SubtlexUS SUBTLCD value       0.347       111       0.146       91       0.764       28       0.392       88       0.515       52         18       LxSem Word Familiarity       to_SbCDC_C       total SubtlexUS CD# value       0.347       110       0.146       90       0.764       29       0.392       87       0.515       51  |       |        | -                |               |                                  |           |           |          |          |           |
| 18       AdSem Wiki Knowledge       WNois20_S       Semantic Noise, 200 topics extracted from Wiki       0.492       41       0.032       190       0.566       75       0.572       37       0.025       221         18       Synta       Part-of-Speech       to_Surag_C       total # Subordinating Conjunction POS tags       0.483       0.193       56       0.691       48       0.406       79       0.292       106         18       LxSem Word Familiarity       to_SbSBC_C       total SubtlexUS SUBTLCD value       0.347       111       0.146       91       0.764       28       0.392       88       0.515       52         18       LxSem Word Familiarity       to_SbCDC_C       total SubtlexUS CD# value       0.347       110       0.146       90       0.764       29       0.392       87       0.515       51  |       |        |                  |               |                                  |           |           |          |          |           |
| 18       Synta       Part-of-Speech       to_SuTag_C       total # Subordinating Conjunction POS tags       0.4       83       0.193       56       0.691       48       0.406       79       0.299       106         18       LxSem       Word Familiarity       to_SbSBC_C       total SubtlexUS SUBTLCD value       0.347       111       0.146       91       0.764       28       0.392       88       0.515       52         18       LxSem       Word Familiarity       to_SbCDC_C       total SubtlexUS CD# value       0.347       110       0.146       90       0.764       29       0.392       87       0.515       51   |       |        |                  |               |                                  |           |           |          |          |           |
| 18         LxSem Word Familiarity to_SbSBC_C         total SubtlexUS SUBTLCD value         0.347         111         0.146         91         0.764         28         0.315         52           18         LxSem Word Familiarity to_SbCDC_C         total SubtlexUS CD# value         0.347         111         0.146         91         0.764         28         0.392         88         0.515         52           18         LxSem Word Familiarity to_SbCDC_C         total SubtlexUS CD# value         0.347         110         0.146         90         0.764         29         0.392         87         0.515         51   |       |        | -                |               | -                                |           |           |          |          |           |
| 18         LxSem Word Familiarity to_SbCDC_C total SubtlexUS CD# value         0.347 110         0.146 90         0.764 29         0.392 87         0.515 51  |       |        | •                | - 0-          | total SubtlexUS SUBTLCD value    |           |           |          |          |           |
| 18         Synta         Part-of-Speech         to_AvTag_C         total # Adverb POS tags         0.342         114         0.17         67         0.726         40         0.352         96         0.469         64   |       |        | -                |               |                                  |           |           |          |          |           |
|   | 18    | Synta  | Part-of-Speech   | to_AvTag_C    | total # Adverb POS tags          | 0.342 114 | 0.17 67   | 0.726 40 | 0.352 96 | 0.469 64  |

Table 10: Part A. The full generalizability ranking of handcrafted linguistic features under Approach A. r: Pearson's correlation between the feature and the dataset. rk: the feature's correlation ranking on the specific dataset.

|          |                |                                    |                          | Feature   | ССВ                    | WBT                    | CAM                    | СКС                    | OSE                    |
|----------|----------------|------------------------------------|--------------------------|---|------------------------|------------------------|------------------------|------------------------|------------------------|
| Score    | Branch         | Subgroup                           | LingFeat Code            | Brief Explanation   | r rk                   |
| 17       | AdSem          | Wiki Knowledge                     | WTopc05_S                | Number of topics, 50 topics extracted from Wiki   | 0.549 22               | 0.033 186              | 0.514 89               | 0.533 50               | 0.042 203              |
| 17       |                | Part-of-Speech                     | as_AvTag_C               | # Adverb POS tags per Sent  | 0.32 129               | 0.292 41               | 0.526 83               | 0.43 67                | 0.415 79               |
| 16<br>16 |                | Type Token Ratio<br>Wiki Knowledge | WRich15_S                | Bi-Logarithmic TTR<br>Semantic Richness, 150 topics extracted from Wiki                               | 0.591 13<br>0.495 39   | 0.062 149<br>0.02 208  | 0.07 200<br>0.48 95    | 0.001 229<br>0.549 45  | 0.523 47<br>0.037 209  |
| 16       |                | Part-of-Speech                     | as_CoTag_C               | # Coordinating Conjunction POS tags per Sent  | 0.38 91                | 0.411 24               | 0.463 97               | 0.442 66               | 0.293 107              |
| 15       | Synta          |                                    | to_AvPhr_C               | total # Adverb phrases  | 0.356 105              | 0.17 66                | 0.705 45               | 0.298 111              | 0.432 73               |
| 15       |                | Shallow                            | TokSenL_S                | log(total # tokens)/log(total # sentence)   | 0.293 137              | 0.352 29               | 0.297 130              | 0.544 46               | 0.198 121              |
| 14<br>13 | AdSem<br>Synta | Wiki Knowledge<br>Phrasal          | at PrPhr C               | Semantic Richness, 200 topics extracted from Wiki<br># prepositional phrases per Word                 | 0.465 50<br>0.57 16    | 0.029 195<br>0.133 103 | 0.446 102<br>0.316 124 | 0.556 41<br>0.323 105  | 0.027 219 0.366 92     |
| 13       |                | Phrasal                            | ra_NoPrP_C               | ratio of Noun phrases # to Prep phrases #   | 0.477 43               | 0.149 83               | 0.34 120               | 0.345 97               | 0.389 87               |
| 13       |                | Entity Grid                        | ra_SNTo_C                | ratio of sn transitions to total  | 0.448 58               | 0.019 210              | 0.514 88               | 0.196 133              | 0.518 49               |
| 13<br>13 |                | Word Familiarity<br>Part-of-Speech | at_SbL1C_C<br>as_SuTag_C | SubtlexUS Lg10CD value per Word<br># Subordinating Conjunction POS tags per Sent                      | 0.408 78<br>0.366 97   | 0.161 75<br>0.295 39   | 0.541 78<br>0.407 105  | 0.204 130<br>0.427 68  | 0.392 86<br>0.151 131  |
| 13       | -              | Shallow                            | TokSenS_S                | # Subordinating Conjunction POS tags per Sent<br>sqrt(total # tokens x total # sentence)              | 0.300 97               | 0.293 39               | 0.407 103              | 0.427 08 0.249 121     | 0.131 131 0.498 59     |
| 13       | Synta          | Tree Structure                     | to_TreeH_C               | total Tree height of all sentences  | 0.27 145               | 0.069 143              | 0.755 33               | 0.309 108              | 0.515 50               |
| 13       |                | Phrasal                            | as_AvPhr_C               | # Adverb phrases per Sent   | 0.244 152              | 0.328 34               | 0.427 103              | 0.38 92                | 0.356 93               |
| 12<br>12 |                | Entity Grid<br>Phrasal             | ra_NSTo_C<br>to_AjPhr_C  | ratio of ns transitions to total<br>total # Adjective phrases   | 0.426 73<br>0.339 120  | 0.033 187<br>0.182 62  | 0.516 87<br>0.682 53   | 0.266 117<br>0.327 101 | 0.505 56<br>0.271 111  |
| 11       |                | Wiki Knowledge                     | -                        | Semantic Noise, 50 topics extracted from Wiki   | 0.462 53               | 0.061 150              | 0.455 100              | 0.412 75               | 0.118 151              |
| 11       | Synta          | Phrasal                            | ra_PrNoP_C               | ratio of Prep phrases # to Noun phrases #   | 0.421 75               | 0.162 74               | 0.276 135              | 0.344 98               | 0.37 90                |
| 11       |                | Shallow                            | TokSenM_S                | total # tokens x total # sentence   | 0.189 173              | 0.112 116              | 0.674 55               | 0.177 140              | 0.486 60               |
| 10<br>10 |                | Phrasal<br>Entity Density          | ra_VeNoP_C<br>at_UEnti_C | ratio of Verb phrases # to Noun phrases #<br>number of unique Entities per Word                       | 0.46 54<br>0.127 197   | 0.164 70<br>0.307 37   | 0.124 174<br>0.548 77  | 0.041 209<br>0.253 119 | 0.027 220 0.124 149    |
| 9        |                | Variation Ratio                    | SimpNoV_S                | Noun Variation-1  | 0.499 38               | 0.087 130              | 0.038 212              | 0.031 213              | 0.337 95               |
| 9        |                | Part-of-Speech                     | at_VeTag_C               | # Verb POS tags per Word  | 0.431 69               | 0.187 61               | 0.076 196              | 0.111 171              | 0.011 224              |
| 9        |                | Word Familiarity                   |                          | SubtlexUS Lg10WF value per Word   | 0.399 84               | 0.089 129              | 0.531 81               | 0.24 123               | 0.412 80               |
| 9<br>9   |                | Part-of-Speech<br>Word Familiarity | ra_VeNoT_C<br>at_SbSBC_C | ratio of Verb POS # to Noun POS #<br>SubtlexUS SUBTLCD value per Word                                 | 0.397 86<br>0.37 94    | 0.198 53<br>0.032 192  | 0.234 142<br>0.492 93  | 0.171 142<br>0.324 103 | 0.067 186<br>0.435 71  |
| 9        |                | Word Familiarity                   |                          | SubtlexUS CD# value per Word  | 0.37 95                | 0.032 192              | 0.492 94               | 0.324 103              | 0.435 70               |
| 9        | Synta          |                                    | as_AjPhr_C               | # Adjective phrases per Sent  | 0.323 128              | 0.239 46               | 0.387 106              | 0.357 95               | 0.157 127              |
| 9        |                | WB Knowledge                       | BClar15_S                | Semantic Clarity, 150 topics extracted from WeeBit  | 0.025 221              | 0.161 76               | 0.38 108               | 0.481 57               | 0.315 100              |
| 8<br>8   |                | Wiki Knowledge<br>Entity Density   | WNois15_S<br>at_EntiM_C  | Semantic Noise, 150 topics extracted from Wiki<br>number of Entities Mentions #s per Word             | 0.388 88<br>0.17 180   | 0.033 188<br>0.204 49  | 0.454 101<br>0.501 92  | 0.454 63<br>0.292 112  | 0.006 226 0.127 146    |
| 8        |                | WB Knowledge                       | BClar20_S                | Semantic Clarity, 200 topics extracted from WeeBit  | 0.004 227              | 0.147 88               | 0.3 129                | 0.462 60               | 0.308 104              |
| 7        | -              | Phrasal                            | ra_PrVeP_C               | ratio of Prep phrases # to Verb phrases #   | 0.485 42               | 0.055 157              | 0.184 158              | 0.189 136              | 0.219 117              |
| 7<br>7   |                | Word Familiarity<br>Part-of-Speech | at_SbCDL_C<br>ra_CoNoT_C | SubtlexUS CDlow value per Word  | 0.362 102 0.02 224     | 0.047 166<br>0.277 42  | 0.474 96<br>0.159 163  | 0.31 107<br>0.013 222  | 0.431 74<br>0.132 142  |
| 7        |                | Part-of-Speech                     | at_CoTag_C               | ratio of Coordinating Conjunction POS # to Noun POS #<br># Coordinating Conjunction POS tags per Word | 0.02 224 0.218 161     | 0.277 42 0.267 44      | 0.02 220               | 0.013 222 0.111 172    | 0.132 142 0.087 169    |
| 7        |                | Part-of-Speech                     | ra_NoCoT_C               | ratio of Noun POS # to Coordinating Conjunction #   | 0.022 222              | 0.254 45               | 0.019 221              | 0.053 201              | 0.109 157              |
| 6        |                | Phrasal                            | ra_VePrP_C               | ratio of Verb phrases # to Prep phrases #   | 0.475 46               | 0.018 211              | 0.301 127              | 0.255 118              | 0.249 114              |
| 6<br>6   |                | Entity Grid<br>WB Knowledge        | ra_XNTo_C<br>BTopc15_S   | ratio of xn transitions to total<br>Number of topics, 150 topics extracted from WeeBit                | 0.339 119<br>0.133 193 | 0.103 124<br>0.146 92  | 0.658 59<br>0.209 151  | 0.327 100<br>0.416 73  | 0.29 108<br>0.03 217   |
| 6        |                | Word Familiarity                   |                          | SubtlexUS SUBTLWF value per Word  | 0.181 175              | 0.196 54               | 0.095 184              | 0.021 220              | 0.109 156              |
| 6        |                | Word Familiarity                   | at_SbFrQ_C               | SubtlexUS FREQ# value per Word  | 0.181 174              | 0.196 55               | 0.095 183              | 0.021 219              | 0.109 155              |
| 5<br>5   |                | Part-of-Speech                     | ra_NoVeT_C               | ratio of Noun POS # to Verb POS #   | 0.432 66<br>0.364 100  | 0.118 111<br>0.002 232 | 0.149 168<br>0.33 123  | 0.112 170<br>0.411 76  | 0.051 197 0.041 206    |
| 5        |                | Wiki Knowledge<br>Entity Grid      | ra_NXTo_C                | Semantic Richness, 100 topics extracted from Wiki ratio of nx transitions to total                    | 0.339 118              | 0.002 232              | 0.62 66                | 0.28 116               | 0.041 200              |
| 5        |                | Part-of-Speech                     | at_FuncW_C               | # Function words per Word   | 0.28 142               | 0.04 175               | 0.181 159              | 0.461 62               | 0.032 215              |
| 5        |                | WB Knowledge                       | BTopc20_S                | Number of topics, 200 topics extracted from WeeBit  | 0.25 150               | 0.135 99               | 0.025 215              | 0.418 71               | 0.044 198              |
| 5<br>5   |                | Variation Ratio<br>Part-of-Speech  | SimpVeV_S<br>ra_VeCoT_C  | Verb Variation-1<br>ratio of Verb POS # to Coordinating Conjunction #                                 | 0.286 139<br>0.192 172 | 0.048 165<br>0.172 64  | 0.081 193<br>0.134 171 | 0.003 226 0.022 218    | 0.48 62<br>0.054 194   |
| 5        |                | Word Familiarity                   |                          | SubtlexUS FREQIow value per Word  | 0.176 178              |                        | 0.061 203              | 0.001 228              | 0.09 165               |
| 4        | Synta          | Phrasal                            | at_NoPhr_C               | # Noun phrases per Word   | 0.424 74               | 0.066 146              | 0.089 188              | 0.005 224              | 0.042 202              |
| 4<br>4   |                | Type Token Ratio                   |                          | unique tokens/total tokens (TTR)<br>Semantic Noice, 100 topics extracted from Wikin                   | 0.375 92               | 0.025 200              | 0.367 113              | 0.163 147              | 0.344 94               |
| 4        |                | Wiki Knowledge<br>Phrasal          | at_SuPhr_C               | Semantic Noise, 100 topics extracted from Wikip<br># Subordinate Clauses per Word                     | 0.34 117<br>0.204 165  | 0.021 207<br>0.157 78  | 0.376 109<br>0.246 140 | 0.426 69<br>0.314 106  | 0.03 216<br>0.073 182  |
| 4        | Synta          | Phrasal                            | ra_SuNoP_C               | ratio of Subordinate Clauses # to Noun phrases #  | 0.081 203              | 0.163 72               | 0.224 146              | 0.307 109              | 0.086 170              |
| 4        |                | WB Knowledge                       | BNois15_S                | Semantic Noise, 150 topics extracted from WeeBit  | 0.035 214              | 0.162 73               | 0.341 119              | 0.221 127              | 0.091 164              |
| 3<br>3   |                | Part-of-Speech<br>Phrasal          | ra_AjVeT_C<br>ra_NoVeP_C | ratio of Adjective POS # to Verb POS #<br>ratio of Noun phrases # to Verb phrases #                   | 0.411 77<br>0.406 79   | 0.034 185<br>0.068 145 | 0.133 172 0.069 201    | 0.156 150<br>0.031 212 | 0.005 227<br>0.019 223 |
| 3        | 2              | Wiki Knowledge                     |                          | Semantic Richness, 50 topics extracted from Wiki  | 0.405 80               | 0.063 145              | 0.347 117              | 0.301 110              | 0.019 223              |
| 3        | Synta          | -                                  | ra_AvPrP_C               | ratio of Adv phrases # to Prep phrases #  | 0.4 82                 | 0.014 217              | 0.222 147              | 0.196 135              | 0.115 152              |
| 3        |                | Variation Ratio                    | SimpAjV_S                | Adjective Variation-1   | 0.398 85               | 0.109 118              | 0.279 134              | 0.073 192              | 0.201 120              |
| 3<br>3   |                | Phrasal<br>Part-of-Speech          | ra_NoSuP_C<br>ra_NoAjT_C | ratio of Noun phrases # to Subordinate Clauses #<br>ratio of Noun POS # to Adjective POS #            | 0.157 185<br>0.121 199 | 0.153 80<br>0.152 81   | 0.228 145<br>0.125 173 | 0.052 205<br>0.114 169 | 0.04 207<br>0.004 228  |
| 3        |                | Part-of-Speech                     | ra_SuNoT_C               | ratio of Subordinating Conjunction POS # to Noun POS #  | 0.085 202              | 0.132 81 0.149 82      | 0.039 211              | 0.114 109              | 0.004 228              |
| 3        | AdSem          | WB Knowledge                       | BNois20_S                | Semantic Noise, 200 topics extracted from WeeBit  | 0.129 196              | 0.148 85               | 0.202 153              | 0.167 144              | 0.032 214              |
| 2        |                | Phrasal                            | ra_VeSuP_C               | ratio of Verb phrases # to Subordinate Clauses #  | 0.349 106              | 0.137 98               | 0.307 126              | 0.127 167              | 0.043 200              |
| 2<br>2   | -              | Phrasal<br>Part-of-Speech          | ra_SuVeP_C<br>ra_CoFuW_C | ratio of Subordinate Clauses # to Verb phrases #<br>ratio of Content words to Function words          | 0.345 113<br>0.284 141 | 0.052 160<br>0.023 203 | 0.343 118<br>0.2 154   | 0.376 93<br>0.376 94   | 0.083 172<br>0.042 201 |
| 2        | Disco          | Entity Grid                        | ra_ONTo_C                | ratio of on transitions to total  | 0.333 123              | 0.023 203              | 0.288 133              | 0.06 199               | 0.383 88               |
| 2        | Disco          | Entity Grid                        | ra_NOTo_C                | ratio of no transitions to total  | 0.348 108              | $0.022\ 204$           | 0.383 107              | 0.056 200              | 0.378 89               |
| 2        |                | WB Knowledge                       | BRich10_S                | Semantic Richness, 100 topics extracted from WeeBit   | 0.196 170              | 0.044 171              | 0.369 111              | 0.035 210              | 0.336 96               |
| 2<br>2   | Disco<br>Synta | Entity Density<br>Part-of-Speech   | to_UEnti_C<br>ra_AjCoT_C | total number of unique Entities<br>ratio of Adjective POS # to Coordinating Conjunction #             | 0.308 134 0.0 229      | 0.132 105<br>0.148 86  | 0.3 128<br>0.049 207   | 0.023 216<br>0.091 181 | 0.31 102 0.077 177     |
| 2        | Synta          | Part-of-Speech                     | ra_AjNoT_C               | ratio of Adjective POS # to Coordinating Conjunction #  | 0.074 205              | 0.146 89               | 0.031 213              | 0.068 195              | 0.041 205              |
|          |                | -                                  | . –                      |   |                        |                        |                        |                        |                        |

Table 11: Part B. The full generalizability ranking of handcrafted linguistic features under Approach A.

|       |          |                  |               | Feature  | ССВ       | WB    | Т   | CAM       | СКС       | OSE       |
|-------|----------|------------------|---------------|--|-----------|-------|-----|-----------|-----------|-----------|
| Score | e Branch | Subgroup         | LingFeat Code | Brief Explanation                                      | r rk      | r     | rk  | r rk      | r rk      | r rk      |
| 1     | Synta    | Part-of-Speech   | ra_SuVeT_C    | ratio of Subordinating Conjunction POS # to Verb POS # | 0.36 103  | 0.053 | 159 | 0.109 177 | 0.282 115 | 0.137 139 |
| 1     | Synta    | Part-of-Speech   | ra_AjAvT_C    | ratio of Adjective POS # to Adverb POS #               | 0.357 104 | 0.042 | 172 | 0.056 204 | 0.091 180 | 0.044 199 |
| 1     | LxSem    | Psycholinguistic | at_AABrL_C    | lemmas AoA of lemmas, Bristol norm per Word            | 0.333 124 | 0.029 | 194 | 0.462 98  | 0.284 113 | 0.217 118 |
| 1     | LxSem    | Psycholinguistic | at_AACoL_C    | AoA of lemmas, Cortese and Khanna norm per Word        | 0.333 125 | 0.029 | 193 | 0.462 99  | 0.284 114 | 0.217 119 |
| 1     | AdSem    | WB Knowledge     | BNois10_S     | Semantic Noise, 100 topics extracted from WeeBit       | 0.193 171 | 0.036 | 180 | 0.37 110  | 0.161 149 | 0.33 97   |
| 1     | AdSem    | WB Knowledge     | BNois05_S     | Semantic Noise, 50 topics extracted from WeeBit        | 0.158 184 | 0.011 | 219 | 0.351 116 | 0.15 153  | 0.325 98  |
| 1     | AdSem    | WB Knowledge     | BTopc10_S     | Number of topics, 100 topics extracted from WeeBit     | 0.197 169 | 0.038 | 179 | 0.364 114 | 0.166 145 | 0.323 99  |
| 1     | Disco    | Entity Density   | to_EntiM_C    | total number of Entities Mentions #s                   | 0.139 191 | 0.02  | 209 | 0.335 122 | 0.0 230   | 0.312 101 |
| 1     | AdSem    | WB Knowledge     | BRich05_S     | Semantic Richness, 50 topics extracted from WeeBit     | 0.126 198 | 0.051 | 162 | 0.24 141  | 0.051 207 | 0.309 103 |
| 1     | LxSem    | Psycholinguistic | at_AABiL_C    | lemmas AoA of lemmas, Bird norm per Word               | 0.203 166 | 0.11  | 117 | 0.266 138 | 0.053 202 | 0.302 105 |
| 1     | Synta    | Tree Structure   | at_FTree_C    | length of flattened Trees per Word                     | 0.28 143  | 0.14  | 96  | 0.097 182 | 0.1 177   | 0.152 130 |
| 1     | Synta    | Phrasal          | at_VePhr_C    | # Verb phrases per Word                                | 0.31 132  | 0.138 | 97  | 0.079 194 | 0.032 211 | 0.009 225 |
| 1     | Synta    | Part-of-Speech   | ra_NoAvT_C    | ratio of Noun POS # to Adverb POS #                    | 0.261 147 | 0.133 | 102 | 0.101 180 | 0.052 204 | 0.034 212 |
| 1     | Synta    | Part-of-Speech   | ra_CoVeT_C    | ratio of Coordinating Conjunction POS # to Verb POS #  | 0.302 135 | 0.133 | 104 | 0.023 218 | 0.133 164 | 0.088 168 |

Table 12: Part C. The full generalizability ranking of handcrafted linguistic features under Approach A.

| Feature  |         |                                      |                          | eature   | ССВ               | WBT               | CAM                  | СКС               | OSE                  |
|----------|---------|--------------------------------------|--------------------------|--|-------------------|-------------------|----------------------|-------------------|----------------------|
| Score    | Branch  | Subgroup                             | LingFeat Code            | Brief Explanation  | r rk              | r rk              | r rk                 | r rk              | r rk                 |
| 35       | LxSem   | Psycholinguistic                     | as_AAKuL_C               | lemmas AoA of lemmas per Sent                                  | 0.54 25           | 0.505 1           | 0.722 42             | 0.711 4           | 0.601 25             |
| 35       | LxSem   | Psycholinguistic                     | as_AAKuW_C               | AoA of words per Sent  | 0.537 28          | 0.502 2           | 0.722 41             | 0.711 6           | 0.602 24             |
| 33       | ShaTr   | Shallow                              | as_Chara_C               | # characters per Sent  | 0.539 27          | 0.487 4           | 0.696 46             | 0.711 5           | 0.613 20             |
| 33       | Synta   | Tree Structure                       | as_FTree_C               | length of flattened Trees per Sent                             | 0.505 37          | 0.485 5           | 0.677 54             | 0.719 2           | 0.622 16             |
| 32       | LxSem   | Psycholinguistic                     | at_AAKuL_C               | lemmas AoA of lemmas per Word                                  | 0.723 2           | 0.323 35          | 0.785 19             | 0.65 20           | 0.453 67             |
| 32       | LxSem   | Psycholinguistic                     | at_AAKuW_C               | AoA of words per Word  | 0.703 5           | 0.308 36          | 0.784 20             | 0.643 21          | 0.455 66             |
| 31       | Synta   | Phrasal                              | as_NoPhr_C               | # Noun phrases per Sent  | 0.55 20           | 0.406 25          | 0.66 58              | 0.673 18          | 0.582 35             |
| 31       |         | Shallow                              | as_Sylla_C               | # syllables per Sent   | 0.541 24          | 0.461 10          | 0.686 50             | 0.697 11          | 0.59 31              |
| 31       | Synta   | Part-of-Speech                       | as_ContW_C               | # Content words per Sent                                       | 0.534 29          | 0.453 13          | 0.667 56             | 0.688 14          | 0.544 43             |
| 31       | Synta   |                                      | as_PrPhr_C               | # prepositional phrases per Sent                               | 0.513 35          | 0.417 23          | 0.607 70             | 0.608 28          | 0.59 32              |
| 31       |         | Shallow                              | as_Token_C               | # tokens per Sent  | 0.494 40          | 0.464 9           | 0.65 60              | 0.709 7           | 0.58 36              |
| 31       |         | Part-of-Speech                       | as_FuncW_C               | # Function words per Sent                                      | 0.468 48          | 0.471 8           | 0.662 57             | 0.673 17          | 0.614 19             |
| 31       |         | Psycholinguistic                     | to_AAKuL_C               | total lemmas AoA of lemmas                                     | 0.428 71          | 0.189 59          | 0.835 3              | 0.627 22          | 0.716 5              |
| 31       |         | Psycholinguistic                     | to_AAKuW_C               |  | 0.427 72          | 0.189 60          | 0.835 4              | 0.625 23          | 0.715 6              |
| 30       |         | Type Token Ratio                     | _                        | Corrected TTR  | 0.745 1           | 0.006 228         | 0.846 1              | 0.445 65          | 0.692 7              |
| 30       |         | Variation Ratio                      | CorrNoV_S                | Corrected Noun Variation-1                                     | 0.717 3           | 0.086 131         | 0.842 2              | 0.406 78          | 0.612 21             |
| 30       |         | Tree Structure                       | as_TreeH_C               | Tree height per Sent   | 0.55 21           | 0.341 30          | 0.686 51             | 0.699 9           | 0.541 44             |
| 30       | Synta   |                                      | to_PrPhr_C               | total # prepositional phrases                                  | 0.47 47           | 0.189 58          | 0.808 11             | 0.58 36           | 0.729 3              |
| 30       |         | Word Familiarity                     | as_SbL1C_C               | SubtlexUS Lg10CD value per Sent                                | 0.467 49          | 0.43 20           | 0.612 69             | 0.699 10          | 0.533 45             |
| 30       |         |                                      | as_SbL1W_C               | SubtlexUS Lg10WF value per Sent                                | 0.462 52          | 0.437 19          | 0.605 71             | 0.693 12          | 0.523 48             |
| 29       |         | Variation Ratio                      | SquaNoV_S                | Squared Noun Variation-1                                       | 0.645 9           | 0.124 109         | 0.815 7              | 0.401 84          | 0.583 34             |
| 29       |         | Variation Ratio                      | CorrVeV_S                | Corrected Verb Variation-1                                     | 0.602 11          | 0.058 155         | 0.801 15             | 0.393 86          | 0.737 2              |
| 29       |         | Part-of-Speech                       | as_NoTag_C               | # Noun POS tags per Sent                                       | 0.551 19          | 0.304 38          | 0.624 65             | 0.608 29          | 0.48 61              |
| 29       |         | Psycholinguistic                     | to_AABrL_C               | total lemmas AoA of lemmas, Bristol norm                       | 0.451 56          | 0.134 100         | 0.808 10             | 0.561 38          | 0.637 12             |
| 29       |         | Psycholinguistic                     | to_AACoL_C               | total AoA of lemmas, Cortese and Khanna norm                   | 0.451 57          | 0.134 101         | 0.808 9              | 0.561 39          | 0.637 13             |
| 29       |         | Part-of-Speech                       | to_NoTag_C               | total # Noun POS tags  | 0.441 61          | 0.129 107         | 0.805 13             | 0.55 44           | 0.636 15             |
| 29       | Synta   |                                      | to_NoPhr_C               | total # Noun phrases   | 0.416 76          | 0.148 84          | 0.809 8              | 0.527 52          | 0.659 9              |
| 29       |         | Part-of-Speech                       | to_ContW_C               | total # Content words  | 0.402 81          | 0.163 71          | 0.804 14             | 0.558 40          | 0.654 11             |
| 28       |         | Psycholinguistic                     | as_AACoL_C               | AoA of lemmas, Cortese and Khanna norm per Sent                | 0.532 30          | 0.339 32          | 0.649 61             | 0.597 32          | 0.499 58             |
| 28<br>28 |         | Psycholinguistic                     | as_AABrL_C               | lemmas AoA of lemmas, Bristol norm per Sent                    | 0.532 31 0.459 55 | 0.339 31 0.458 11 | 0.649 62             | 0.597 31 0.653 19 | 0.499 57<br>0.443 69 |
| 28<br>28 |         | Psycholinguistic<br>Word Familiarity | as_AABiL_C<br>as SbCDL C | lemmas AoA of lemmas, Bird norm per Sent                       | 0.439 55          | 0.458 11          | 0.582 73<br>0.527 82 | 0.633 19          | 0.443 69             |
| 28<br>28 |         | Word Familiarity                     |                          | SubtlexUS CDIow value per Sent<br>SubtlexUS CD# value per Sent | 0.432 63          | 0.441 14          | 0.527 82             | 0.623 26          | 0.401 83             |
| 28       |         | Word Familiarity                     |                          | SubtlexUS SUBTLCD value per Sent                               | 0.431 67          | 0.437 17          | 0.525 84             | 0.624 24          | 0.404 82             |
| 28<br>28 |         | Part-of-Speech                       | as_VeTag_C               | # Verb POS tags per Sent                                       | 0.431 08          | 0.437 18          | 0.525 85             | 0.588 34          | 0.404 83             |
| 28<br>28 |         | Tree Structure                       | to_FTree_C               | total length of flattened Trees                                | 0.428 70          | 0.470 0           | 0.378 74 0.805 12    | 0.538 54          | 0.505 55             |
| 28<br>27 |         | Variation Ratio                      | SquaVeV_S                | Squared Verb Variation-1                                       | 0.559 17          | 0.100 09          | 0.803 12             | 0.338 49          | 0.076 8              |
| 27       |         | Variation Ratio                      | SquaVeV_S                | Squared Adjective Variation-1                                  | 0.531 32          | 0.141 94          | 0.754 34             | 0.407 77          | 0.573 37             |
| 27       |         | Part-of-Speech                       | as_AjTag_C               | # Adjective POS tags per Sent                                  | 0.506 36          | 0.353 28          | 0.553 76             | 0.533 51          | 0.404 84             |
| 27       |         | Word Familiarity                     | as_SbFrL_C               | SubtlexUS FREQlow value per Sent                               | 0.443 60          | 0.335 20          | 0.52 86              | 0.552 42          | 0.425 77             |
| 27       |         | Part-of-Speech                       | to_AjTag_C               | total # Adjective POS tags                                     | 0.441 62          | 0.120 21          | 0.777 23             | 0.504 54          | 0.525 46             |
| 27       |         | •                                    |                          | SubtlexUS SUBTLWF value per Sent                               | 0.44 63           | 0.441 15          | 0.509 91             | 0.542 48          | 0.425 76             |
| 27       |         | Word Familiarity                     |                          | SubtlexUS FREQ# value per Sent                                 | 0.44 64           | 0.441 16          | 0.509 90             | 0.542 47          | 0.425 75             |
| 27       | Synta   |                                      | as_VePhr_C               | # Verb phrases per Sent  | 0.383 90          | 0.455 12          | 0.59 72              | 0.586 35          | 0.505 54             |
| 26       |         | Shallow                              | at_Sylla_C               | # syllables per Word   | 0.66 7            | 0.106 120         | 0.627 64             | 0.505 53          | 0.37 91              |
| 26       |         | Variation Ratio                      | CorrAjV_S                | Corrected Adjective Variation-1                                | 0.591 12          | 0.078 134         | 0.779 21             | 0.422 70          | 0.584 33             |
| 26       |         | Entity Grid                          | ra_NNTo_C                | ratio of nn transitions to total                               | 0.476 44          | 0.078 135         | 0.754 35             | 0.451 64          | 0.602 23             |
| 26       |         | Tree Structure                       | at_TreeH_C               | Tree height per Word   | 0.476 45          | 0.419 22          | 0.416 104            | 0.597 33          | 0.41 81              |
| 26       |         | Word Familiarity                     |                          | total SubtlexUS Lg10CD value                                   | 0.37 93           | 0.14 95           | 0.797 16             | 0.491 56          | 0.621 17             |
| 26       |         | Word Familiarity                     |                          | total SubtlexUS Lg10WF value                                   | 0.365 99          | 0.144 93          | 0.795 17             | 0.477 58          | 0.611 22             |
| 26       |         | Word Familiarity                     |                          | total SubtlexUS FREQIow value                                  | 0.348 109         |                   | 0.774 24             | 0.414 74          | 0.555 40             |
| 26       |         | Word Familiarity                     |                          | total SubtlexUS SUBTLWF value                                  | 0.34 115          |                   | 0.77 27              | 0.403 81          | 0.551 42             |
| 26       |         | Word Familiarity                     |                          | total SubtlexUS FREQ# value                                    | 0.34 116          |                   | 0.77 26              | 0.403 82          | 0.551 41             |
|          | Labelli |                                      |                          |  | 5.5. 110          | 1 3.200 10        | 1 5.1.7 20           | 1 01100 02        |                      |

Table 13: Part A. The full generalizability ranking of handcrafted linguistic features under Approach A. r: Pearson's correlation between the feature and the dataset. rk: the feature's correlation ranking on the specific dataset.

|          |        |                                    | ]                        | Feature   | ССВ                    | WBT                    | CAM                    | СКС                    | OSE                    |
|----------|--------|------------------------------------|--------------------------|---|------------------------|------------------------|------------------------|------------------------|------------------------|
| Score    | Branch | Subgroup                           | LingFeat Code            | Brief Explanation   | r rk                   |
| 25       | ShaTr  | Shallow                            | at_Chara_C               | # characters per Word   | 0.443 59               | 0.2 52                 | 0.619 67               | 0.402 83               | 0.443 68               |
| 25       |        | Phrasal                            | to_SuPhr_C               | total # Subordinate Clauses   | 0.367 96               | 0.202 50               | 0.721 43               | 0.462 61               | 0.419 78               |
| 25       |        | Psycholinguistic                   | to_AABiL_C               | total lemmas AoA of lemmas, Bird norm   | 0.365 98<br>0.364 101  | 0.155 79<br>0.268 43   | 0.786 18<br>0.728 39   | 0.473 59<br>0.406 80   | 0.565 39<br>0.434 72   |
| 25<br>25 |        | Part-of-Speech<br>Part-of-Speech   | to_CoTag_C<br>to_FuncW_C | total # Coordinating Conjunction POS tags<br>total # Function words                                   | 0.33 126               | 0.208 43               | 0.728 39               | 0.385 89               | 0.434 72               |
| 25       |        | Phrasal                            | to_VePhr_C               | total # Verb phrases  | 0.324 127              | 0.169 68               | 0.76 31                | 0.416 72               | 0.57 38                |
| 24       | LxSem  | Variation Ratio                    | CorrAvV_S                | Corrected AdVerb Variation-1  | 0.542 23               | 0.059 154              | 0.71 44                | 0.333 99               | 0.474 63               |
| 24       |        | Word Familiarity                   |                          | total SubtlexUS CDlow value   | 0.348 107              | 0.148 87               | 0.764 30               | 0.394 85               | 0.513 53               |
| 24       |        | Word Familiarity                   |                          | total SubtlexUS CD# value   | 0.347 110              | 0.146 90               | 0.764 29               | 0.392 87               | 0.515 51               |
| 24<br>23 |        | Word Familiarity<br>Wiki Knowledge |                          | total SubtlexUS SUBTLCD value<br>Number of topics, 200 topics extracted from Wikipedia                | 0.347 111<br>0.584 14  | 0.146 91<br>0.015 214  | 0.764 28<br>0.616 68   | 0.392 88 0.617 27      | 0.515 52 0.137 138     |
| 23       |        | Wiki Knowledge                     | -                        | Number of topics, 200 topics extracted from Wikipedia   | 0.58 15                | 0.007 227              | 0.645 63               | 0.605 30               | 0.191 122              |
| 23       |        | Variation Ratio                    | SquaAvV_S                | Squared AdVerb Variation-1  | 0.515 34               | 0.093 128              | 0.686 52               | 0.326 102              | 0.46 65                |
| 23       | -      | Part-of-Speech                     | to_AvTag_C               | total # Adverb POS tags   | 0.342 114              | 0.17 67                | 0.726 40               | 0.352 96               | 0.469 64               |
| 23       | -      | Part-of-Speech                     | as_AvTag_C               | # Adverb POS tags per Sent  | 0.32 129               | 0.292 41               | 0.526 83               | 0.43 67                | 0.415 79               |
| 23<br>22 | -      | Part-of-Speech<br>Phrasal          | to_VeTag_C<br>as_SuPhr_C | total # Verb POS tags<br># Subardineta Clauses per Sont   | 0.288 138<br>0.387 89  | 0.173 63<br>0.357 26   | 0.738 38<br>0.532 80   | 0.383 91<br>0.495 55   | 0.597 27<br>0.265 112  |
| 22       |        | Part-of-Speech                     | as_CoTag_C               | <ul><li># Subordinate Clauses per Sent</li><li># Coordinating Conjunction POS tags per Sent</li></ul> | 0.38 91                | 0.337 20               | 0.332 80               | 0.493 55               | 0.203 112 0.293 107    |
| 22       |        | Phrasal                            | to_AvPhr_C               | total # Adverb phrases  | 0.356 105              | 0.17 66                | 0.705 45               | 0.298 111              | 0.432 73               |
| 22       | -      | Tree Structure                     | to_TreeH_C               | total Tree height of all sentences  | 0.27 145               | 0.069 143              | 0.755 33               | 0.309 108              | 0.515 50               |
| 21       |        | Entity Grid                        | ra_NSTo_C                | ratio of ns transitions to total  | 0.426 73               | 0.033 187              | 0.516 87               | 0.266 117              | 0.505 56               |
| 21       |        | Part-of-Speech                     | to_SuTag_C               | total # Subordinating Conjunction POS tags  | 0.4 83                 | 0.193 56               | 0.691 48               | 0.406 79               | 0.299 106              |
| 20<br>20 |        | Type Token Ratio                   |                          | Uber Index<br># propositional phrases per Word  | 0.646 8<br>0.57 16     | 0.041 174<br>0.133 103 | 0.369 112<br>0.316 124 | 0.109 173<br>0.323 105 | 0.599 26<br>0.366 92   |
| 20       |        | Phrasal<br>Wiki Knowledge          | at_PrPhr_C<br>WTopc05_S  | # prepositional phrases per Word<br>Number of topics, 50 topics extracted from Wiki                   | 0.57 10                | 0.033 186              | 0.510 124              | 0.523 105              | 0.300 92 0.042 203     |
| 20       |        | Wiki Knowledge                     |                          | Number of topics, 100 topics extracted from Wiki  | 0.52 33                | 0.004 229              | 0.532 79               | 0.552 43               | 0.075 180              |
| 20       |        | Entity Grid                        | ra_SNTo_C                | ratio of sn transitions to total  | 0.448 58               | 0.019 210              | 0.514 88               | 0.196 133              | 0.518 49               |
| 20       |        | Word Familiarity                   | at_SbL1C_C               | SubtlexUS Lg10CD value per Word   | 0.408 78               | 0.161 75               | 0.541 78               | 0.204 130              | 0.392 86               |
| 20       |        | Entity Grid                        | ra_XNTo_C                | ratio of xn transitions to total  | 0.339 119              | 0.103 124              | 0.658 59               | 0.327 100              | 0.29 108               |
| 20<br>20 |        | Phrasal<br>Phrasal                 | to_AjPhr_C               | total # Adjective phrases   | 0.339 120<br>0.244 152 | 0.182 62<br>0.328 34   | 0.682 53<br>0.427 103  | 0.327 101<br>0.38 92   | 0.271 111 0.356 93     |
| 20       | 2      | Shallow                            | as_AvPhr_C<br>TokSenS_S  | # Adverb phrases per Sent<br>sqrt(total # tokens x total # sentence)                                  | 0.244 132 0.241 154    | 0.328 34 0.064 147     | 0.427 103              | 0.38 92 0.249 121      | 0.330 93               |
| 19       |        | Wiki Knowledge                     |                          | Semantic Noise, 200 topics extracted from Wiki  | 0.492 41               | 0.032 190              | 0.566 75               | 0.572 37               | 0.025 221              |
| 19       |        | Phrasal                            | ra_NoPrP_C               | ratio of Noun phrases # to Prep phrases #   | 0.477 43               | 0.149 83               | 0.34 120               | 0.345 97               | 0.389 87               |
| 19       |        | Word Familiarity                   |                          | SubtlexUS Lg10WF value per Word   | 0.399 84               | 0.089 129              | 0.531 81               | 0.24 123               | 0.412 80               |
| 19       |        | Word Familiarity                   |                          | SubtlexUS SUBTLCD value per Word  | 0.37 94                | 0.032 192              | 0.492 93               | 0.324 103              | 0.435 71               |
| 19<br>19 |        | Word Familiarity<br>Part-of-Speech | as_SuTag_C               | SubtlexUS CD# value per Word<br># Subordinating Conjunction POS tags per Sent                         | 0.37 95<br>0.366 97    | 0.032 191<br>0.295 39  | 0.492 94<br>0.407 105  | 0.324 104<br>0.427 68  | 0.435 70<br>0.151 131  |
| 19       | -      | Word Familiarity                   |                          | SubtlexUS CDlow value per Word  | 0.362 102              | 0.047 166              | 0.474 96               | 0.31 107               | 0.431 74               |
| 18       |        | Wiki Knowledge                     |                          | Semantic Richness, 150 topics extracted from Wiki   | 0.495 39               | 0.02 208               | 0.48 95                | 0.549 45               | 0.037 209              |
| 18       | AdSem  | Wiki Knowledge                     | WRich20_S                | Semantic Richness, 200 topics extracted from Wiki   | 0.465 50               | 0.029 195              | 0.446 102              | 0.556 41               | 0.027 219              |
| 18       |        | Wiki Knowledge                     |                          | Semantic Noise, 50 topics extracted from Wiki   | 0.462 53               | 0.061 150              | 0.455 100              | 0.412 75               | 0.118 151              |
| 18<br>18 |        | Phrasal<br>Entity Grid             | ra_PrNoP_C<br>ra_NXTo_C  | ratio of Prep phrases # to Noun phrases #<br>ratio of nx transitions to total                         | 0.421 75<br>0.339 118  | 0.162 74 0.097 127     | 0.276 135<br>0.62 66   | 0.344 98<br>0.28 116   | 0.37 90<br>0.278 110   |
| 18       |        | Shallow                            | TokSenL_S                | log(total # tokens)/log(total # sentence)   | 0.293 137              | 0.352 29               | 0.02 00 0.297 130      | 0.28 110               | 0.198 121              |
| 18       |        | Shallow                            | TokSenM_S                | total # tokens x total # sentence   | 0.189 173              | 0.112 116              | 0.674 55               | 0.177 140              | 0.486 60               |
| 17       | Synta  | Phrasal                            | as_AjPhr_C               | # Adjective phrases per Sent  | 0.323 128              | 0.239 46               | 0.387 106              | 0.357 95               | 0.157 127              |
| 17       |        | Entity Density                     | at_UEnti_C               | number of unique Entities per Word  | 0.127 197              | 0.307 37               | 0.548 77               | 0.253 119              | 0.124 149              |
| 16       |        | Phrasal<br>Wiki Knowledge          | ra_VePrP_C               | ratio of Verb phrases # to Prep phrases #<br>Semantic Noise, 150 topics extracted from Wiki           | 0.475 46<br>0.388 88   | 0.018 211<br>0.033 188 | 0.301 127<br>0.454 101 | 0.255 118<br>0.454 63  | 0.249 114<br>0.006 226 |
|          |        | Psycholinguistic                   |                          |   | 0.333 124              |                        |                        |                        |                        |
|          |        |                                    |                          | AoA of lemmas, Cortese and Khanna norm per Word   | 0.333 125              | 0.029 193              | 0.462 99               | 0.284 114              | 0.217 119              |
| 16       | Disco  | Entity Density                     | at_EntiM_C               | number of Entities Mentions #s per Word   | 0.17 180               | 0.204 49               | 0.501 92               | 0.292 112              | 0.127 146              |
| 16       |        | WB Knowledge                       |                          | Semantic Clarity, 150 topics extracted from WeeBit  | 0.025 221              | 0.161 76               | 0.38 108               | 0.481 57               | 0.315 100              |
| 15       |        | Type Token Ratio                   |                          | Bi-Logarithmic TTR  | 0.591 13               | 0.062 149              | 0.07 200               | 0.001 229              | 0.523 47               |
| 15<br>15 |        | Wiki Knowledge<br>Type Token Ratio |                          | Semantic Richness, 50 topics extracted from Wiki<br>TTR   | 0.405 80<br>0.375 92   | 0.063 148 0.025 200    | 0.347 117<br>0.367 113 | 0.301 110<br>0.163 147 | 0.035 211<br>0.344 94  |
| 15       |        | Wiki Knowledge                     | -                        | Semantic Richness, 100 topics extracted from Wiki   | 0.373 92               | 0.023 200              | 0.33 123               | 0.103 147              | 0.041 206              |
| 15       |        | Wiki Knowledge                     |                          | Semantic Noise, 100 topics extracted from Wiki  | 0.34 117               | 0.021 207              | 0.376 109              | 0.426 69               | 0.03 216               |
| 15       |        | Entity Density                     | to_UEnti_C               | total number of unique Entities   | 0.308 134              | 0.132 105              | 0.3 128                | 0.023 216              | 0.31 102               |
| 15       |        | WB Knowledge                       | BClar20_S                | Semantic Clarity, 200 topics extracted from WeeBit  | 0.004 227              | 0.147 88               | 0.3 129                | 0.462 60               | 0.308 104              |
| 14       |        | Entity Grid                        | ra_NOTo_C                | ratio of no transitions to total  | 0.348 108              | 0.022 204              | 0.383 107              | 0.056 200              | 0.378 89               |
| 14<br>13 | Synta  | Phrasal<br>Phrasal                 | ra_SuVeP_C<br>ra_PrVeP_C | ratio of Subordinate Clauses # to Verb phrases #<br>ratio of Prep phrases # to Verb phrases #         | 0.345 113<br>0.485 42  | 0.052 160<br>0.055 157 | 0.343 118<br>0.184 158 | 0.376 93<br>0.189 136  | 0.083 172<br>0.219 117 |
| 13       |        | Variation Ratio                    | SimpAjV_S                | Adjective Variation-1   | 0.398 85               | 0.109 118              | 0.279 134              | 0.073 192              | 0.201 120              |
| 13       |        | Phrasal                            | ra_VeSuP_C               | ratio of Verb phrases # to Subordinate Clauses #  | 0.349 106              | 0.137 98               | 0.307 126              | 0.127 167              | 0.043 200              |
| 13       | -      | Part-of-Speech                     | at_NoTag_C               | # Noun POS tags per Word  | 0.347 112              | 0.104 122              | 0.295 131              | 0.148 154              | 0.107 159              |
| 13       |        | Entity Grid                        | ra_ONTo_C                | ratio of on transitions to total  | 0.333 123              | 0.04 178               | 0.288 133              | 0.06 199               | 0.383 88               |
| 13<br>13 | -      | Phrasal<br>Psycholinguistic        | at_SuPhr_C<br>at_AABiL_C | # Subordinate Clauses per Word<br>lemmas AoA of lemmas, Bird norm per Word                            | 0.204 165<br>0.203 166 | 0.157 78<br>0.11 117   | 0.246 140<br>0.266 138 | 0.314 106<br>0.053 202 | 0.073 182 0.302 105    |
| 13       |        | WB Knowledge                       | at_AABIL_C<br>BTopc10_S  | Number of topics, 100 topics extracted from WeeBit  | 0.203 100              | 0.038 179              | 0.260 138              | 0.166 145              | 0.323 99               |
| 13       |        | WB Knowledge                       | BNois10_S                | Semantic Noise, 100 topics extracted from WeeBit  | 0.193 171              | 0.036 180              | 0.37 110               | 0.161 149              | 0.33 97                |
| 13       |        | WB Knowledge                       | BNois05_S                | Semantic Noise, 50 topics extracted from WeeBit   | 0.158 184              | 0.011 219              | 0.351 116              | 0.15 153               | 0.325 98               |
| 13       | AdSem  | WB Knowledge                       | BTopc15_S                | Number of topics, 150 topics extracted from WeeBit  | 0.133 193              | 0.146 92               | 0.209 151              | 0.416 73               | 0.03 217               |

Table 14: Part B. The full generalizability ranking of handcrafted linguistic features under Approach B.

|          |        |                                    |                          | Feature   | ССВ                    | WBT                    | CAM                    | СКС                 | OSE                    |
|----------|--------|------------------------------------|--------------------------|---|------------------------|------------------------|------------------------|---------------------|------------------------|
| Score    | Branch | Subgroup                           | LingFeat Code            | Brief Explanation   | r rk                   | r rk                   | r rk                   | r rk                | r rk                   |
| 12       |        | Variation Ratio                    | SimpNoV_S                | Noun Variation-1  | 0.499 38               | 0.087 130              | 0.038 212              | 0.031 213           | 0.337 95               |
| 12<br>12 | •      | Part-of-Speech<br>Phrasal          | ra_NoVeT_C<br>ra_AvPrP_C | ratio of Noun POS # to Verb POS #<br>ratio of Adv phrases # to Prep phrases #                             | 0.432 66<br>0.4 82     | 0.118 111<br>0.014 217 | 0.149 168<br>0.222 147 | 0.112 170 0.196 135 | 0.051 197<br>0.115 152 |
| 12       |        | Part-of-Speech                     | ra_VeNoT_C               | ratio of Verb POS # to Noun POS #   | 0.4 82                 | 0.198 53               | 0.222 147 0.234 142    | 0.170 133           | 0.067 186              |
|          | -      | Part-of-Speech                     | ra_SuVeT_C               | ratio of Subordinating Conjunction POS # to Verb POS #  | 0.36 103               | 0.053 159              | 0.109 177              | 0.282 115           | 0.137 139              |
| 12       |        | Entity Density                     | as_UEnti_C               | number of unique Entities per Sent  | 0.337 121              | 0.114 113              | 0.273 136              | 0.066 196           | 0.157 128              |
| 12       | -      | Part-of-Speech                     | at_AjTag_C               | # Adjective POS tags per Word   | 0.334 122              | 0.117 112              | 0.216 149              | 0.197 132           | 0.037 210              |
|          | •      | Phrasal<br>Part-of-Speech          | ra_SuAvP_C<br>at_FuncW_C | ratio of Subordinate Clauses # to Adv phrases #<br># Function words per Word                              | 0.309 133 0.28 142     | 0.008 226<br>0.04 175  | 0.141 170<br>0.181 159 | 0.241 122 0.461 62  | 0.111 153<br>0.032 215 |
|          |        | WB Knowledge                       | BTopc20_S                | Number of topics, 200 topics extracted from WeeBit  | 0.23 142               | 0.135 99               | 0.025 215              | 0.401 02            | 0.032 213              |
|          |        | Wiki Knowledge                     | WClar05_S                | Semantic Clarity, 50 topics extracted from Wiki   | 0.212 164              | 0.014 218              | 0.214 150              | 0.235 124           | 0.102 161              |
|          |        | WB Knowledge                       | BRich10_S                | Semantic Richness, 100 topics extracted from WeeBit   | 0.196 170              | 0.044 171              | 0.369 111              | 0.035 210           | 0.336 96               |
|          |        | WB Knowledge                       | BClar05_S                | Semantic Clarity, 50 topics extracted from WeeBit   | 0.14 190               | 0.041 173              | 0.339 121              | 0.164 146           | 0.289 109              |
| 12<br>11 |        | Entity Density<br>Phrasal          | to_EntiM_C<br>ra_VeNoP_C | total number of Entities Mentions<br>ratio of Verb phrases # to Noun phrases #                            | 0.139 191<br>0.46 54   | 0.02 209<br>0.164 70   | 0.335 122<br>0.124 174 | 0.0 230 0.041 209   | 0.312 101<br>0.027 220 |
| 11       |        | Part-of-Speech                     | at_VeTag_C               | # Verb POS tags per Word  | 0.40 54 0.431 69       | 0.187 61               | 0.076 196              | 0.111 171           | 0.027 220              |
| 11       |        | Part-of-Speech                     | ra_AjVeT_C               | ratio of Adjective POS # to Verb POS #  | 0.411 77               | 0.034 185              | 0.133 172              | 0.156 150           | 0.005 227              |
| 11       | •      | Part-of-Speech                     | ra_SuAvT_C               | ratio of Subordinating Conjunction POS # to Adverb POS #  | 0.314 131              | 0.021 206              | 0.106 178              | 0.148 156           | 0.18 124               |
|          |        | Variation Ratio                    | SimpVeV_S                | Verb Variation-1  | 0.286 139              | 0.048 165              | 0.081 193              | 0.003 226           | 0.48 62                |
| 11<br>11 | -      | Part-of-Speech<br>Part-of-Speech   | ra_CoFuW_C<br>at_SuTag_C | ratio of Content words to Function words<br># Subordinating Conjunction POS tags per Word                 | 0.284 141<br>0.259 148 | 0.023 203 0.13 106     | 0.2 154<br>0.085 192   | 0.376 94 0.252 120  | 0.042 201<br>0.135 141 |
| 11       | •      | Wiki Knowledge                     | WClar20_S                | Semantic Clarity, 200 topics extracted from Wikipedia   | 0.144 187              | 0.016 212              | 0.308 125              | 0.232 120           | 0.034 213              |
|          |        | WB Knowledge                       | BTopc05_S                | Number of topics, 50 topics extracted from WeeBit   | 0.139 192              | 0.009 224              | 0.291 132              | 0.144 160           | 0.222 116              |
| 11       | AdSem  | WB Knowledge                       | BRich05_S                | Semantic Richness, 50 topics extracted from WeeBit  | 0.126 198              | 0.051 162              | 0.24 141               | 0.051 207           | 0.309 103              |
| 11       | Synta  |                                    | ra_SuNoP_C               | ratio of Subordinate Clauses # to Noun phrases #  | 0.081 203              | 0.163 72               | 0.224 146              | 0.307 109           | 0.086 170              |
| 11<br>10 |        | WB Knowledge<br>Part-of-Speech     | BNois15_S<br>ra_CoVeT_C  | Semantic Noise, 150 topics extracted from WeeBit<br>ratio of Coordinating Conjunction POS # to Verb POS # | 0.035 214 0.302 135    | 0.162 73<br>0.133 104  | 0.341 119<br>0.023 218 | 0.221 127 0.133 164 | 0.091 164<br>0.088 168 |
| 10       | •      | Phrasal                            | ra_AvSuP_C               | ratio of Adv phrases # to Subordinate Clauses #   | 0.299 136              | 0.06 151               | 0.025 218              | 0.133 104           | 0.088 108              |
| 10       | -      | Tree Structure                     | at_FTree_C               | length of flattened Trees per Word  | 0.28 143               | 0.14 96                | 0.097 182              | 0.1 177             | 0.152 130              |
| 10       |        | Entity Density                     | as_EntiM_C               | number of Entities Mentions #s per Sent   | 0.242 153              | 0.015 215              | 0.219 148              | 0.051 206           | 0.168 125              |
| 10       |        | Entity Grid                        | _                        | Local Coherence distance for PW score   | 0.239 155              | 0.002 230              | 0.195 156              | 0.143 161           | 0.141 136              |
| 10<br>10 |        | Entity Grid<br>Part-of-Speech      | LoCoDPA_S                | Local Coherence distance for PA score   | 0.239 156 0.218 161    | 0.002 231<br>0.267 44  | 0.195 157<br>0.02 220  | 0.143 162 0.111 172 | 0.141 135<br>0.087 169 |
| 10       |        | Variation Ratio                    | at_CoTag_C<br>SimpAvV_S  | # Coordinating Conjunction POS tags per Word<br>AdVerb Variation-1  | 0.218 101 0.214 163    | 0.207 44 0.098 126     | 0.02 220               | 0.021 221           | 0.087 109              |
| 10       | Synta  |                                    | ra_AjPrP_C               | ratio of Adj phrases # to Prep phrases #  | 0.201 168              | 0.036 181              | 0.155 164              | 0.095 178           | 0.252 113              |
| 10       | AdSem  | WB Knowledge                       | BNois20_S                | Semantic Noise, 200 topics extracted from WeeBit  | 0.129 196              | 0.148 85               | 0.202 153              | 0.167 144           | 0.032 214              |
| 10       |        | WB Knowledge                       | BRich20_S                | Semantic Richness, 200 topics extracted from WeeBit   | 0.047 211              | 0.104 121              | 0.112 176              | 0.221 126           | 0.143 134              |
| 9<br>9   | •      | Phrasal<br>Phrasal                 | at_NoPhr_C<br>ra_NoVeP_C | # Noun phrases per Word<br>ratio of Noun phrases # to Verb phrases #                                      | 0.424 74<br>0.406 79   | 0.066 146<br>0.068 145 | 0.089 188<br>0.069 201 | 0.005 224 0.031 212 | 0.042 202<br>0.019 223 |
| 9        |        | Phrasal                            | ra_PrAvP_C               | ratio of Prop phrases # to Adv phrases #  | 0.32 130               | 0.003 145              | 0.021 219              | 0.176 141           | 0.071 183              |
| 9        | -      | Phrasal                            | at_VePhr_C               | # Verb phrases per Word   | 0.31 132               | 0.138 97               | 0.079 194              | 0.032 211           | 0.009 225              |
| 9        |        | Part-of-Speech                     | ra_CoAvT_C               | ratio of Coordinating Conjunction POS # to Adverb POS #   | 0.284 140              | 0.04 176               | 0.16 162               | 0.079 189           | 0.119 150              |
| 9        |        | Part-of-Speech                     | ra_NoAvT_C               | ratio of Noun POS # to Adverb POS #   | 0.261 147              | 0.133 102              | 0.101 180              | 0.052 204           | 0.034 212              |
| 9<br>9   |        | Entity Grid<br>Entity Grid         | LoCohPW_S<br>LoCohPA_S   | Local Coherence for PW score<br>Local Coherence for PA score  | 0.229 159<br>0.229 160 | 0.034 183<br>0.034 184 | 0.012 227<br>0.012 226 | 0.146 157 0.146 158 | 0.148 133<br>0.148 132 |
| 9        |        | Phrasal                            | ra_SuPrP_C               | ratio of Subordinate Clauses # to Prep phrases #  | 0.218 162              | 0.048 164              | 0.012 220              | 0.07 194            | 0.227 115              |
| 9        |        | Part-of-Speech                     | ra_VeAjT_C               | ratio of Verb POS # to Adjective POS #  | 0.177 177              | 0.059 153              | 0.203 152              | 0.162 148           | 0.042 204              |
| 9        |        | Part-of-Speech                     | at_ContW_C               | # Content words per Word  | 0.161 183              | 0.057 156              | 0.23 143               | 0.183 139           | 0.055 193              |
| 9        |        | Phrasal                            | ra_NoSuP_C               | ratio of Noun phrases # to Subordinate Clauses #  | 0.157 185              | 0.153 80               | 0.228 145              | 0.052 205           | 0.04 207               |
| 9<br>9   | -      | Phrasal<br>Part-of-Speech          | ra_PrAjP_C<br>ra_NoAjT_C | ratio of Prep phrases # to Adj phrases #<br>ratio of Noun POS # to Adjective POS #                        | 0.142 189<br>0.121 199 | 0.035 182<br>0.152 81  | 0.017 223<br>0.125 173 | 0.207 128 0.114 169 | 0.136 140<br>0.004 228 |
| 9        |        | WB Knowledge                       | BClar10_S                | Semantic Clarity, 100 topics extracted from WeeBit  | 0.079 204              | 0.015 216              | 0.125 175 0.269 137    | 0.148 155           | 0.181 123              |
| 9        |        | Part-of-Speech                     | _                        | ratio of Coordinating Conjunction POS # to Noun POS #   | 0.02 224               | 0.277 42               | 0.159 163              | 0.013 222           | 0.132 142              |
| 8        |        | Part-of-Speech                     | ra_AjAvT_C               | ratio of Adjective POS # to Adverb POS #  | 0.357 104              | 0.042 172              | 0.056 204              | 0.091 180           | 0.044 199              |
| 8        |        | Part-of-Speech                     | ra_SuCoT_C               | ratio of Subordinating Conj POS # to Coordinating Conj #  | 0.274 144              | 0.054 158              | 0.019 222              | 0.143 163           | 0.077 179              |
| 8<br>8   |        | Part-of-Speech<br>Phrasal          | ra_VeSuT_C<br>ra_AvNoP_C | ratio of Verb POS # to Subordinating Conjunction #<br>ratio of Adv phrases # to Noun phrases #            | 0.266 146 0.257 149    | 0.046 169<br>0.128 108 | 0.09 186<br>0.072 199  | 0.105 175 0.044 208 | 0.065 188<br>0.051 196 |
| 8        |        | Part-of-Speech                     | ra_SuAjT_C               | ratio of Subordinating Conjunction POS # to Adjective POS #   | 0.237 149              | 0.008 225              | 0.072 199              | 0.044 208           | 0.138 137              |
| 8        |        | Phrasal                            | ra_NoAvP_C               | ratio of Noun phrases # to Adv phrases #  | 0.235 157              | 0.102 125              | 0.09 187               | 0.082 188           | 0.071 185              |
| 8        | -      | Phrasal                            | ra_AjAvP_C               | ratio of Adj phrases # to Adv phrases #   | 0.232 158              | 0.016 213              | 0.046 209              | 0.094 179           | 0.156 129              |
| 8        |        | Part-of-Speech                     | ra_AvSuT_C               | ratio of Adverb POS # to Subordinating Conjunction #  | 0.202 167              | 0.024 201              | 0.003 230              | 0.114 168           | 0.067 187              |
| 8<br>8   |        | Part-of-Speech<br>Word Familiarity | ra_VeCoT_C<br>at_SbFrO_C | ratio of Verb POS # to Coordinating Conjunction #<br>SubtlexUS FREQ# value per Word                       | 0.192 172<br>0.181 174 | 0.172 64<br>0.196 55   | 0.134 171<br>0.095 183 | 0.022 218 0.021 219 | 0.054 194<br>0.109 155 |
| 8        |        | Word Familiarity                   |                          | SubtlexUS SUBTLWF value per Word  | 0.181 174 0.181 175    | 0.196 53               | 0.095 185              | 0.021 219           | 0.109 155              |
| 8        |        | Wiki Knowledge                     |                          | Semantic Clarity, 100 topics extracted from Wiki  | 0.178 176              | 0.01 223               | 0.153 167              | 0.171 143           | 0.084 171              |
| 8        |        | Wiki Knowledge                     |                          | Semantic Clarity, 150 topics extracted from Wiki  | 0.165 182              | 0.011 221              | 0.161 161              | 0.185 138           | 0.074 181              |
| 8        |        | Entity Grid                        | LoCohPU_S                | Local Coherence for PU score  | 0.129 195              | 0.023 202              | 0.103 179              | 0.084 184           | 0.13 144               |
| 8<br>8   |        | Part-of-Speech<br>Part-of-Speech   | ra_VeAvT_C<br>ra_SuNoT_C | ratio of Verb POS # to Adverb POS #<br>ratio of Subordinating Conjunction POS # to Noun POS #             | 0.108 200<br>0.085 202 | 0.078 136<br>0.149 82  | 0.229 144<br>0.039 211 | 0.025 215 0.155 151 | 0.079 174<br>0.158 126 |
| 8        |        | WB Knowledge                       | BRich15_S                | Semantic Richness, 150 topics extracted from WeeBit   | 0.085 202              | 0.149 82 0.059 152     | 0.039 211 0.154 166    | 0.135 131           | 0.138 126              |
| 8        |        | Part-of-Speech                     | ra_NoCoT_C               | ratio of Noun POS # to Coordinating Conjunction #   | 0.022 222              | 0.254 45               | 0.019 221              | 0.053 201           | 0.109 157              |
| 8        | LxSem  | Type Token Ratio                   | MTLDTTR_S                | Measure of Textual Lexical Diversity (default TTR = $0.72$ )  | 0.0 230                | 0.103 123              | 0.119 175              | 0.151 152           | 0.0 231                |
|          |        |                                    |                          |   |                        |                        |                        |                     |                        |

Table 15: Part C. The full generalizability ranking of handcrafted linguistic features under Approach B.

|       |        |                  |               | Feature  | CCI   | В   | WBT       | CAM      |          | CKC    |       | Е   |
|-------|--------|------------------|---------------|--|-------|-----|-----------|----------|----------|--------|-------|-----|
| Score | Branch | Subgroup         | LingFeat Code | Brief Explanation  | r     | rk  | r rk      | r r      | k   ::   | rk     | r     | rk  |
| 7     | LxSem  | Word Familiarity | at_SbFrL_C    | SubtlexUS FREQlow value per Word                           | 0.176 | 178 | 0.171 65  | 0.061 20 | 0.0      | 01 228 | 0.09  | 165 |
| 7     | Synta  | Part-of-Speech   | ra_AvNoT_C    | ratio of Adverb POS # to Noun POS #                        | 0.171 | 179 | 0.108 119 | 0.076 19 | 05 0.0   | 84 185 | 0.023 | 222 |
| 7     | Disco  | Entity Grid      | LoCoDPU_S     | Local Coherence distance for PU score                      | 0.154 | 186 | 0.032 189 | 0.086 19 | 0.0      | 87 182 | 0.111 | 154 |
| 7     | Synta  | Phrasal          | at_AvPhr_C    | # Adverb phrases per Word                                  | 0.144 | 188 | 0.113 115 | 0.047 20 | 0.0   80 | 29 214 | 0.058 | 191 |
| 7     | Synta  | Phrasal          | ra_AjSuP_C    | ratio of Adj phrases # to Subordinate Clauses #            | 0.133 | 194 | 0.04 177  | 0.195 15 | 55 0.0   | 01 227 | 0.079 | 173 |
| 7     | Synta  | Phrasal          | ra_AjVeP_C    | ratio of Adj phrases # to Verb phrases #                   | 0.104 | 201 | 0.01 222  | 0.055 20 | 0.0      | 83 186 | 0.124 | 148 |
| 7     | Synta  | Part-of-Speech   | ra_CoAjT_C    | ratio of Coordinating Conjunction POS # to Adjective POS # | 0.068 | 206 | 0.051 161 | 0.176 10 | 50 0.0   | 74 191 | 0.104 | 160 |
| 7     | Synta  | Part-of-Speech   | ra_AvCoT_C    | ratio of Adverb POS # to Coordinating Conjunction #        | 0.029 | 216 | 0.119 110 | 0.024 2  | 6 0.0    | 22 217 | 0.107 | 158 |
| 7     | Synta  | Part-of-Speech   | ra_AjSuT_C    | ratio of Adjective POS # to Subordinating Conjunction #    | 0.025 | 219 | 0.001 233 | 0.024 2  | 7 0.2    | 04 131 | 0.057 | 192 |
| 7     | Synta  | Phrasal          | ra_SuAjP_C    | ratio of Subordinate Clauses # to Adj phrases #            | 0.02  | 223 | 0.022 205 | 0.05 20  | 0.2      | 04 129 | 0.029 | 218 |
| 7     | Synta  | Phrasal          | ra_PrSuP_C    | ratio of Prep phrases # to Subordinate Clauses #           | 0.002 | 228 | 0.076 139 | 0.143 10 | 59 O.    | 07 193 | 0.13  | 143 |
| 6     | Synta  | Part-of-Speech   | ra_AvVeT_C    | ratio of Adverb POS # to Verb POS #                        | 0.168 | 181 | 0.011 220 | 0.097 18 | 31 0.0   | 53 203 | 0.053 | 195 |
| 6     | Synta  | Part-of-Speech   | ra_AjNoT_C    | ratio of Adjective POS # to Noun POS #                     | 0.074 | 205 | 0.146 89  | 0.031 2  | 3 0.0    | 68 195 | 0.041 | 205 |
| 6     | Synta  | Phrasal          | ra_VeAjP_C    | ratio of Verb phrases # to Adj phrases #                   | 0.067 | 207 | 0.072 142 | 0.087 19 | 0 0.1    | 04 176 | 0.064 | 189 |
| 6     | Synta  | Part-of-Speech   | ra_AvAjT_C    | ratio of Adverb POS # to Adjective POS #                   | 0.061 | 208 | 0.049 163 | 0.088 18 | 39 0.1   | 07 174 | 0.039 | 208 |
| 6     | Synta  | Phrasal          | ra_NoAjP_C    | ratio of Noun phrases # to Adj phrases #                   | 0.05  | 209 | 0.084 132 | 0.073 19 | 0.1      | 28 166 | 0.062 | 190 |
| 6     | Synta  | Part-of-Speech   | ra_NoSuT_C    | ratio of Noun POS # to Subordinating Conjunction #         | 0.049 | 210 | 0.075 140 | 0.004 22 | 29 0.1   | 86 137 | 0.077 | 178 |
| 6     | Synta  | Phrasal          | ra_VeAvP_C    | ratio of Verb phrases # to Adv phrases #                   | 0.039 | 213 | 0.084 133 | 0.155 10 | 65 0.0   | 65 198 | 0.097 | 163 |
| 6     | Synta  | Part-of-Speech   | ra_CoSuT_C    | ratio of Coordinating Conj POS # to Subordinating Conj #   | 0.03  | 215 | 0.076 137 | 0.044 2  | 0 0.1    | 96 134 | 0.001 | 229 |
| 6     | Synta  | Phrasal          | at_AjPhr_C    | # Adjective phrases per Word                               | 0.027 | 218 | 0.046 167 | 0.029 2  | 4 0.0    | 76 190 | 0.126 | 147 |
| 6     | Synta  | Phrasal          | ra_AjNoP_C    | ratio of Adj phrases # to Noun phrases #                   | 0.01  | 226 | 0.046 168 | 0.013 22 | 25 0.0   | 66 197 | 0.127 | 145 |
| 6     | Synta  | Part-of-Speech   | ra_AjCoT_C    | ratio of Adjective POS # to Coordinating Conjunction #     | 0.0   | 229 | 0.148 86  | 0.049 20 | 0.0 0.0  | 91 181 | 0.077 | 177 |
| 5     | Synta  | Phrasal          | ra_AvAjP_C    | ratio of Adv phrases # to Adj phrases #                    | 0.044 | 212 | 0.044 170 | 0.066 20 | 0.0      | 86 183 | 0.088 | 167 |
| 5     | Synta  | Part-of-Speech   | at_AvTag_C    | # Adverb POS tags per Word                                 | 0.029 | 217 | 0.072 141 | 0.095 18 | 35 0.0   | 11 223 | 0.078 | 175 |
| 5     | Synta  | Phrasal          | ra_AvVeP_C    | ratio of Adv phrases # to Verb phrases #                   | 0.02  | 225 | 0.068 144 | 0.005 22 | 28   0.0 | 03 225 | 0.071 | 184 |
| 5     | Disco  | Entity Grid      | ra_XXTo_C     | ratio of xx transitions to total                           | 0.0   | 231 | 0.025 198 | 0.0 23   | 31 0     | 0 231  | 0.0   | 230 |
| 5     | Disco  | Entity Grid      | ra_XSTo_C     | ratio of xs transitions to total                           | 0.0   | 232 | 0.025 197 | 0.0 23   | 32 0     | 0 232  | 0.0   | 232 |
| 5     | Disco  | Entity Grid      | ra_SSTo_C     | ratio of ss transitions to total                           | 0.0   | 233 | 0.025 199 | 0.0 23   | 33 0     | 0 233  | 0.0   | 233 |

Table 16: Part D. The full generalizability ranking of handcrafted linguistic features under Approach B.