# Learning Domain-Specific, L1-Specific Measures of Word Readability

**Shane Bergsma — David Yarowsky**

*Dept. of Computer Science and Human Language Technology Center of Excellence*
*Johns Hopkins University*
shane.a.bergsma@gmail.com, yarowsky@cs.jhu.edu

ABSTRACT. *Improved readability ratings for second-language readers could have a huge impact in areas such as education, advertising, and information retrieval. We propose ways to adapt readability measures for users who (a) are proficient in a particular domain, and (b) have a particular native language (L1). Specifically, we predict the readability of individual words. Our learned models use a range of creative features based on diverse statistical, etymological, lexical, and morphological information. We evaluate on a corpus of computational linguistics articles divided according to seven L1s; we show that we can accurately predict the target readability scores in this domain. Our technique improves over several reasonable baselines. We provide an in-depth analysis showing which kinds of information are most predictive of word difficulty in different L1s, and show how this differs for style and content words.*

RÉSUMÉ. *Une amélioration au niveau de la lisibilité linguistique pour les lecteurs de langue seconde pourrait avoir un impact énorme dans les domaines tels que celui de l'éducation, la publicité et des recherches d'information. Nous proposons des moyens d'adapter des mesures de lisibilité pour des utilisateurs qui (a) sont compétents dans un domaine particulier, et (b) ont une langue maternelle spécifique (L1). Plus précisément, nous prévoyons la lisibilité linguistique de mots individuels. Nos fonctions de prédiction utilisent une gamme de caractéristiques basée sur différentes informations statistiques, étymologique, lexicales et morphologiques. En évaluant sur un corpus d'articles en linguistique informatique répartis en sept L1s, nous démontrons que nous pouvons prédire avec précision une cible du niveau de lisibilité dans ce domaine. Nous fournissons une analyse en profondeur démontrant quels types d'informations sont plus prédictifs de la difficulté des mots dans différentes L1s. De plus, nous démontrons comment ceci diffère pour les mots de contenu et les mots grammaticaux.*

KEYWORDS: *Second-language learning, readability.*

MOTS-CLÉS : *L'apprentissage d'une langue seconde, lisibilité linguistique.*

TAL Volume 54 – n° 1/2013, pages 203 à 226

## 1. Introduction

Methods for automatically predicting the difficulty of texts have been intensely developed since the 1920s (Chall, 1988). Educators have widely used the resulting formulas to provide students with reading material at an appropriate level of difficulty. Readability measures have also recently been explored as a "valuable new relevance signal" for search engines (Collins-Thompson *et al.*, 2011), and as a component of several intelligent tutoring systems and teacher-support tools (Schwarm and Ostendorf, 2005; Heilman *et al.*, 2007; Burstein *et al.*, 2007; Miltsakaki and Troutt, 2008).

This paper explores new readability measures for second-language (L2) readers. Like others, our approach is based on predicting the readability of individual English words. Measures of word familiarity have long been the primary component of automatic readability measures, such as the widely-used Flesch-Kincaid and Dale-Chall formulas (Klare, 1974). Many studies have confirmed word difficulty as an "excellent predictor of reading difficulty" (Collins-Thompson and Callan, 2005) that is "highly correlated" with human readability ratings (Pitler and Nenkova, 2008). The key idea of our work is that word difficulty is sensitive to the native language (L1) of the reader, and that there are L1-specific patterns of difficulty that can be exploited to provide better readability scoring for specific populations.

Our study uses a corpus of computational linguistics articles from the ACL Anthology, semi-automatically divided into L1 populations according to author names and affiliations (Section 2.1). We use this data to learn which words are most difficult for the different L1 populations, and why.

Consider the following pair of sentences:

**(1)** We *claim* that this *notation* is *problematic*.

**(2)** We *propose* that new *terminology* should be *adopted*.

To a native English speaker, the sentences are equally readable: the content words *claim*, *notation*, and *problematic* in (1) and *propose*, *terminology*, and *adopted* in (2) all occur hundreds of times in the English portion of our ACL data. But to a native *Chinese* speaker, the sentences are very different: in papers by Chinese authors, the words in (2) also occur hundreds of times, but the words *claim*, *notation*, and *problematic* in (1) hardly occur at all (only a dozen times each).

These differences in frequency not only indicate which words native Chinese speakers *use*, but also which words they *understand*. Frequency has always been used as the primary indicator of word difficulty, and the close correlation between frequency and difficulty has been validated in many studies (Tamayo, 1987; Chall, 1988; Breland, 1996; Crossley *et al.*, 2008). Moreover, careful analysis of academic writing by English-L2 learners has shown that there is a close correlation between a learner's L2 word *usage* and independent, direct measures of their L2 *vocabulary* (Laufer and Nation, 1995). This is a very significant finding; it means that frequencies of L2 words in

texts written by specific populations provide excellent indicators of L2 word difficulty for those populations. [1]

However, we show that the *domain* has a much stronger influence on word frequency than does the population L1. For example, while the word *terminology* is obviously understood by Chinese ACL authors, that same word is absent from the Chinese portion of the International Corpus of Learner English (Granger *et al.*, 2009). Readability measures should therefore also be sensitive to the domain-knowledge of the readers. Unfortunately, we usually lack extensive text written by each L1 population in each domain, and thus lack a domain-specific, L1-specific source of frequency/difficulty ratings.

We thus focus on *learning* L1-specific predictors from a small number of in-domain judgments (perhaps obtained from direct vocabulary tests or sparse in-domain data), and *generalizing* from this observed data to make judgements on unseen words (Section 3). Overall, our work provides a greater understanding of how domain and L1 affect readability. Our main contributions are:

1) providing an efficient statistical framework for learning readability measures, via the technique of support vector regression (Section 4);

2) introducing and evaluating a wide range of new features for making readability predictions, with a detailed analysis across seven distinct L1s (Section 5);

3) demonstrating that we can generalize from observed readability scores to make predictions for new words (Section 7);

4) demonstrating that both domain and L1 are important parameters for readability (Section 7).

## 2. Data

### 2.1. *Dividing ACL Data by L1*

Our main corpus comprises papers from the ACL Anthology Network (Radev *et al.*, 2009, Release 2011). As we discuss in Section 8, the ACL Anthology Network (AAN) has increasingly been used for studies of linguistic style, however we are the first to study style within the AAN's specific L1 populations.

We converted AAN papers to text using the Linux utility `pdftotext` and removed non-sentences such as author names, references and tables. We focused on the eight most-common L1s in the AAN: Japanese, Chinese, Spanish, Italian, French, German, Dutch and English.

1. Of course, word knowledge is not binary. Non-native speakers can often understand more words than they are comfortable using; neglecting to use a word may reflect uncertainty and lack of confidence, leading to a preference for simpler alternatives (Laufer and Nation, 1999). These preferences, and the resulting frequencies, therefore still reflect degrees of word difficulty.

| Language | ACL | | ICLE | |
|---|---|---|---|---|
| | Docs. | Words | Docs. | Words |
| English | 1,573 | 6.4M | - | - |
| Japanese | 992 | 3.3M | 364 | 200K |
| Chinese | 444 | 1.3M | 982 | 493K |
| German | 409 | 1.4M | 443 | 236K |
| French | 260 | 791K | 458 | 287K |
| Dutch | 133 | 511K | 287 | 272K |
| Italian | 101 | 342K | 397 | 228K |
| Spanish | 82 | 274K | 258 | 204K |
| Total | 3,994 | 14.2M | 6,260* | 7.8M* |

**Table 1.** *Statistics on our ACL data and comparison to the (smaller) ICLE. *Note the ICLE contains 9 additional languages, which are included in the* Total

We identified papers written by these L1s using the AAN's manually-curated author name/affiliation meta-data (for papers $\leq$2009). The AAN links the author of each paper with a country, state or province. We manually labeled each of these regions-of-affiliation with the L1 predominantly spoken there. We also leveraged the first name of the authors – an imperfect but useful indicator of an author's L1. To mark native-English names, we collected a list of common first names in the U.S. (via `www.census.gov/genealogy/names/names_files.html`) and expanded it with related nicknames, e.g., Dan for Daniel, Rob for Robert, etc. For non-English names, we extracted lists of common male and female names for our target regions (via `www.20000-names.com`). We created our L1 sets using the following rules: to be labeled as native-English, all author names and affiliations must be English. To be labeled with another L1, (1) every author must have a country-of-affiliation where that L1 is spoken, and (2) every author's name must agree with that L1 *or* be of unknown origin.

These heuristics provide a decent trade-off between precision and recall. They assign 3,994 of the 15,192 total papers in the ACL anthology to one of our L1s. Our corpus provides more data for each target L1 population than is provided by the widely-used International Corpus of Learner English (Table 1). In terms of precision, we manually checked 100 random assignments, and found that in only 5 cases was the assignment fully invalid. In 86 cases, the paper was clearly written by a native speaker of the assigned L1, while in 7 other cases, at least the second (and usually more senior) author speaks the target L1. In 2 cases the L1 of the author was not clear.

While our semi-automatic approach provides decent precision, we believe that some of these annotation errors could have an effect on our results, particularly for those languages with smaller numbers of papers and which are thus more susceptible to label noise. Future work may wish to explore more labor-intensive, fully-manual annotation efforts.

We make our annotations of L1s in the ACL Anthology papers available online as an attachment to this article, available through the TAL website on the same page as the electronic version of this paper.

### 2.2. *Target Words for Evaluation: Stopwords vs. Content Words*

In this paper, we learn and evaluate L1 readability measures separately for both stopwords and content words.

For stopwords, we target mainly function words: relatively common English words that reflect author style rather than content. We started with the standard set of 524 stopwords originally used in the SMART IR system (following Tomokiyo and Jones [2001] who looked at non-native speech patterns). We then removed words that were among the 50 most common words in English (e.g. ubiquitous words such as *the*, *of*, *and*, etc.), and those that occurred less than 10 times in the English ACL data (e.g. unusual words such as *hereupon*, *hither*, *thanx*, etc.), resulting in 348 final stopwords. We then randomly assigned half of these into a training set, a quarter into a test set and a quarter into a development set (for both preliminary testing and tuning the hyperparameters of our learned model [Section 6.1]).

For content words, we use words that are more likely to reflect topic, but also words which authors have some flexibility in using or not. We looked at all the words in the Moby thesaurus,[2] and filtered those that occur less than 5 times in the papers of each of our target ACL L1 populations. We then included only those words with a synonym that is also on our filtered list. This last step helps us select words with flexibility in usage: one author might say *objective*, while another might say *goal*, *purpose*, or *aim*. We arrive at a final list of 838 content words, which we also divide half into training and a quarter for development and a quarter for testing.

We use our ACL frequency information to calculate readability measures as described below (Section 3) for each of our training, test and development words in each of our target L1s; we use this as our gold standard data for experiments.

All of the words used in our evaluation, as well as their readability scores, are also available as an attachment to this article, available through the TAL website.

## 3. The Readability Prediction Task

Our task is to predict the readability of an English word for a particular L1 reader, a reader assumed to be knowledgeable in the domain of computational linguistics. In theory, our gold-standard readability scores could be obtained directly from members of the target populations, e.g. by asking them which words they know (Dolch, 1932; Tilley, 1936) or by administering vocabulary tests such as cloze deletion

2. Available at `www.gutenberg.org/dirs/etext02/mthes10.zip`

tests and semantic association evaluations (Read, 1993; Laufer and Nation, 1995; Breland, 1996, *inter alia*). In this work, since we have extensive writing by the L1 populations in the target domains, we compute our gold-standard scores using a *log-odds-ratio*, which capture the difference in odds of the word being used by a native-English speaker, and the odds of the word being used by the L1 group.

### 3.1. *The Log-Odds-Ratio*

The log-odds-ratio is calculated as follows. Let $N_1$ and $N_2$ be the number of tokens in the native-English and L2-English corpora, respectively, while let $f_1(w)$ and $f_2(w)$ be the frequencies of word $w$ in the native-English and L2 data. The odds of word $w$ in the two sections are:

$$O_1(w) = \frac{f_1(w)}{N_1 - f_1(w)}, \quad O_2(w) = \frac{f_2(w)}{N_2 - f_2(w)} \qquad [1]$$

The log-odds-ratio for word $w$ is then $log(\frac{O_1(w)}{O_2(w)})$. To handle sparsity, we smooth all of our frequency counts using add-1 smoothing.

The log-odds-ratio is an appropriate measure for several reasons. Since it incorporates L1-specific frequencies, it exploits the observation that, when their English usage frequencies are available, these frequencies are valid and reliable indicators of word difficulty for learner populations (Laufer and Nation, 1995). Secondly, since the log-odds-ratio also incorporates the *native-English* ACL frequencies, it is sensitive to the domain of the writing. Finally, while frequency is regarded as the "most influential variable" in much lexical decision research in applied linguistics and psychology (Kuperman *et al.*, 2012) and the "single most important variable... in virtually every task used to measure word recognition, from tachistoscopic recognition to semantic judgment" (Gough, 1984), many of these findings show that the degree of difficulty is often most correlated with the *logarithm* of the frequency (Gough, 1984). By also being based on logarithms, the log-odds-ratio scales with human sensitivity, and is thus easy to interpret.

On the other hand, like other readability measures, the log-odds-ratio is only a proxy for word difficulty, and may not correlate perfectly with readability scores obtained in other ways. Our methodology is therefore based on developing a formula (a regression) to predict something else (a log-odds-ratio) that is an imperfect proxy for what we are really interested in ("true" readability). The important point, however, is that when the frequencies are available, a log-odds-ratio is as good a measure of "true" readability as one is likely to obtain. Designing valid measures of word readability has long been recognized as a tricky undertaking (Dolch, 1932; Tilley, 1936; Read, 1993); questions on direct readability tests often measure the difficulty of the question's hints and context as much as the difficulty of the words themselves. Measures like the log-odds-ratio, on the other hand, based on word frequencies, are not only valid and reliable indicators of readability, but are simple to calculate and apply. By learning to predict these ratios, we not only develop our own formula for readability (applicable

| | | Jap. | Chin. | Spa. | Ital. | Fr. | Ger. | Dut. |
|---|---|---|---|---|---|---|---|---|
| Stop- | *besides* | -0.74 | **-1.96** | -1.54 | -1.05 | -0.85 | -1.23 | -1.11 |
| words | *mainly* | -1.72 | -2.11 | -1.89 | **-2.19** | -1.80 | -1.68 | -1.56 |
| | *presumably* | 1.78 | **3.19** | 0.75 | 0.52 | 3.30 | 0.98 | 1.86 |
| | *perhaps* | **2.22** | 1.86 | 1.51 | 1.91 | 0.95 | 1.51 | 0.85 |
| | *somewhat* | 1.29 | 1.11 | 0.82 | **1.73** | 1.12 | 0.56 | 0.16 |
| | *whereas* | 0.29 | 0.56 | -0.44 | -0.51 | -0.27 | **-0.84** | -0.78 |
| Content | *claim* | 1.02 | **2.44** | 1.20 | 0.33 | 0.74 | 0.70 | 0.56 |
| words | *composed* | -0.48 | -0.85 | -0.69 | -0.93 | **-1.28** | -0.14 | 0.24 |
| | *complementary* | 0.69 | 0.06 | -1.04 | -0.29 | **-1.09** | -0.31 | -0.11 |
| | *considerably* | 0.56 | **1.31** | -0.13 | 0.23 | 0.66 | -0.38 | -0.97 |
| | *notion* | 0.71 | **2.10** | 1.62 | -0.15 | -0.56 | 0.25 | -0.33 |
| | *obtain* | -0.93 | -0.88 | **-1.44** | -0.72 | -0.74 | -0.10 | -0.52 |

**Table 2.** *Native- vs. L2-English* **log-odds-ratios** *for various words in the ACL data. Negative values (in red) are relatively more frequent in the given L1 group, positive values (in green) are relatively more frequent among native-English speakers*

to any word, with or without in-domain frequency data), but also, through our feature analysis (Section 7.2), we begin to understand *why* a word might be difficult for a particular population.

### 3.2. *Examples of Log-Odds-Ratios by L1*

Table 2 gives some words and log-odds-ratios from the ACL data. Note that since the native-English odds are in the numerator of the ratio, positive values are relatively more common in native-English, meaning they are relatively more difficult for those particular non-natives, while negative values are more common in the particular L1 group, meaning they are relatively more readable for those particular non-natives. Our task is essentially to predict the values in Table 2 on the basis of known log-odds-ratios for other words, exploiting patterns of readability that are specific to each L1.

Note that in many cases, the ratios are consistent across all the L1s. E.g., the words *besides* and *mainly* are used much more frequently by all non-native-L1 groups, while *presumably* and *perhaps* are much more common in native writing. In other cases, the historical relationships among languages can play a role. E.g., the ratios for *composed* are close to zero (equal usage) for Dutch and German, but lean toward the L1 speakers for the Romance languages Spanish, Italian and French. We might have predicted this, given there is a close cognate for *composed* in each of Spanish, Italian and French (i.e., *composto*, *compuesto*, *composé*, resp.). In Section 4, we introduce a model that allows us to incorporate such knowledge as features in a learned predictor of English word readability.

Our analysis of such factors as predictors of word difficulty goes well beyond the limited prior explorations of L2 word readability, which we describe in more detail in Section 8. In short, prior work has generally regarded being a non-native speaker as a *single* variable that affects readability (Crossley and McNamara, 2009; Rosa and Eskenazi, 2011), whereas the particular L1 of the speaker has not been taken into consideration.

## 4. A Support Vector Regression Model of Word Readability

Since domain- and L1-specific readability scores will generally not be available for every word, we aim to *predict* these scores on the basis of information that is at hand. We take a machine learning approach to this problem. For each word, we encode available information in a feature vector, $\bar{x}$, and learn a function, $f$, to map this vector to a predicted readability score, $f : \bar{x} \rightarrow \hat{y} \in \mathbb{R}$. The following section (Section 5) describes the rich set of features that we encode in $\bar{x}$. To make our predictions sensitive to the L1, we train a separate predictor $f^i$ for each L1 (indexed by $i$). We assume some $\bar{x}, y$ pairs are available for training the model in each L1. In this paper we use log-odds-ratios as our target $y$'s, but other measures can also be used.

We train $f^i(\bar{x})$ using the technique of Support Vector Regression. SVR is related to estimators from robust statistics and has had "excellent performances" on various regression tasks (Smola and Schölkopf, 2004). We use an $\epsilon$-SVR formulation; a set of weights, $\bar{w}$, is found such that the prediction $\hat{y} = \bar{w} \cdot \bar{x}$ is within $\epsilon$ of the true $y$ for each training point. Deviations from these constraints are allowed, subject to a penalty involving a regularization parameter, $C$. At test time, we apply the learned (L1-specific) $\bar{w}$ to our new feature vectors, and produce estimates $\hat{y}$.

Like classification SVMs, the SVR framework allows us to use kernel functions for non-linear models, implicitly performing the regression in higher-dimensional mappings of feature space.

## 5. Word Readability Features

The success of SVR hinges on the quality of the information encoded in the feature vector. Since each instance is a unique word, our features must be general enough to allow generalization from one word to another. This section describes the five types of features used in our experiments. [3]

The SVR system is able to use these features to model both domain-specific and L1-specific preferences. No matter which feature class we are using, we always include a feature for the log-frequency of the word in the native-English ACL data (that

3. The selection of the features, and their final forms (e.g. binary vs. real-valued, etc.), were based on performance in development experiments.

is, in general we assume domain-specific native writing is always available). This enables each readability prediction to be made relative to the domain-specific frequency of the word. We also have one set of features (L1s-ACL) that directly encodes properties of the domain of interest. Aside from these two feature types, most of the features described below focus on general properties of words. During training, it is clear that the SVR system can learn how these etymological, morphological and psychological properties influence word usage by the different L1s. However, it is worth emphasizing that while some features may seem to focus on either domain-specific or L1-specific patterns, all the features are always weighted according to the readability scores observed during training. And since the training data reflects both domain-specific and L1-specific preferences, all features are ultimately weighted in light of how both of these parameters affect word usage. For example, the SVR system could learn that in a particular domain, a particular L1-group tends to use Latinate words (reflected in the ETYM features) or tends to use a more British style of writing (CNTS features). Features that capture general L1 preferences can therefore also be used to capture domain effects and vice versa.

### 5.1. *Etymological Features:* ETYM

These features encode the intuition from Section 3: a reader might readily understand an English word if that word is cognate with a word in their L1. Uitdenbogerd (2005) found that manually-identified cognates "slightly improve" a readability measure for L1-English readers of French, while Burstein *et al.* (2007) mark cognates as part of their text adaptation toolkit. Nicolai *et al.* (2013) recently explored cognate features for the related task of native-language identification.

Cognates arise through both common ancestors and via word borrowings. To capture the ancestral relationships, we used the etymological dictionary at `www.etymonline.com`. For a predictor $f^i$ for L1=$i$, we include a feature if the English word originated from language $i$. We also include a feature if the word is of Latin origin.[4]

We also extracted cognates from bilingual dictionaries between English and the given L1. We use in-house electronic bilingual dictionaries collected over a number of years from sources such as Freelang and Wiktionary. These dictionaries include orthographic but not phonetic information, although the latter is clearly a useful source of information for automatic cognate identification (Kondrak, 2001; Kondrak and Sherif, 2006). We have a feature for the (normalized) *string edit similarity* between the English word and its translation in the L1 (the string edit distance divided by the length of the longer word, with this quotient subtracted from 1). We also look for words that might be "false friends" with an L1 word (and therefore difficult for those L1 readers); we have a feature for the *average edit distance* between the English

4. More detailed features (with date of origin, etc.) did not improve performance in development experiments.

word and its 10-most-similar *non*-translations in the L1. For example, for L1=French, the scores identify the word *compute* as a false friend – e.g. with French words like *compte* (*account*). On the other hand, the English word *likelihood* is unlike any French word in our dictionaries, and has a low false-friend score. [5]

Finally, we have a feature for the *translation ambiguity* of the word; we simply count how many different translations that word has in the target L1. E.g., the word *clear* has 40 translations in our Chinese-English dictionary, while, ironically, the word *confusing* has only 4.

### 5.2. *Morphological Features:* MORF

In development, we found only 3 types of morphological features useful for readability prediction: (1) the number of characters in the word, (2) the two-character suffix of the word (capturing some syntax e.g. *-ly* or *-ng*), and (3) whether the word begins with the prefix *re-* or *un-*.

### 5.3. *Lexical Resource Features:* LEXI

This group of features encodes information derived from psycholinguistic lexical research. One important repository for such data is the English Lexicon Project (Balota *et al.*, 2007), which contains results of human lexical decision and speeded naming experiments for over 40,000 English words. For example, humans are shown a string of letters and "asked to press one button if the string is a word and another button if the string is a nonword." This is referred to as the *lexical naming* task. For each word in our data, we include features giving the average human participant's (a) reaction time and (b) accuracy on this task.

We also include features for two indicators that have been found predictive in word recognition research. First, we include a feature for the age-of-acquisition of each English word, leveraging Kuperman *et al.* (2012)'s ratings for over 30,000 words. Secondly, we include a feature for the orthographic similarity between the given English word and other English words, via Yarkoni *et al.* (2008)'s OLD20 scores. OLD20

5. As an interesting aside, note that false friends between pairs of languages can arise because of accidental orthographic or phonetic similarities, or due to shifts in meaning of words with common ancestors. Languages that have a closer historical relationship may have both more similarities in spelling and pronunciation, as well as more cognates, and thus one might expect there to be a higher rate of false friends between such languages. However, a quick search of the literature yielded no formal evidence for such a trend. It may therefore be useful to report that our computed false friend scores were significantly higher for languages that have a closer relationship with English (such as French and Dutch) and lower for Chinese and Japanese. In the end, of course, these inter-language differences are not germane to our central idea; ultimately, our learned systems will determine how much the false friend scores affect word difficulty on a language-by-language and domain-by-domain basis.

scores capture the intra-language presence of similar words to a target word, taking the average edit distance between a target word and its 20 closest orthographic neighbours; these similar words might interfere with recognition and use of the target. When a target word is not present in one of the above data sets, we simply omit the feature from the model; we found this worked better than having a separate missing-value indicator feature.

### 5.4. *Count Features:* CNTS

Some of our most innovative features capture the frequency with which the target words are used in other text domains. If L1 has an impact on the word preferences of L2-English writers, then one would expect many words to consistently be used more or less frequently than average in the L2 writing of different L1 populations, regardless of the domain. For example, if a common function word is used relatively infrequently in the Chinese portion of the ICLE learner corpus, we might expect that word to also be infrequent in other domains. For each predictor $f^i$ for L1=$i$, we therefore include a feature for the log-frequency of the target word in the L1=$i$ portion of the ICLE. We also include a feature for the log-frequency of the target word in the entire, combined ICLE (designed to capture preferences universal to all English L2 learners).

We also developed a technique to estimate the count of our words in online English documents (web pages, PDFs, etc.) that are likely to be written by speakers of specific L1s. We simply note the number of pages retrieved by an Internet search engine for a query restricted to one of the top-level domains `.nl`, `.fr`, `.de`, `.it`, `.es`, `.jp` and `.cn`.[6] We query for our target word as well as some common English words to ensure the returned pages are in English. For example, to count the frequency of *presumably* on pages that might be authored by native-Dutch speakers, we issue the query "***presumably*** *the and a of site:.nl*" and note how many pages are returned. Analogous to our ICLE features, we provide web-search features for the log-frequency of the target word in the L1=$i$ top-level domain, and also a feature for the log-frequency of the word using the pooled counts of all our foreign domains.

We also investigate whether the Europarl (Koehn, 2005) parallel corpus can provide useful lexical statistics for our task. The English portion of Europarl helpfully indicates the original language of all the utterances (i.e., originally-in-English or translated-to-English from language $i$). For each predictor $f^i$ where L1=$i$ is a European language, we include a feature for the log-count of each word in the portions of Europarl translated to English from original language $i$. These translations might

6. We use `blekko.com` as our search engine; Blekko provides a publicly-available search-API. Note also that independently and in parallel with our own work, Cook and Hirst (2012) addressed the question of whether top-level Internet domains can provide representative text for different groups of *native* speakers of English. Together with our own positive results, this technique shows promise as a low-cost tool for building large and low-cost world-English corpora of both native and non-native varieties.

be helpful in that translated text is known to reflect L1 preferences; that is, translations often directly echo the words in the text to be translated, as opposed to what a native speaker would generate on their own (Koppel and Ordan, 2011). Note that we again also have features for the log-count of the word across all the non-native portions of English Europarl (i.e., excluding those originally in English).

Finally, we explored the use of counts from domains that are otherwise unconnected with a target L1; we presently describe each domain and explain our motivation for using counts from each of them. First, we provide features for the log-count of words in the British National Corpus. [7] Our hypothesis is that some L2 writing will reflect a preference for a British writing style. We also provide the log-frequencies obtained from (1) an English web-scale N-gram corpus from Google (Lin *et al.*, 2010), and (2) articles in the Xinhua section of the AQUAINT corpus (Vorhees, 2002) (Xinhua is the official news agency of the People's Republic of China, and hence may somewhat reflect Chinese lexical preferences). Finally, each predictor also has features for the log-frequency of words in different components of the International Corpus of English (Greenbaum and Nelson, 1996). [8] In development experiments with the ICE, we saw gains when using count features derived from its sections for (1) written and spoken Canadian English, (2) written United States English, (3) written Hong Kong English, and (4) written Singapore English. The idea is that some of our target L1s speak English either similarly or dissimilarly to how English is spoken in Canada, the United States, Hong Kong, and Singapore, and our learned system can exploit these relationships to make better word difficulty predictions.

### 5.5. *Other L1-Based ACL Counts: L1s-*ACL

Recall that the readability scores that we are trying to predict are based on a log-odds-ratio; this ratio involves the frequency of the target word in ACL papers written by a particular L1. We assume these frequencies are *unknown* for our test words. Our final set of features are based on the idea: What if we have access to word frequencies from ACL papers of *other* L1s? For example, if we knew a word was used relatively often in German ACL papers, could that help us predict the readability scores in Dutch or French? While we ultimately wish to predict readability of words in domains where only native-English text is available, there are cases, like academic domains, where we might have extensive frequency information for certain L1s (like Japanese or German) but not others (such as Vietnamese or Hungarian). We exploit this intuition by including features in each $f^i$ for the count of the target word in each of our seven sets of L1-divided ACL data (e.g. Spanish, Italian, German, etc.), excluding counts for language $i$ (the L1 currently under consideration).

7. BNC counts were obtained directly from the data hosted at `www.kilgarriff.co.uk/bnc-readme.html`

8. The ICE comprises writing by *native* speakers of English and should not be confused with the IC*LE*.

## 6. Experiments

The aim of our experiments is to test our ability to predict the readability of unseen words, using various types of feature information. We compare our approach to a variety of baseline systems (Section 6.2) using two evaluation measures described below (Section 6.1).

### 6.1. *SVR Software and Performance Measures*

For evaluating our support vector regression technique, we use an efficient implementation of $\epsilon$-SVR that exists as part of the SVM$^{light}$ toolkit (Joachims, 1999). We train the predictor on the training examples and optimize the regularization parameter ($C$) on the development data, choosing the model that performs best over the range $C$=$10^{-4}, 10^{-3}...10^{3}$. We also investigated using different kernels (with different parameters) and different $\epsilon$ values, but ultimately went with the default settings which performed well.

We evaluate using two performance measures:

1) Mean Squared Error (MSE): the average squared difference between the true log-odds-ratio and that predicted by our system;

2) %-Closer: percentage of instances where the trained system is closer than the *English ACL* baseline (see Section 6.2).

We also evaluate over two different data collections. In the *ALL* setting, we compute our performance measures as an average over all seven target-L1 test sets in the ACL data. In the *CJG* setting, we restrict our measures to only those target L1s where we have more than a million words of data: Chinese, Japanese, and German. With more data, the log-odds-ratios are more likely to reflect true preferences.

### 6.2. *Comparison Approaches*

Our comparison approaches provide simple methods for estimating the target log-odds-ratios; they allow us to to determine what can be gained by using our more complex trained models. Each comparison approach uses the following strategy: we estimate the log-odds-ratios using the English ACL frequencies for $O_1(w)$ (see Equation 1), but we estimate $O_2(w)$ (the non-native odds) using some other set of frequencies.

First, we estimate $O_2(w)$ using frequencies from the portion of the ICLE that has the corresponding L1. One would expect these estimates to be quite good if word frequencies in the ICLE reflect consistent patterns of usage among different L1 populations, *regardless of domain*. We refer to these estimates as ***TargetL1-ICLE*** in our results.

Historically, word readability scores have been based on general corpus frequencies, uncustomized for domain or L1 (Klare, 1974). We replicate this standard approach by calculating $O_2(w)$ using frequencies from our web-scale Google N-gram Corpus (referred to as ***English Google***).

We also consider two domain-specific approaches. First, we calculate $O_2(w)$ using the same English ACL frequencies that we used for $O_1(w)$. In practice, this results in a log-odds-ratio of **zero** for every word. We refer to this as ***English ACL*** in our results. We shall see that this simple approach is a very competitive baseline for our task. Secondly, we calculate $O_2(w)$ from the pooled set of all our non-native ACL data, excluding the L1 under consideration. We refer to this as **L1s-ACL Combo**. As with our predictor using L1s-ACL *features*, this baseline represents an idealized setting where large amounts of in-domain non-native data are assumed to be available. Of course, the comparison approach described here differs from the learned model using L1s-ACL features in that the learned model can weight some languages differently than others in the prediction. We will see that this can make a big difference in terms of performance.

Finally, we devised an approach to provide a kind of upper bound on the performance we might expect on this task. We start by randomly dividing the papers in each of our L1-ACL data sets into two equal halves. We then calculate the log-odds-ratios separately on both halves, for every word in our test set. We then calculate the MSE between the two halves for each test word. We repeated this procedure 100 times for each L1 and averaged the results to give us our final ***Oracle**** score. The Oracle* score gives us an indication of how accurately we can predict our readability scores given both in-domain and L1-specific data. Note that while the Oracle* considers the same words as those used in our other predictions, they are computed from different data, i.e. 50% samples of the data used to compute the true target scores. However, we shall see that the Oracle* performance can still be quite instructive.

## 7. Results

### 7.1. *Main Results*

**The Importance of Domain:** Table 3 gives the MSE for our predictors and comparison systems. Looking first at the comparison approaches, a key take-home message of our work is that domain has a much bigger impact on word usage than L1. The *TargetL1-ICLE* predictions (using data matched for L1) are much worse than the predictions of *English ACL* (using data matched for domain). This is especially important in light of the growing body of work designing automatic *classifiers* for author-L1 (Koppel *et al.*, 2005; Tsur and Rappoport, 2007; Wong and Dras, 2011). Since most of these classifiers are trained on ICLE data, it is unlikely they would transfer to other domains where, as we have demonstrated, non-native authors have very different preferences, even on function words.

| | | Stopwords | | Content | |
|---|---|---|---|---|---|
| | | ALL | CJG | ALL | CJG |
| **Comparison Systems** | Oracle* | 0.22 | 0.16 | 0.42 | 0.21 |
| | *TargetL1-ICLE* | 2.16 | 2.53 | 2.40 | 2.57 |
| | *English Google* | 1.94 | 2.18 | 1.73 | 1.88 |
| | *English ACL* | 0.43 | 0.48 | 0.40 | 0.45 |
| | L1s-ACL Combo | **0.27** | **0.20** | **0.34** | **0.32** |
| **Trained Systems** | $f$(ETYM) | 0.41 | 0.44 | 0.38 | **0.39** |
| | $f$(MORF) | 0.41 | 0.43 | 0.38 | **0.39** |
| | $f$(LEXI) | 0.41 | 0.44 | 0.38 | 0.40 |
| | $f$(CNTS) | **0.36** | **0.38** | **0.37** | 0.40 |
| | $f$(L1s-ACL) | **0.22** | **0.20** | **0.28** | **0.26** |

**Table 3.** *Mean-squared error for the Oracle*, comparison systems, and trained SVR models*

| | | Stopwords | | Content | |
|---|---|---|---|---|---|
| | | ALL | CJG | ALL | CJG |
| **Comparison** | L1s-ACL Combo | 59.7 | 64.4 | 54.1 | 57.3 |
| **Trained Systems** | $f$(ETYM) | 52.8 | 55.7 | 55.4 | 58.1 |
| | $f$(MORF) | 55.0 | 53.8 | 55.9 | 60.3 |
| | $f$(LEXI) | 54.7 | 52.3 | 55.2 | 58.1 |
| | $f$(CNTS) | 59.1 | 58.0 | 53.1 | 53.0 |
| | $f$(L1s-ACL) | **65.9** | **68.6** | **60.1** | **62.7** |

**Table 4.** *Percentage of words where predictions are closer than the predictions of the baseline* English ACL *model*

Turning back to Table 3, we see that the *English ACL* approach also strongly outperforms the traditional approach (*English Google*) that uses open-domain word frequencies. Taken together with the *TargetL1-ICLE* performance, these results provide strong motivation for making all readability measures sensitive to the domain of the writing. Finally, the best comparison approach is L1s-ACL Combo, which shows what can be achieved if domain-specific data is available from other, distinct L2 populations.

**Success of Learned Models:** Turning our attention to the learned predictors ($f([features])$), we see that all approaches are able to achieve lower MSE than the *English ACL* baseline. Table 4 gives the percentage of cases where the trained model is closer to the true log-odds-ratio than the *English ACL* baseline. The fact these scores are all well above 50% is an important result: we have shown that models trained on

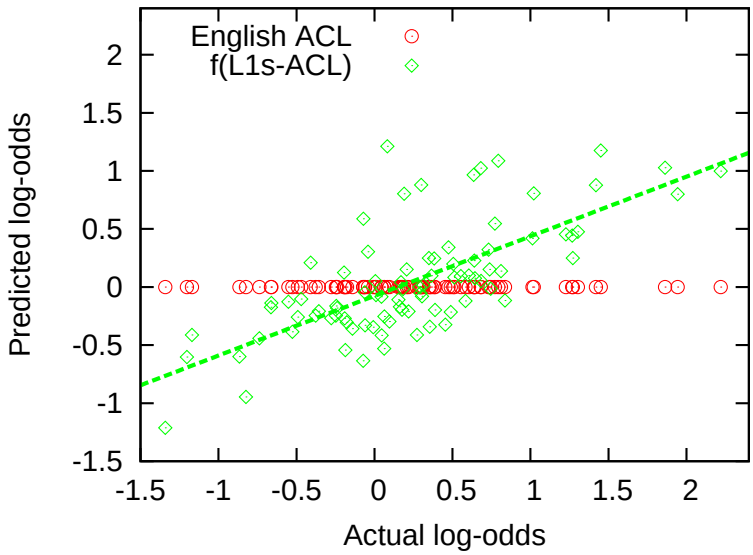


**Figure 1.** *Baseline model scores (red squares) and learned predictions (green diamonds) for* Japanese *words in the ACL data*

some words are able to generalize from the feature information to make good predictions on new, unseen words. In all cases, the best learned predictor is $f$(L1s-ACL), which leverages the counts from specific, distinct ACL-L1 populations (an idealized setting for readability prediction, as discussed above). Indeed its MSE results are close to, and sometimes superior to, the results of the Oracle* (although on the more robust CJG scores, $f$(L1s-ACL) remains lower).

**Results on Stopwords vs. Content Words:** Interestingly, of the non-idealized predictors, $f$(CNTS) performs best on stopwords, but is not as helpful on content words. The count features, based on statistics from other domains and corpora, are more useful on stopwords because stopword usage is less dependent on domain (although domain still plays a surprisingly large role in stopword usage, as is indicated by the large gap in performance between the ACL-specific and baseline systems as noted above). For content words, clues like morphology and etymology provide the best information. Note we also tried models that combined our different feature types, but unfortunately observed no improvement on development data. This suggests that the readability scores for certain words can be recovered with learned predictions using any features, but other words are difficult to handle for any feature type.

Figure 1 provides some more insight into our models and their errors. The figure plots, for an L1 of Japanese, the actual log-odds-ratios on the $x$-axis and predicted scores on the $y$-axis (using the *English ACL* and $f$(L1s-ACL) predictors). A perfect

| Class | Type | **Jap.** | **Chin.** | **Spa.** | **Ital.** | **Fr.** | **Ger.** | **Dut.** |
|---|---|---|---|---|---|---|---|---|
| ETYM | Transl. string-sim. | -0.01 | 0.02 | -0.02 | -0.00 | **-0.05** | -0.01 | 0.01 |
| ETYM | Latin origin | -0.01 | -0.01 | 0.02 | 0.02 | **0.03** | 0.00 | 0.00 |
| LEXI | Age of acquisition | **0.01** | 0.00 | 0.00 | 0.00 | -0.00 | 0.01 | 0.00 |
| LEXI | Lex.-Naming Acc. | -0.49 | **-0.58** | -0.43 | -0.20 | -0.28 | -0.27 | -0.28 |
| MORF | Num. characters | -0.06 | -0.04 | -0.05 | **-0.09** | -0.03 | -0.04 | -0.01 |
| CNTS | Freq-ICLE *[Target-L1]* | -0.04 | -0.02 | **-0.18** | -0.04 | -0.16 | -0.18 | -0.09 |
| CNTS | Freq-Web *[Target]* domain | -0.07 | **-0.98** | -0.44 | -0.12 | -0.29 | -0.28 | 0.11 |
| CNTS | Freq-Europarl *[Target]* | N/A | N/A | 0.01 | -0.07 | -0.07 | -0.15 | **-0.26** |
| CNTS | Freq-Xinhua News | -0.20 | **-0.20** | -0.18 | -0.01 | -0.10 | -0.03 | -0.05 |
| CNTS | Freq-ICE, HongKong-Wri. | **-0.13** | -0.07 | 0.13 | 0.05 | -0.02 | -0.05 | 0.10 |
| CNTS | Freq-ICE, Canadian-Writ. | 0.31 | **0.48** | 0.17 | 0.05 | 0.27 | 0.18 | 0.16 |
| CNTS | Freq-ICE, Canadian-Spok. | 0.05 | 0.06 | 0.12 | **0.14** | 0.12 | 0.13 | 0.12 |
| L1s-ACL | Freq-Jap. ACL | N/A | **-0.54** | 0.17 | -0.27 | -0.10 | 0.07 | 0.10 |
| L1s-ACL | Freq-Chin. ACL | **-0.44** | N/A | -0.33 | 0.17 | 0.04 | -0.05 | -0.18 |
| L1s-ACL | Freq-Spa. ACL | 0.08 | **-0.22** | N/A | -0.21 | -0.03 | 0.01 | -0.04 |
| L1s-ACL | Freq-Ital. ACL | -0.21 | 0.06 | **-0.25** | N/A | -0.18 | -0.17 | 0.06 |
| L1s-ACL | Freq-Fr. ACL | -0.05 | -0.00 | -0.02 | -0.13 | N/A | -0.13 | **-0.18** |
| L1s-ACL | Freq-Ger. ACL | 0.14 | -0.10 | -0.14 | -0.31 | -0.26 | N/A | **-0.32** |
| L1s-ACL | Freq-Dut. ACL | 0.05 | **-0.27** | -0.03 | -0.02 | -0.18 | -0.19 | N/A |

**Table 5.** *Predictive features for each language on the stopword data. Positive feature weights (in green) predict L2 difficulty; negative weights (in red) predict L2 readability*

set of predictions would result in the line $y = x$. The *English ACL* predictor predicts zero for each score, and suffers much of its error in the extremely low and high actual scores. The $f$(L1s-ACL) predictor does a better job on the extreme points, predicting the correct *sign* of the score for almost every point outside of $\pm 0.5$. However, many of the actual log-odds are close to zero, and it might be difficult to model these small log-odds-ratios accurately, regardless of the features.

### 7.2. *Feature Analysis*

Another key objective of our work is to better understand what specific kinds of information are helpful in automated readability measures. We can get some intuition about this by looking at the weights assigned to features after SVR training. More

predictive features will generally get higher weights, and the sign of the weights will tell us if a feature is being used the way we expected. To this end, Table 5 gives the weights on some of the highest-weighted features in each category, for each L1, for models trained on the stopwords.

For ETYM features, we see that a word generally gets a higher readability score if it is cognate with English (i.e. high translation string-similarity), but the effect is only pronounced for Spanish and French. On the other hand, the opposite is true for words of Latin origin; the log-odds-ratio/word-difficulty is *increased* for Spanish and French for Latinate words. This is intuitive in the sense that translations with high string similarity are cognates with the same form and *meaning*. The meaning of Latinate words, on the other hand, can diverge from their ancestors over time (e.g., *librairie* in French means *bookstore* as opposed to *library*). Such divergences from their native language may deter non-natives from using such words.

For LEXI features, we see that age-of-acquisition has no effect on readability in this domain. Lexical naming accuracy, on the other hand, is highly predictive of non-native readability. This again shows a connection between readability and frequency: words that are easy to recognize (and read) are used more frequently by non-native speakers.

For MORF features, we see a somewhat counter-intuitive effect: longer stopwords are used *relatively* more frequently by non-native speakers (compared to native speakers). It has long been observed that function-word frequency is a decreasing function of word length (Miller *et al.*, 1958), and this remains true for both the native and non-native stopwords, however it appears that this effect is relatively stronger in native writing.

Next, we make some observations about the CNTS features. First, note that our L1-specific counts (ICLE, Search-engine [Web] hits, and Europarl) work as expected: if a word has a higher count in one of these L1-specific corpora, that word gets a higher readability score in the ACL data. Thus L1 statistics from other domains *can be useful for modeling readability*, if these statistics are weighted appropriately (as opposed to being trusted entirely as in the *TargetL1-ICLE* system). A high frequency in Xinhua news also correlates with readability, especially for Chinese and Japanese speakers, while a high count in the Hong Kong ICE predicts readability for Chinese/Japanese speakers but difficulty for Spanish/Dutch ones. Furthermore, frequency in the Canadian ICE is universally predictive of difficulty for non-natives. This observation could perhaps be partly attributed to the fact that, as the language of a colony as opposed to an imperial power, Canadian English (unlike U.S. or British English) has had little impact on world English usage, and is therefore somewhat idiosyncratic to native English speakers.

Finally, observe that for the L1s-ACL features, the feature weights seem to reflect geographic or ancestral relationships between the L1 languages. That is, counts in the native-Chinese ACL data are predictive of readability for native-Japanese speakers, and vice versa, while German and Dutch are also mutually indicative. These results

motivate the development of new readability models for L1s that do not currently have sufficient data to create gold standard log-odds-ratios. For example, by learning from a small training set, we might learn how to exploit the numerous German and Dutch papers to better rate word difficulty for Norwegian speakers, or leverage the popular Romance languages to better rate difficulty for Portuguese or Romanian readers.

## 8. Related Work

Most prior work on readability assessment has focused on labeling documents with an appropriate reading level (e.g. a school grade level). The dominant approach has long been a two-variable formula, with one variable for word difficulty and one for sentence difficulty, with the word-difficulty variable being the more predictive component (Klare, 1974). Recent research has shown that for "an educated adult audience" the traditional formulas do not correlate well with human readability ratings (Pitler and Nenkova, 2008). Pitler and Nenkova (2008) combine lexical, syntactic and discourse features in a linear regression for readability. Schwarm and Ostendorf (2005) use an SVM classifier to combine readability indicators, while Collins-Thompson *et al.* (2005) and Heilman *et al.* (2007) use language-models trained on texts marked for reading level. Feng *et al.* (2009) use a linear regression model to assess readability for adults with intellectual disabilities.

While much of the prior work on readability has looked at the relationship between grade level and readability, few studies have considered native-language as a component of personalized readability measures. Greenfield (1999) investigated ways to scale the traditional two-variable readability measures for Japanese readers. Crossley *et al.* (2008) used Greenfield's data to show that models trained with lexical, syntactic and discourse features could improve on the traditional approaches to readability. Crossley and McNamara (2009) investigated differences in L1 and L2 writing using the Spanish section of the ICLE (and a matched English corpus), and confirmed previous findings that L2 writers are "lexically less proficient" and have "less lexical variation and sophistication in their writing." Rosa and Eskenazi (2011) show the effect of two simple measures of phonetic and semantic complexity on the acquisition of individual words by L2 learners (undifferentiated for L1).

Of course, L2 acquisition can depend on many factors of the reader beyond their L1, such as their age, previous language background, and order of acquisition (Rosa and Eskenazi, 2011) as well as psychological factors such as their world and topic knowledge, word-decoding accuracy and speed, etc (Zakaluk and Samuels, 1988). Our efforts at incorporating domain knowledge and L1 into trained readability measures can be viewed as steps toward better personalized readability tools. Zakaluk and Samuels (1988) designed a readability-assessment tool that considers both a text's difficulty as well as the reader's knowledge of the topic and general word recognition skill. Miltsakaki and Troutt (2008) also consider the domain knowledge of readers; they collect word counts in different "thematic areas" and let the user profile affect how these counts are used to assess difficulty. Collins-Thompson *et al.* (2011) es-

timate a user's reading proficiency based on the user's "current and past [Internet] search behavior."

Like us, other researchers have used the ACL Anthology to study writing differences between different groups of academic authors. For example, differences have been noted in papers divided according to individual authors (Johri *et al.*, 2011; Feng *et al.*, 2012), author gender (Sarawgi *et al.*, 2011; Bergsma *et al.*, 2012; Vogel and Jurafsky, 2012), likelihood of future citation (Yogatama *et al.*, 2011), and native language (Bergsma *et al.*, 2012; Post and Bergsma, 2013). Bergsma *et al.* (2012) study syntactic and lexical differences between native-English-speaking and L2-English writing in the ACL Anthology as part of a classification task, but do not differentiate non-native writing according to the specific L1. Post and Bergsma (2013) recently explored the prediction of author L1 in the ACL Anthology as a sub-task in a paper comparing different strategies for exploiting syntax for text classification. Dale and Kilgarriff (2010) developed a shared task on editing the writing of papers by non-native speakers.

There is also a large body of work on *correcting* errors in L2 writing, with a specific focus on difficulties in preposition and article usage (Han *et al.*, 2006; Chodorow *et al.*, 2007; Felice and Pulman, 2007; Tetreault and Chodorow, 2008; Gamon, 2010).

## 9. Conclusion and Future Work

We have introduced a statistical framework for training readability measures based on support vector regression, and we proposed and evaluated a range of creative features for our model. It is possible to accurately predict an unseen word's readability based on these features. We have shown that domain knowledge is more important than L1 for readability, but both can be important and effective components of a system that weights them appropriately. We also showed that frequency-features were most helpful for function words, while etymological, morphological and psychological information is more helpful on content words. Our feature analysis also showed how the value of features can depend on L1. It is clear that a monolithic approach to L2 readability will fail to consider the important and regular differences between L1s, and thus will not achieve optimal readability-prediction accuracy.

In future work, we are interested in extending our analysis to see if *syntactic* or *discourse* differences between L1s can also be predicted using a machine learning approach. Syntax in particular has previously been shown to play an important role in L2 readability (Heilman *et al.*, 2007). We are also interested in extending our predictions into other domains such as social media, where user ethnicity has been shown to correlate with linguistic style (Eisenstein *et al.*, 2011; Rao *et al.*, 2011). Finally, our results can immediately help systems choose the best lexical simplifications for specific L1 populations (Yatskar *et al.*, 2010; Biran *et al.*, 2011) or guide the substitution of synonyms in teacher-support tools (Burstein *et al.*, 2007).

## 10. References


Balota D., Yap M., Hutchison K., Cortese M., Kessler B., Loftis B., Neely J., Nelson D., Simpson G., Treiman R., "The English Lexicon Project", *Behavior Research Methods*, vol. 39, p. 445-459, 2007.

Bergsma S., Post M., Yarowsky D., "Stylometric Analysis of Scientific Articles", *Proc. NAACL-HLT*, p. 327-337, 2012.

Biran O., Brody S., Elhadad N., "Putting it Simply: a Context-Aware Approach to Lexical Simplification", *Proc. ACL*, p. 496-501, 2011.

Breland H. M., "Word Frequency and Word Difficulty: A Comparison of Counts in Four Corpora", *Psychological Science*, vol. 7, n° 2, p. 96-99, 1996.

Burstein J., Shore J., Sabatini J., Lee Y.-W., Ventura M., "The Automated Text Adaptation Tool", *Proc. NAACL-HLT*, p. 3-4, 2007.

Chall J. S., "The Beginning Years", *in* B. L. Zakaluk, S. J. Samuels (eds), *Readability: Its Past, Present and Future*, Newark: International Reading Association, p. 2-13, 1988.

Chodorow M., Tetreault J. R., Han N.-R., "Detection of Grammatical Errors Involving Prepositions", *Proc. ACL-SIGSEM Workshop on Prepositions*, p. 25-30, 2007.

Collins-Thompson K., Bennett P. N., White R. W., de la Chica S., Sontag D., "Personalizing Web Search Results by Reading Level", *Proc. CIKM*, p. 403-412, 2011.

Collins-Thompson K., Callan J., "Predicting Reading Difficulty with Statistical Language Models", *Journal of the American Society for Information Science and Technology*, vol. 56, n° 13, p. 1448-1462, 2005.

Cook P., Hirst G., "Do Web-Corpora from Top-Level Domains Represent National Varieties of English?", *Proc. 11th International Conference on the Statistical Analysis of Textual Data*, 2012.

Crossley S. A., Greenfield J., Mcnamara D. S., "Assessing Text Readability Using Cognitively Based Indices", *TESOL Quarterly*, vol. 42, n° 3, p. 475-493, 2008.

Crossley S. A., McNamara D. S., "Computational Assessment of Lexical Differences in L1 and L2 Writing", *Journal of Second Language Writing*, vol. 18, n° 2, p. 119-135, 2009.

Dale R., Kilgarriff A., "Helping our Own: Text Massaging for Computational Linguistics as a New Shared Task", *Proc. 6th International Natural Language Generation Conference*, p. 261-265, 2010.

Dolch E. W., "Testing Word Difficulty", *The Journal of Educational Research*, vol. 26, n° 1, p. 22-27, 1932.

Eisenstein J., Smith N. A., Xing E. P., "Discovering Sociolinguistic Associations with Structured Sparsity", *Proc. ACL*, p. 1365-1374, 2011.

Felice R. D., Pulman S. G., "Automatically Acquiring Models of Preposition Use", *Proc. ACL-SIGSEM Workshop on Prepositions*, p. 45-50, 2007.

Feng L., Elhadad N., Huenerfauth M., "Cognitively Motivated Features for Readability Assessment", *Proc. EACL*, p. 229-237, 2009.

Feng S., Banerjee R., Choi Y., "Characterizing Stylistic Elements in Syntactic Structure", *Proc. EMNLP-CoNLL*, p. 1522-1533, 2012.

Gamon M., "Using Mostly Native Data to Correct Errors in Learners' Writing: a Meta-Classifier Approach", *Proc. HLT-NAACL*, p. 163-171, 2010.

Gough P. B., "Word Recognition", *in* P. Pearson (ed.), *Handbook of Reading Research*, New York: Longman, 1984.

Granger S., Dagneaux E., Meunier F., Paquot M., "The International Corpus of Learner English. Version 2. Handbook and CD-Rom", 2009.

Greenbaum S., Nelson G., "The International Corpus of English (ICE) Project", *World Englishes*, vol. 15, n° 1, p. 3-15, 1996.

Greenfield G. R., Classic Readability Formulas in an EFL Context: Are They Valid for Japanese Speakers?, PhD thesis, Temple University, 1999.

Han N.-R., Chodorow M., Leacock C., "Detecting Errors in English Article Usage by Non-Native Speakers", *Nat. Lang. Eng.*, vol. 12, n° 2, p. 115-129, 2006.

Heilman M., Collins-Thompson K., Callan J., Eskenazi M., "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts", *Proc. HLT-NAACL*, p. 460-467, 2007.

Joachims T., "Making Large-Scale Support Vector Machine Learning Practical", *in* B. Schölkopf, C. Burges (eds), *Advances in Kernel Methods: Support Vector Machines*, MIT-Press, 1999.

Johri N., Ramage D., McFarland D., Jurafsky D., "A Study of Academic Collaborations in Computational Linguistics using a Latent Mixture of Authors Model", *Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, p. 124-132, 2011.

Klare G. R., "Assessing Readability", *Reading Research Quarterly*, vol. 10, n° 1, p. 62-102, 1974.

Koehn P., "Europarl: A Parallel Corpus for Statistical Machine Translation", *Proc. MT Summit X*, p. 79-86, 2005.

Kondrak G., "Identifying Cognates by Phonetic and Semantic Similarity", *Proc. NAACL*, 2001.

Kondrak G., Sherif T., "Evaluation of Several Phonetic Similarity Algorithms on The Task of Cognate Identification", *Proc. Coling-ACL Workshop on Linguistic Distances*, 2006.

Koppel M., Ordan N., "Translationese and Its Dialects", *Proc. ACL*, p. 1318-1326, 2011.

Koppel M., Schler J., Zigdon K., "Determining an Author's Native Language by Mining a Text for Errors", *Proc. KDD*, p. 624-628, 2005.

Kuperman V., Stadthagen-Gonzalez H., Brysbaert M., "Age-of-Acquisition Ratings for 30 Thousand English Words", *Behavior Research Methods*, vol. 44, n° 4, p. 978-990, 2012.

Laufer B., Nation P., "Vocabulary Size and Use: Lexical Richness in L2 Written Production", *Applied Linguistics*, vol. 16, n° 3, p. 307-322, 1995.

Laufer B., Nation P., "A Vocabulary-Size Test of Controlled Productive Ability", *Language Testing*, vol. 16, n° 1, p. 33-51, 1999.

Lin D., Church K., Ji H., Sekine S., Yarowsky D., Bergsma S., Patil K., Pitler E., Lathbury R., Rao V., Dalwani K., Narsale S., "New Tools for Web-Scale N-grams", *Proc. LREC*, p. 2221-2227, 2010.

Miller G. A., Newman E. B., Friedman E. A., "Length-Frequency Statistics for Written English", *Information and Control*, vol. 1, n° 4, p. 370-389, 1958.

Miltsakaki E., Troutt A., "Real Time Web Text Classification and Analysis of Reading Difficulty", *Proc. 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, p. 89-97, 2008.

Nicolai G., Hauer B., Salameh M., Yao L., Kondrak G., "Cognate and Misspelling Features for Natural Language Identification", *Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 140-145, 2013.

Pitler E., Nenkova A., "Revisiting Readability: A Unified Framework for Predicting Text Quality", *Proc. EMNLP*, p. 186-195, 2008.

Post M., Bergsma S., "Explicit and Implicit Syntactic Features for Text Classification", *Proc. ACL (Volume 2: Short Papers)*, p. 866-872, 2013.

Radev D. R., Muthukrishnan P., Qazvinian V., "The ACL Anthology Network Corpus", *Proc. ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, p. 54-61, 2009.

Rao D., Paul M., Fink C., Yarowsky D., Oates T., Coppersmith G., "Hierarchical Bayesian Models for Latent Attribute Detection in Social Media", *Proc. ICWSM*, p. 598-601, 2011.

Read J., "The Development of a New Measure of L2 Vocabulary Knowledge", *Language Testing*, vol. 10, n° 3, p. 355-371, 1993.

Rosa K. D., Eskenazi M., "Effect of Word Complexity on L2 Vocabulary Learning", *Proc. 6th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 76-80, 2011.

Sarawgi R., Gajulapalli K., Choi Y., "Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre", *Proc. CoNLL*, p. 78-86, 2011.

Schwarm S. E., Ostendorf M., "Reading Level Assessment Using Support Vector Machines and Statistical Language Models", *Proc. ACL*, p. 523-530, 2005.

Smola A. J., Schölkopf B., "A Tutorial on Support Vector Regression", *Statistics and Computing*, vol. 14, p. 199-222, 2004.

Tamayo J. M., "Frequency of Use as a Measure of Word Difficulty in Bilingual Vocabulary Test Construction and Translation", *Educational and Psychological Measurement*, vol. 47, n° 4, p. 893-902, 1987.

Tetreault J. R., Chodorow M., "The Ups and Downs of Preposition Error Detection in ESL Writing", *Proc. Coling*, p. 865-872, 2008.

Tilley H. C., "A Technique for Determining the Relative Difficulty of Word Meanings Among Elementary School Children", *The Journal of Experimental Education*, vol. 5, n° 1, p. 61-64, 1936.

Tsur O., Rappoport A., "Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words", *Proc. Workshop on Cognitive Aspects of Computational Language Acquisition*, p. 9-16, 2007.

Uitdenbogerd S., "Readability of French as a Foreign Language and its Uses", *Proceedings of the Australian Document Computing Symposium*, 2005.

Vogel A., Jurafsky D., "He Said, She Said: Gender in the ACL Anthology", *Proc. ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, p. 33-41, 2012.

Vorhees E., "Overview of the TREC 2002 Question Answering Track", *Proceedings of the 11th Text REtrieval Conference (TREC)*, 2002.

Wong S.-M. J., Dras M., "Exploiting Parse Structures for Native Language Identification", *Proc. EMNLP*, p. 1600-1610, 2011.

Yarkoni T., Balota D., Yap M., "Moving Beyond Coltheart's N: A New Measure of Orthographic Similarity", *Psychonomic Bulletin and Review*, vol. 15, n° 5, p. 971-979, 2008.

Yatskar M., Pang B., Danescu-Niculescu-Mizil C., Lee L., "For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia", *Proc. HLT-NAACL*, p. 365-368, 2010.

Yogatama D., Heilman M., O'Connor B., Dyer C., Routledge B. R., Smith N. A., "Predicting a Scientific Community's Response to an Article", *Proc. EMNLP*, p. 594-604, 2011.

Zakaluk B. L., Samuels S. J., "Toward a New Approach to Predicting Text Comprehensibility", *in* B. L. Zakaluk, S. J. Samuels (eds), *Readability: Its Past, Present and Future*, Newark: International Reading Association, p. 121-144, 1988.