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 Despite the recent machine learning-based readability assessment models, many researchers insist on using the outdated formulas.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inaccurate formulas, we: : **1.** introduce Common Readability Formula (CoRF), **2.** recalibrate outdated formulas (Flesch-Kincaid Grade Level, Fog Index, SMOG Index, Coleman-Liau Index, and Automated Readability Index), **3.** evaluate the formulas, and **4.**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

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-the-art (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?"

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,

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

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 [1]. 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 linguistically-lacking 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

[1] corestandards.org

Properties	CCB	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

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 $625\frac{\text{item}}{\text{class}}$ 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 non-native country's official ESL education curriculum.

2.5 OneStopEnglish (OSE) - indep. & paired

OSE is another well-regarded dataset in RA. It aims at ESL (English as Second Language) learners and consists of three paraphrased versions of an article from The Guardian Newspaper. Along with the original OSE dataset, we created a paired version (OSE-Pair). This variation has 189 items and each item 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).

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.

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

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 over-generalization that "all" words with more than two syllables are difficult. Indeed, "banana" is quite an easy word.

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

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

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

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

Coleman-Liau Index (COLE) is more of a lesser-used 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.

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

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.

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

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 story-type 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. $\frac{\text{\#letter}}{\text{\#word}}$, $\frac{\text{\#word}}{\text{\#sent}}$, $\frac{\text{\#syllable}}{\text{\#word}}$) 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, c). The calculation was done by using non-linear least squares in fitting functions (Virtanen et al., 2020). Additional details available in appendix 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.

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 $\frac{\text{\#letter}}{\text{\#word}}$.
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					CCB		WBT		CAM		CKC		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	CorrTTR_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					CCB		WBT		CAM		CKC		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).

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

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

4.4 Approach B - Absolute Correlation

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 correlation-dependent. 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 correlation-based (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 entity-related) 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 direct classification of word difficulties always outperformed vague approaches like SubtlexUS word frequency (Brysbaert and New, 2009).

Back to feature selection, we follow the steps below. We hoped to ensure linguistic breadth and meaningful correlation performance.

1. From top to bottom, go through ranking (table 3 & 4) to sort out the features that performed the best in each linguistic subgroup.

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.

The steps above produce the same results for 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 familiarity), CorrTTR_S (type token ratio). CorrNov_S (variation) only appeared under approach B, and we did not include it. Putting together gives:

$$\textbf{CoRF} = \frac{a \cdot \sum \text{AoA} + b \cdot \sum \text{familiarity}}{\#\text{sentence}} + \frac{c \cdot \#\text{CW} + d \cdot \#\text{NP} + e \cdot \sum \text{TH}}{\#\text{sentence}} + f \cdot \frac{\#\text{unique word}}{\sqrt{\#\text{word}}} + g$$

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.

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)

a	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 (for output 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

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

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.

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%).

a) Adv-Ele	CoRF	FKGL	FOGI	SMOG	COLE	AUTO
Accuracy	N.A.	74.2%	64.9%	11.4%	66.0%	78.0%
Accuracy	77.4%	62.7%	51.8%	11.4%	71.1%	65.2%
b) Adv-Int	**CoRF**	**FKGL**	**FOGI**	**SMOG**	**COLE**	**AUTO**
Accuracy	N.A.	70.2%	63.0%	12.2%	63.6%	74.7%
Accuracy	77.8%	60.4%	51.3%	12.2%	67.7%	65.9%
c) Int-Ele	**CoRF**	**FKGL**	**FOGI**	**SMOG**	**COLE**	**AUTO**
Accuracy	N.A.	69.8%	61.3%	9.02%	61.9%	73.2%
Accuracy	73.1%	59.7%	48.9%	9.02%	66.5%	62.1%

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

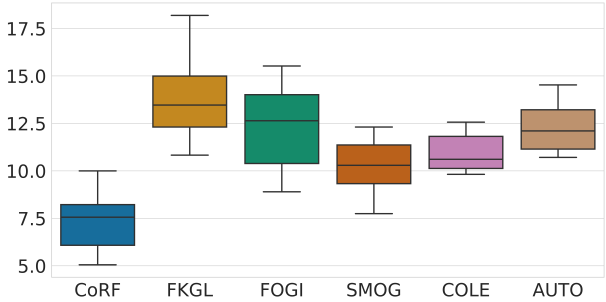

Figure 1: On medical texts. CoRF, against five others.

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 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.

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 over-inflated 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.

References

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 efl 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.

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.

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.

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 efl 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.

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.

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-of-thumb 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:

```
import <Anonymous>
new_object = <Anonymous>.request("string")
readability1 = new_object.CoRF()
readability2 = new_object.FKGL()
readability3 = new_object.FOGI()
readability4 = new_object.SMOG()
readability5 = new_object.COLE()
readability6 = new_object.AUTO()
time_to_read = new_object.RT()
```

CoRF(), FKGL(), FOGI(), SMOG(), COLE(), AUTO(), RT() are shortcut functions. It can be *slightly* faster to directly call in the full forms as:

common_readability_formula()
flesch_kincaid_grade_level()
fog_index()
smog_index()
coleman_liau_index()
automated_readability_index()
read_time()

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 passed without some city or other registering in its archives an event of this kind. There were nobles, who made war against each other; there was the king, who made war against the cardinal; there was Spain, which made war against the king. Then, in addition to these concealed or public, secret or open wars, there were robbers, mendicants, Huguenots, wolves, and scoundrels, who made war upon everybody. The citizens always took up arms readily against thieves, wolves or scoundrels, often against nobles or Huguenots, sometimes against the king, but never against the cardinal or Spain. It resulted, then, from this habit that on the said first Monday 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

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 Health, Ontario Canada

ITEM C

BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language 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 job in keeping calculation speed short for longer texts as originally intended. Second, CoRF's calculation speed linearly increases in respect to the text length. Though, we believe that CoRF's speed is decent in its wide linguistic coverage, it seems true that the speed is weakness when compared to the other readability formulas.

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.

B External Resources

B.1 Python Libraries

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).

Sci-Kit Learn v.1.0.1
Calculation, r2 score and MAE in Tables 2 and 6.
SciPy v.1.7.3
Calculation of Pearson's r for Tables 2 and 6. Fitting function (scipy.optimize.curve_fit()) used to recalibrate traditional readability formulas and give coefficients for CoRF in Tables 2 and 5.
NLTK v.3.6.5
Calculation of tree height for CoRF.
LingFeat v.1.0.0-beta.19
Extraction of handcrafted linguistic features.

B.2 Datasets

New Class	CCB	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[2]. WBT was obtained from the authors[3] in HTML format. We conducted basic preprocessing and manipulated WBT in CSV format. CAM was retrieved from an existing archive[4]. CKC was retrieved from a South Korean educational company[5]. OSE was retrieved from a public archive[6]. NSL was obtained from an American educational company[7]. AUGS medical texts (refer to Section 6.3) were manually scraped from the official website[8]. ASSET was obtained from a public repository[9]. 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
0: Gable earned an Academy Award nomination for portraying Fletcher Christian in Mutiny on the Bounty.
1: Gable also earned an Oscar nomination when he portrayed Fletcher Christian in 1935's Mutiny on the Bounty.
2: Gable won an Academy Award vote when he acted in 1935's Mutiny on the Bounty as Fletcher Christian.
3: Gable also won an Academy Award nomination when he played Fletcher Christian in the 1935 film Mutiny on the Bounty.
4: Gable was nominated for an Academy Award for portraying Fletcher Christian in 1935's Mutiny on the Bounty.
5: Gable also earned an Academy Award nomination in 1935 for playing Fletcher Christian in "Mutiny on the Bounty.
6: 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).
8: Gable earned an Academy Award nomination for his role as Fletcher Christian in the 1935 film Mutiny on the Bounty.
9: Gable also got an Academy Award nomination when he played Fletcher Christian in 1935's movie, Mutiny on the Bounty.
10: 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

[2] corestandards.org/assets/Appendix_B.pdf
[3] Dr. Sowmya Vajjala, National Research Council, Canada
[4] ilexir.co.uk/datasets/index.html
[5] Bruce W. Lee, LAIR, LXPER Inc., South Korea
[6] github.com/nishkalavallabhi/OneStopEnglishCorpus
[7] Luke Orland, Newsela Inc., New York, U.S.A.
[8] augs.org/patient-fact-sheets/
[9] github.com/facebookresearch/asset

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

OSE-Sent (INT-ELE)
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 become 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 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. 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.

C CCB Human Predictions

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

E Computing Power

Single CPU chip. **Architecture**: x86_64; **CPU(s)**: 16; **Model name**: Intel(R) Core(TM) i9-9900KF CPU @ 3.60GHz; **CPU MHz**: 800.024

Feature					CCB		WBT		CAM		CKC		OSE	
Score	Branch	Subgroup	LingFeat Code	Brief Explanation	r	rk	r	rk	r	rk	r	rk	r	rk
43	ShaTr	Shallow	as_Sylla_C	# syllables per Sent	0.541	24	0.461	10	0.686	50	0.697	11	0.59	31
43	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
43	ShaTr	Shallow	as_Chara_C	# characters per Sent	0.539	27	0.487	4	0.696	46	0.711	5	0.613	20
43	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
42	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
40	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
40	Synta	Tree Structure	as_TreeH_C	Tree height per Sent	0.55	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
40	ShaTr	Shallow	as_Token_C	# tokens per Sent	0.494	40	0.464	9	0.65	60	0.709	7	0.58	36
39	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
39	Synta	Phrasal	as_NoPhr_C	# Noun phrases per Sent	0.55	20	0.406	25	0.66	58	0.673	18	0.582	35
39	Synta	Phrasal	to_PrPhr_C	total # prepositional phrases	0.47	47	0.189	58	0.808	11	0.58	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 lemmas AoA of lemmas	0.428	71	0.189	59	0.835	3	0.627	22	0.716	5
38	LxSem	Psycholinguistic	to_AAKuW_C	total AoA (Age of Acquisition) of words	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	0.59	32
36	LxSem	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
35	LxSem	Type Token Ratio	CorrTTR_S	Corrected TTR	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
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
34	LxSem	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
34	LxSem	Psycholinguistic	as_AABrL_C	lemmas AoA of lemmas, Bristol norm per Sent	0.532	31	0.339	31	0.649	62	0.597	31	0.499	57
34	LxSem	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
33	LxSem	Psycholinguistic	as_AABiL_C	lemmas AoA of lemmas, Bird norm per Sent	0.459	55	0.458	11	0.582	73	0.653	19	0.443	69
33	Synta	Phrasal	to_NoPhr_C	total # Noun phrases	0.416	76	0.148	84	0.809	8	0.527	52	0.659	9
33	Synta	Part-of-Speech	to_ContW_C	total # Content words	0.402	81	0.163	71	0.804	14	0.558	40	0.654	11
32	LxSem	Variation Ratio	CorrNoV_S	Corrected Noun Variation-1	0.717	3	0.086	131	0.842	2	0.406	78	0.612	21
32	LxSem	Variation Ratio	CorrVeV_S	Corrected Verb Variation-1	0.602	11	0.058	155	0.801	15	0.393	86	0.737	2
32	LxSem	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
32	Synta	Part-of-Speech	as_VeTag_C	# Verb POS tags per Sent	0.428	70	0.476	6	0.578	74	0.588	34	0.505	55
32	Synta	Tree Structure	to_FTree_C	total length of flattened Trees	0.396	87	0.166	69	0.805	12	0.538	49	0.676	8
31	LxSem	Variation Ratio	SquaNoV_S	Squared Noun Variation-1	0.645	9	0.124	109	0.815	7	0.401	84	0.583	34
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
31	Synta	Part-of-Speech	to_AjTag_C	total # Adjective POS tags	0.441	62	0.191	57	0.777	23	0.504	54	0.525	46
30	LxSem	Variation Ratio	SquaVeV_S	Squared Verb Variation-1	0.559	17	0.076	138	0.777	22	0.384	90	0.716	4
30	Synta	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
30	Synta	Phrasal	as_VePhr_C	# Verb phrases per Sent	0.383	90	0.455	12	0.59	72	0.586	35	0.505	54
29	LxSem	Word Familiarity	as_SbCDL_C	SubtlexUS CDlow value per Sent	0.432	65	0.441	14	0.527	82	0.623	26	0.401	85
28	Synta	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
28	Disco	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
28	Synta	Tree Structure	at_TreeH_C	Tree height per Word	0.476	45	0.419	22	0.416	104	0.597	33	0.41	81
28	LxSem	Word Familiarity	as_SbCDC_C	SubtlexUS 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	SubtlexUS SUBTLCD value per Sent	0.431	68	0.437	18	0.525	85	0.624	25	0.404	83
28	LxSem	Word Familiarity	to_SbL1C_C	total SubtlexUS Lg10CD value	0.37	93	0.14	95	0.797	16	0.491	56	0.621	17
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	SubtlexUS FREQlow value per Sent	0.443	60	0.426	21	0.52	86	0.552	42	0.425	77
26	LxSem	Word Familiarity	as_SbSBW_C	SubtlexUS SUBTLWF value per Sent	0.44	63	0.441	15	0.509	91	0.542	48	0.425	76
26	LxSem	Word Familiarity	as_SbFrQ_C	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_SbL1W_C	total SubtlexUS Lg10WF value	0.365	99	0.144	93	0.795	17	0.477	58	0.611	22
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	0.565	39
25	LxSem	Word Familiarity	to_SbFrL_C	total SubtlexUS FREQlow value	0.348	109	0.201	51	0.774	24	0.414	74	0.555	40
24	LxSem	Word Familiarity	to_SbFrQ_C	total SubtlexUS FREQ# value	0.34	116	0.206	48	0.77	26	0.403	82	0.551	41
24	LxSem	Word Familiarity	to_SbSBW_C	total SubtlexUS SUBTLWF value	0.34	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_SuPhr_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	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	0.2	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_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	0.584	14	0.015	214	0.616	68	0.617	27	0.137	138
20	LxSem	Variation Ratio	SquaAvV_S	Squared AdVerb Variation-1	0.515	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 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	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_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	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					CCB		WBT		CAM		CKC		OSE	
Score	Branch	Subgroup	LingFeat Code	Brief Explanation	r	rk	r	rk	r	rk	r	rk	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	Synta	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	LxSem	Type Token Ratio	BiLoTTR_S	Bi-Logarithmic TTR	0.591	13	0.062	149	0.07	200	0.001	229	0.523	47
16	AdSem	Wiki Knowledge	WRich15_S	Semantic Richness, 150 topics extracted from Wiki	0.495	39	0.02	208	0.48	95	0.549	45	0.037	209
16	Synta	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	Phrasal	to_AvPhr_C	total # Adverb phrases	0.356	105	0.17	66	0.705	45	0.298	111	0.432	73
15	ShaTr	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	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
13	Synta	Phrasal	at_PrPhr_C	# prepositional phrases per Word	0.57	16	0.133	103	0.316	124	0.323	105	0.366	92
13	Synta	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	Disco	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	LxSem	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
13	Synta	Part-of-Speech	as_SuTag_C	# Subordinating Conjunction POS tags per Sent	0.366	97	0.295	39	0.407	105	0.427	68	0.151	131
13	ShaTr	Shallow	TokSenS_S	sqrt(total # tokens x total # sentence)	0.241	154	0.064	147	0.758	32	0.249	121	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	Synta	Phrasal	as_AvPhr_C	# Adverb phrases per Sent	0.244	152	0.328	34	0.427	103	0.38	92	0.356	93
12	Disco	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
12	Synta	Phrasal	to_AjPhr_C	total # Adjective phrases	0.339	120	0.182	62	0.682	53	0.327	101	0.271	111
11	AdSem	Wiki Knowledge	WNois05_S	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	ShaTr	Shallow	TokSenM_S	total # tokens x total # sentence	0.189	173	0.112	116	0.674	55	0.177	140	0.486	60
10	Synta	Phrasal	ra_VeNoP_C	ratio of Verb phrases # to Noun phrases #	0.46	54	0.164	70	0.124	174	0.041	209	0.027	220
10	Disco	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
9	LxSem	Variation Ratio	SimpNoV_S	Noun Variation-1	0.499	38	0.087	130	0.038	212	0.031	213	0.337	95
9	Synta	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	LxSem	Word Familiarity	at_SbL1W_C	SubtlexUS Lg10WF value per Word	0.399	84	0.089	129	0.531	81	0.24	123	0.412	80
9	Synta	Part-of-Speech	ra_VeNoT_C	ratio of Verb POS # to Noun POS #	0.397	86	0.198	53	0.234	142	0.171	142	0.067	186
9	LxSem	Word Familiarity	at_SbSBC_C	SubtlexUS SUBTLCD value per Word	0.37	94	0.032	192	0.492	93	0.324	103	0.435	71
9	LxSem	Word Familiarity	at_SbCDC_C	SubtlexUS CD# value per Word	0.37	95	0.032	191	0.492	94	0.324	104	0.435	70
9	Synta	Phrasal	as_AjPhr_C	# Adjective phrases per Sent	0.323	128	0.239	46	0.387	106	0.357	95	0.157	127
9	AdSem	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	AdSem	Wiki Knowledge	WNois15_S	Semantic Noise, 150 topics extracted from Wiki	0.388	88	0.033	188	0.454	101	0.454	63	0.006	226
8	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
8	AdSem	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	Synta	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	LxSem	Word Familiarity	at_SbCDL_C	SubtlexUS CDlow value per Word	0.362	102	0.047	166	0.474	96	0.31	107	0.431	74
7	Synta	Part-of-Speech	ra_CoNoT_C	ratio of Coordinating Conjunction POS # to Noun POS #	0.02	224	0.277	42	0.159	163	0.013	222	0.132	142
7	Synta	Part-of-Speech	at_CoTag_C	# Coordinating Conjunction POS tags per Word	0.218	161	0.267	44	0.02	220	0.111	172	0.087	169
7	Synta	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	Synta	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	Disco	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
6	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
6	LxSem	Word Familiarity	at_SbSBW_C	SubtlexUS SUBTLWF value per Word	0.181	175	0.196	54	0.095	184	0.021	220	0.109	156
6	LxSem	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	Synta	Part-of-Speech	ra_NoVeT_C	ratio of Noun POS # to Verb POS #	0.432	66	0.118	111	0.149	168	0.112	170	0.051	197
5	AdSem	Wiki Knowledge	WRich10_S	Semantic Richness, 100 topics extracted from Wiki	0.364	100	0.002	232	0.33	123	0.411	76	0.041	206
5	Disco	Entity Grid	ra_NXTo_C	ratio of nx transitions to total	0.339	118	0.097	127	0.62	66	0.28	116	0.278	110
5	Synta	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	AdSem	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	LxSem	Variation Ratio	SimpVeV_S	Verb Variation-1	0.286	139	0.048	165	0.081	193	0.003	226	0.48	62
5	Synta	Part-of-Speech	ra_VeCoT_C	ratio of Verb POS # to Coordinating Conjunction #	0.192	172	0.172	64	0.134	171	0.022	218	0.054	194
5	LxSem	Word Familiarity	at_SbFrL_C	SubtlexUS FREQlow value per Word	0.176	178	0.171	65	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	LxSem	Type Token Ratio	SimpTTR_S	unique tokens/total tokens (TTR)	0.375	92	0.025	200	0.367	113	0.163	147	0.344	94
4	AdSem	Wiki Knowledge	WNois10_S	Semantic Noise, 100 topics extracted from Wikip	0.34	117	0.021	207	0.376	109	0.426	69	0.03	216
4	Synta	Phrasal	at_SuPhr_C	# Subordinate Clauses per Word	0.204	165	0.157	78	0.246	140	0.314	106	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	AdSem	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	Synta	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
3	Synta	Phrasal	ra_NoVeP_C	ratio of Noun phrases # to Verb phrases #	0.406	79	0.068	145	0.069	201	0.031	212	0.019	223
3	AdSem	Wiki Knowledge	WRich05_S	Semantic Richness, 50 topics extracted from Wiki	0.405	80	0.063	148	0.347	117	0.301	110	0.035	211
3	Synta	Phrasal	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	LxSem	Variation Ratio	SimpAjV_S	Adjective Variation-1	0.398	85	0.109	118	0.279	134	0.073	192	0.201	120
3	Synta	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
3	Synta	Part-of-Speech	ra_NoAjT_C	ratio of Noun POS # to Adjective POS #	0.121	199	0.152	81	0.125	173	0.114	169	0.004	228
3	Synta	Part-of-Speech	ra_SuNoT_C	ratio of Subordinating Conjunction POS # to Noun POS #	0.085	202	0.149	82	0.039	211	0.155	151	0.158	126
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	Synta	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	Synta	Phrasal	ra_SuVeP_C	ratio of Subordinate Clauses # to Verb phrases #	0.345	113	0.052	160	0.343	118	0.376	93	0.083	172
2	Synta	Part-of-Speech	ra_CoFuW_C	ratio of Content words to Function words	0.284	141	0.023	203	0.2	154	0.376	94	0.042	201
2	Disco	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
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	AdSem	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	Disco	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
2	Synta	Part-of-Speech	ra_AjCoT_C	ratio of Adjective POS # to Coordinating Conjunction #	0.0	229	0.148	86	0.049	207	0.091	181	0.077	177
2	Synta	Part-of-Speech	ra_AjNoT_C	ratio of Adjective POS # to Noun POS #	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					CCB		WBT		CAM		CKC		OSE	
Score	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					CCB		WBT		CAM		CKC		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	ShaTr	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	Phrasal	as_PrPhr_C	# prepositional phrases per Sent	0.513	35	0.417	23	0.607	70	0.608	28	0.59	32
31	ShaTr	Shallow	as_Token_C	# tokens per Sent	0.494	40	0.464	9	0.65	60	0.709	7	0.58	36
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 lemmas AoA of lemmas	0.428	71	0.189	59	0.835	3	0.627	22	0.716	5
31	LxSem	Psycholinguistic	to_AAKuW_C	total AoA (Age of Acquisition) of words	0.427	72	0.189	60	0.835	4	0.625	23	0.715	6
30	LxSem	Type Token Ratio	CorrTTR_S	Corrected TTR	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	0.086	131	0.842	2	0.406	78	0.612	21
30	Synta	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	Phrasal	to_PrPhr_C	total # prepositional phrases	0.47	47	0.189	58	0.808	11	0.58	36	0.729	3
30	LxSem	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	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
29	LxSem	Variation Ratio	SquaNoV_S	Squared Noun Variation-1	0.645	9	0.124	109	0.815	7	0.401	84	0.583	34
29	LxSem	Variation Ratio	CorrVeV_S	Corrected Verb Variation-1	0.602	11	0.058	155	0.801	15	0.393	86	0.737	2
29	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
29	LxSem	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	LxSem	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	Synta	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	Phrasal	to_NoPhr_C	total # Noun phrases	0.416	76	0.148	84	0.809	8	0.527	52	0.659	9
29	Synta	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	LxSem	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	LxSem	Psycholinguistic	as_AABrL_C	lemmas AoA of lemmas, Bristol norm per Sent	0.532	31	0.339	31	0.649	62	0.597	31	0.499	57
28	LxSem	Psycholinguistic	as_AABiL_C	lemmas AoA of lemmas, Bird norm per Sent	0.459	55	0.458	11	0.582	73	0.653	19	0.443	69
28	LxSem	Word Familiarity	as_SbCDL_C	SubtlexUS CDlow value per Sent	0.432	65	0.441	14	0.527	82	0.623	26	0.401	85
28	LxSem	Word Familiarity	as_SbCDC_C	SubtlexUS 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	SubtlexUS SUBTLCD value per Sent	0.431	68	0.437	18	0.525	85	0.624	25	0.404	83
28	Synta	Part-of-Speech	as_VeTag_C	# Verb POS tags per Sent	0.428	70	0.476	6	0.578	74	0.588	34	0.505	55
28	Synta	Tree Structure	to_FTree_C	total length of flattened Trees	0.396	87	0.166	69	0.805	12	0.538	49	0.676	8
27	LxSem	Variation Ratio	SquaVeV_S	Squared Verb Variation-1	0.559	17	0.076	138	0.777	22	0.384	90	0.716	4
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	Synta	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	LxSem	Word Familiarity	as_SbFrL_C	SubtlexUS FREQlow value per Sent	0.443	60	0.426	21	0.52	86	0.552	42	0.425	77
27	Synta	Part-of-Speech	to_AjTag_C	total # Adjective POS tags	0.441	62	0.191	57	0.777	23	0.504	54	0.525	46
27	LxSem	Word Familiarity	as_SbSBW_C	SubtlexUS SUBTLWF value per Sent	0.44	63	0.441	15	0.509	91	0.542	48	0.425	76
27	LxSem	Word Familiarity	as_SbFrQ_C	SubtlexUS FREQ# value per Sent	0.44	64	0.441	16	0.509	90	0.542	47	0.425	75
27	Synta	Phrasal	as_VePhr_C	# Verb phrases per Sent	0.383	90	0.455	12	0.59	72	0.586	35	0.505	54
26	ShaTr	Shallow	at_Sylla_C	# syllables per Word	0.66	7	0.106	120	0.627	64	0.505	53	0.37	91
26	LxSem	Variation Ratio	CorrAjV_S	Corrected Adjective Variation-1	0.591	12	0.078	134	0.779	21	0.422	70	0.584	33
26	Disco	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	Synta	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	LxSem	Word Familiarity	to_SbL1C_C	total SubtlexUS Lg10CD value	0.37	93	0.14	95	0.797	16	0.491	56	0.621	17
26	LxSem	Word Familiarity	to_SbL1W_C	total SubtlexUS Lg10WF value	0.365	99	0.144	93	0.795	17	0.477	58	0.611	22
26	LxSem	Word Familiarity	to_SbFrL_C	total SubtlexUS FREQlow value	0.348	109	0.201	51	0.774	24	0.414	74	0.555	40
26	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
26	LxSem	Word Familiarity	to_SbFrQ_C	total SubtlexUS FREQ# value	0.34	116	0.206	48	0.77	26	0.403	82	0.551	41

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					CCB		WBT		CAM		CKC		OSE	
Score	Branch	Subgroup	LingFeat Code	Brief Explanation	r	rk	r	rk	r	rk	r	rk	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	Synta	Phrasal	to_SuPhr_C	total # Subordinate Clauses	0.367	96	0.202	50	0.721	43	0.462	61	0.419	78
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	0.565	39
25	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
25	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
25	Synta	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	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
24	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
24	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
23	AdSem	Wiki Knowledge	WTopc20_S	Number of topics, 200 topics extracted from Wikipedia	0.584	14	0.015	214	0.616	68	0.617	27	0.137	138
23	AdSem	Wiki Knowledge	WTopc15_S	Number of topics, 150 topics extracted from Wikipedia	0.58	15	0.007	227	0.645	63	0.605	30	0.191	122
23	LxSem	Variation Ratio	SquaAvV_S	Squared AdVerb Variation-1	0.515	34	0.093	128	0.686	52	0.326	102	0.46	65
23	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
23	Synta	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	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
22	Synta	Phrasal	as_SuPhr_C	# Subordinate Clauses per Sent	0.387	89	0.357	26	0.532	80	0.495	55	0.265	112
22	Synta	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
22	Synta	Phrasal	to_AvPhr_C	total # Adverb phrases	0.356	105	0.17	66	0.705	45	0.298	111	0.432	73
22	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
21	Disco	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	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
20	LxSem	Type Token Ratio	UberTTR_S	Uber Index	0.646	8	0.041	174	0.369	112	0.109	173	0.599	26
20	Synta	Phrasal	at_PrPhr_C	# prepositional phrases per Word	0.57	16	0.133	103	0.316	124	0.323	105	0.366	92
20	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
20	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
20	Disco	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	LxSem	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	Disco	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	Synta	Phrasal	to_AjPhr_C	total # Adjective phrases	0.339	120	0.182	62	0.682	53	0.327	101	0.271	111
20	Synta	Phrasal	as_AvPhr_C	# Adverb phrases per Sent	0.244	152	0.328	34	0.427	103	0.38	92	0.356	93
20	ShaTr	Shallow	TokSenS_S	sqrt(total # tokens x total # sentence)	0.241	154	0.064	147	0.758	32	0.249	121	0.498	59
19	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
19	Synta	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	LxSem	Word Familiarity	at_SbL1W_C	SubtlexUS Lg10WF value per Word	0.399	84	0.089	129	0.531	81	0.24	123	0.412	80
19	LxSem	Word Familiarity	at_SbSBC_C	SubtlexUS SUBTLCD value per Word	0.37	94	0.032	192	0.492	93	0.324	103	0.435	71
19	LxSem	Word Familiarity	at_SbCDC_C	SubtlexUS CD# value per Word	0.37	95	0.032	191	0.492	94	0.324	104	0.435	70
19	Synta	Part-of-Speech	as_SuTag_C	# Subordinating Conjunction POS tags per Sent	0.366	97	0.295	39	0.407	105	0.427	68	0.151	131
19	LxSem	Word Familiarity	at_SbCDL_C	SubtlexUS CDlow value per Word	0.362	102	0.047	166	0.474	96	0.31	107	0.431	74
18	AdSem	Wiki Knowledge	WRich15_S	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	AdSem	Wiki Knowledge	WNois05_S	Semantic Noise, 50 topics extracted from Wiki	0.462	53	0.061	150	0.455	100	0.412	75	0.118	151
18	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
18	Disco	Entity Grid	ra_NXTo_C	ratio of nx transitions to total	0.339	118	0.097	127	0.62	66	0.28	116	0.278	110
18	ShaTr	Shallow	TokSenL_S	log(total # tokens)/log(total # sentence)	0.293	137	0.352	29	0.297	130	0.544	46	0.198	121
18	ShaTr	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	Disco	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	Synta	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
16	AdSem	Wiki Knowledge	WNois15_S	Semantic Noise, 150 topics extracted from Wiki	0.388	88	0.033	188	0.454	101	0.454	63	0.006	226
16	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
16	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
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	AdSem	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
15	LxSem	Type Token Ratio	BiLoTTR_S	Bi-Logarithmic TTR	0.591	13	0.062	149	0.07	200	0.001	229	0.523	47
15	AdSem	Wiki Knowledge	WRich05_S	Semantic Richness, 50 topics extracted from Wiki	0.405	80	0.063	148	0.347	117	0.301	110	0.035	211
15	LxSem	Type Token Ratio	SimpTTR_S	TTR	0.375	92	0.025	200	0.367	113	0.163	147	0.344	94
15	AdSem	Wiki Knowledge	WRich10_S	Semantic Richness, 100 topics extracted from Wiki	0.364	100	0.002	232	0.33	123	0.411	76	0.041	206
15	AdSem	Wiki Knowledge	WNois10_S	Semantic Noise, 100 topics extracted from Wiki	0.34	117	0.021	207	0.376	109	0.426	69	0.03	216
15	Disco	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	AdSem	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	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
14	Synta	Phrasal	ra_SuVeP_C	ratio of Subordinate Clauses # to Verb phrases #	0.345	113	0.052	160	0.343	118	0.376	93	0.083	172
13	Synta	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
13	LxSem	Variation Ratio	SimpAjV_S	Adjective Variation-1	0.398	85	0.109	118	0.279	134	0.073	192	0.201	120
13	Synta	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	Synta	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	Disco	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	Synta	Phrasal	at_SuPhr_C	# Subordinate Clauses per Word	0.204	165	0.157	78	0.246	140	0.314	106	0.073	182
13	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
13	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
13	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
13	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
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					CCB		WBT		CAM		CKC		OSE	
Score	Branch	Subgroup	LingFeat Code	Brief Explanation	r	rk	r	rk	r	rk	r	rk	r	rk
12	LxSem	Variation Ratio	SimpNoV_S	Noun Variation-1	0.499	38	0.087	130	0.038	212	0.031	213	0.337	95
12	Synta	Part-of-Speech	ra_NoVeT_C	ratio of Noun POS # to Verb POS #	0.432	66	0.118	111	0.149	168	0.112	170	0.051	197
12	Synta	Phrasal	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
12	Synta	Part-of-Speech	ra_VeNoT_C	ratio of Verb POS # to Noun POS #	0.397	86	0.198	53	0.234	142	0.171	142	0.067	186
12	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
12	Disco	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	Synta	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
12	Synta	Phrasal	ra_SuAvP_C	ratio of Subordinate Clauses # to Adv phrases #	0.309	133	0.008	226	0.141	170	0.241	122	0.111	153
12	Synta	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
12	AdSem	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
12	AdSem	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
12	AdSem	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
12	AdSem	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	Disco	Entity Density	to_EntiM_C	total number of Entities Mentions	0.139	191	0.02	209	0.335	122	0.0	230	0.312	101
11	Synta	Phrasal	ra_VeNoP_C	ratio of Verb phrases # to Noun phrases #	0.46	54	0.164	70	0.124	174	0.041	209	0.027	220
11	Synta	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
11	Synta	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	Synta	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
11	LxSem	Variation Ratio	SimpVeV_S	Verb Variation-1	0.286	139	0.048	165	0.081	193	0.003	226	0.48	62
11	Synta	Part-of-Speech	ra_CoFuW_C	ratio of Content words to Function words	0.284	141	0.023	203	0.2	154	0.376	94	0.042	201
11	Synta	Part-of-Speech	at_SuTag_C	# Subordinating Conjunction POS tags per Word	0.259	148	0.13	106	0.085	192	0.252	120	0.135	141
11	AdSem	Wiki Knowledge	WClar20_S	Semantic Clarity, 200 topics extracted from Wikipedia	0.144	187	0.016	212	0.308	125	0.23	125	0.034	213
11	AdSem	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	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
11	AdSem	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
10	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
10	Synta	Phrasal	ra_AvSuP_C	ratio of Adv phrases # to Subordinate Clauses #	0.299	136	0.06	151	0.256	139	0.128	165	0.077	176
10	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
10	Disco	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	Disco	Entity Grid	LoCoDPW_S	Local Coherence distance for PW score	0.239	155	0.002	230	0.195	156	0.143	161	0.141	136
10	Disco	Entity Grid	LoCoDPA_S	Local Coherence distance for PA score	0.239	156	0.002	231	0.195	157	0.143	162	0.141	135
10	Synta	Part-of-Speech	at_CoTag_C	# Coordinating Conjunction POS tags per Word	0.218	161	0.267	44	0.02	220	0.111	172	0.087	169
10	LxSem	Variation Ratio	SimpAvV_S	AdVerb Variation-1	0.214	163	0.098	126	0.353	115	0.021	221	0.089	166
10	Synta	Phrasal	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	AdSem	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	Synta	Phrasal	at_NoPhr_C	# Noun phrases per Word	0.424	74	0.066	146	0.089	188	0.005	224	0.042	202
9	Synta	Phrasal	ra_NoVeP_C	ratio of Noun phrases # to Verb phrases #	0.406	79	0.068	145	0.069	201	0.031	212	0.019	223
9	Synta	Phrasal	ra_PrAvP_C	ratio of Prep phrases # to Adv phrases #	0.32	130	0.027	196	0.021	219	0.176	141	0.071	183
9	Synta	Phrasal	at_VePhr_C	# Verb phrases per Word	0.31	132	0.138	97	0.079	194	0.032	211	0.009	225
9	Synta	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	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
9	Disco	Entity Grid	LoCohPW_S	Local Coherence for PW score	0.229	159	0.034	183	0.012	227	0.146	157	0.148	133
9	Disco	Entity Grid	LoCohPA_S	Local Coherence for PA score	0.229	160	0.034	184	0.012	226	0.146	158	0.148	132
9	Synta	Phrasal	ra_SuPrP_C	ratio of Subordinate Clauses # to Prep phrases #	0.218	162	0.048	164	0.015	224	0.07	194	0.227	115
9	Synta	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	Synta	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	Synta	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	Synta	Phrasal	ra_PrAjP_C	ratio of Prep phrases # to Adj phrases #	0.142	189	0.035	182	0.017	223	0.207	128	0.136	140
9	Synta	Part-of-Speech	ra_NoAjT_C	ratio of Noun POS # to Adjective POS #	0.121	199	0.152	81	0.125	173	0.114	169	0.004	228
9	AdSem	WB Knowledge	BClar10_S	Semantic Clarity, 100 topics extracted from WeeBit	0.079	204	0.015	216	0.269	137	0.148	155	0.181	123
9	Synta	Part-of-Speech	ra_CoNoT_C	ratio of Coordinating Conjunction POS # to Noun POS #	0.02	224	0.277	42	0.159	163	0.013	222	0.132	142
8	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
8	Synta	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	Synta	Part-of-Speech	ra_VeSuT_C	ratio of Verb POS # to Subordinating Conjunction #	0.266	146	0.046	169	0.09	186	0.105	175	0.065	188
8	Synta	Phrasal	ra_AvNoP_C	ratio of Adv phrases # to Noun phrases #	0.257	149	0.128	108	0.072	199	0.044	208	0.051	196
8	Synta	Part-of-Speech	ra_SuAjT_C	ratio of Subordinating Conjunction POS # to Adjective POS #	0.244	151	0.008	225	0.074	197	0.082	187	0.138	137
8	Synta	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	Synta	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	Synta	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	Synta	Part-of-Speech	ra_VeCoT_C	ratio of Verb POS # to Coordinating Conjunction #	0.192	172	0.172	64	0.134	171	0.022	218	0.054	194
8	LxSem	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
8	LxSem	Word Familiarity	at_SbSBW_C	SubtlexUS SUBTLWF value per Word	0.181	175	0.196	54	0.095	184	0.021	220	0.109	156
8	AdSem	Wiki Knowledge	WClar10_S	Semantic Clarity, 100 topics extracted from Wiki	0.178	176	0.01	223	0.153	167	0.171	143	0.084	171
8	AdSem	Wiki Knowledge	WClar15_S	Semantic Clarity, 150 topics extracted from Wiki	0.165	182	0.011	221	0.161	161	0.185	138	0.074	181
8	Disco	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	Synta	Part-of-Speech	ra_VeAvT_C	ratio of Verb POS # to Adverb POS #	0.108	200	0.078	136	0.229	144	0.025	215	0.079	174
8	Synta	Part-of-Speech	ra_SuNoT_C	ratio of Subordinating Conjunction POS # to Noun POS #	0.085	202	0.149	82	0.039	211	0.155	151	0.158	126
8	AdSem	WB Knowledge	BRich15_S	Semantic Richness, 150 topics extracted from WeeBit	0.025	220	0.059	152	0.154	166	0.145	159	0.1	162
8	Synta	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					CCB		WBT		CAM		CKC		OSE	
Score	Branch	Subgroup	LingFeat Code	Brief Explanation	r	rk	r	rk	r	rk	r	rk	r	rk
7	LxSem	Word Familiarity	at_SbFrL_C	SubtlexUS FREQlow value per Word	0.176	178	0.171	65	0.061	203	0.001	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	195	0.084	185	0.023	222
7	Disco	Entity Grid	LoCoDPU_S	Local Coherence distance for PU score	0.154	186	0.032	189	0.086	191	0.087	182	0.111	154
7	Synta	Phrasal	at_AvPhr_C	# Adverb phrases per Word	0.144	188	0.113	115	0.047	208	0.029	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	155	0.001	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	205	0.083	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	160	0.074	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	216	0.022	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	217	0.204	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	206	0.204	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	169	0.07	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	181	0.053	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	213	0.068	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	190	0.104	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	189	0.107	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	198	0.128	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	229	0.186	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	165	0.065	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	210	0.196	134	0.001	229
6	Synta	Phrasal	at_AjPhr_C	# Adjective phrases per Word	0.027	218	0.046	167	0.029	214	0.076	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	225	0.066	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	207	0.091	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	202	0.086	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	185	0.011	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	228	0.003	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	231	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	232	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	233	0.0	233	0.0	233

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