

Revisiting Readability: A Unified Framework for Predicting Text Quality

Emily Pitler

Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
epitler@seas.upenn.edu

Ani Nenkova

Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

We combine lexical, syntactic, and discourse features to produce a highly predictive model of human readers' judgments of text readability. This is the first study to take into account such a variety of linguistic factors and the first to empirically demonstrate that discourse relations are strongly associated with the perceived quality of text. We show that various surface metrics generally expected to be related to readability are not very good predictors of readability judgments in our Wall Street Journal corpus. We also establish that readability predictors behave differently depending on the task: predicting text readability or ranking the readability. Our experiments indicate that discourse relations are the one class of features that exhibits robustness across these two tasks.

1 Introduction

The quest for a precise definition of text quality—pinpointing the factors that make text flow and easy to read—has a long history and tradition. Way back in 1944 Robert Gunning Associates was set up, offering newspapers, magazines and business firms consultations on clear writing (Gunning, 1952). In education, teaching good writing technique and grading student writing has always been of key importance (Spandel, 2004; Attali and Burstein, 2006). Linguists have also studied various aspects of text flow, with cohesion-building devices in English (Halliday and Hasan, 1976), rhetorical structure theory (Mann and Thompson, 1988) and centering the-

ory (Grosz et al., 1995) among the most influential contributions.

Still, we do not have unified computational models that capture the interplay between various aspects of readability. Most studies focus on a single factor contributing to readability for a given intended audience. The use of rare words or technical terminology for example can make text difficult to read for certain audience types (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Elhadad and Sutaria, 2007). Syntactic complexity is associated with delayed processing time in understanding (Gibson, 1998) and is another factor that can decrease readability. Text organization (discourse structure), topic development (entity coherence) and the form of referring expressions also determine readability. But we know little about the relative importance of each factor and how they combine in determining perceived text quality.

In our work we use texts from the Wall Street Journal intended for an *educated adult audience* to analyze readability factors including vocabulary, syntax, cohesion, entity coherence and discourse. We study the association between these features and reader assigned readability ratings, showing that discourse and vocabulary are the factors most strongly linked to text quality. In the easier task of text quality ranking, entity coherence and syntax features also become significant and the combination of features allows for ranking prediction accuracy of 88%. Our study is novel in the use of gold-standard discourse features for predicting readability and the simultaneous analysis of various readability factors.

2 Related work

2.1 Readability with respect to intended readers

The definition of what one might consider to be a well-written and readable text heavily depends on the intended audience (Schriver, 1989). Obviously, even a superbly written scientific paper will not be perceived as very readable by a lay person and a great novel might not be appreciated by a third grader. As a result, the vast majority of prior work on readability deals with labeling texts with the appropriate school grade level. A key observation in even the oldest work in this area is that the vocabulary used in a text largely determines its readability. More common words are easier, so some metrics measured text readability by the percentage of words that were not among the N most frequent in the language. It was also observed that frequently occurring words are often short, so word length was used to approximate readability more robustly than using a predefined word frequency list. Standard indices were developed based on the link between word frequency/length and readability, such as Flesch-Kincaid (Kincaid, 1975), Automated Readability Index (Kincaid, 1975), Gunning Fog (Gunning, 1952), SMOG (McLaughlin, 1969), and Coleman-Liau (Coleman and Liau, 1975). They use only a few simple factors that are designed to be easy to calculate and are rough approximations to the linguistic factors that determine readability. For example, Flesch-Kincaid uses the average number of syllables per word to approximate vocabulary difficulty and the average number of words per sentence to approximate syntactic difficulty.

In recent work, the idea of linking word frequency and text readability has been explored for making medical information more accessible to the general public. (Elhadad and Sutaria, 2007) classified words in medical texts as familiar or unfamiliar to a general audience based on their frequencies in corpora. When a description of the unfamiliar terms was provided, the perceived readability of the texts almost doubled.

A more general and principled approach to using vocabulary information for readability decisions has been the use of language models. For any given text, it is easy to compute its likelihood under a given lan-

guage model, i.e. one for text meant for children, or for text meant for adults, or for a given grade level. (Si and Callan, 2001), (Collins-Thompson and Callan, 2004), (Schwarm and Ostendorf, 2005), and (Heilman et al., 2007) used language models to predict the suitability of texts for a given school grade level. But even for this type of task other factors besides vocabulary use are at play in determining readability. Syntactic complexity is an obvious factor: indeed (Heilman et al., 2007) and (Schwarm and Ostendorf, 2005) also used syntactic features, such as parse tree height or the number of passive sentences, to predict reading grade levels. For the task of deciding whether a text is written for an adult or child reader, (Barzilay and Lapata, 2008) found that adding entity coherence to (Schwarm and Ostendorf, 2005)'s list of features improves classification accuracy by 10%.

2.2 Readability as coherence for competent language users

In linguistics and natural language processing, the text properties rather than those of the reader are emphasized. Text coherence is defined as the ease with which a person (tacitly assumed to be a competent language user) understands a text. Coherent text is characterized by various types of cohesive links that facilitate text comprehension (Halliday and Hasan, 1976).

In recent work, considerable attention has been devoted to entity coherence in text quality, especially in relation to information ordering. In many applications such as text generation and summarization, systems need to decide the order in which selected sentences or generated clauses should be presented to the user. Most models attempting to capture local coherence between sentences were based on or inspired by centering theory (Grosz et al., 1995), which postulated strong links between the center of attention in comprehension of adjacent sentences and syntactic position and form of reference. In a detailed study of information ordering in three very different corpora, (Karamanis et al., to appear) assessed the performance of various formulations of centering. Their results were somewhat unexpected, showing that while centering transition preferences were useful, the most successful strategy for information ordering was based on avoid-

ing rough shifts, that is, sequences of sentences that share no entities in common. This supports previous findings that such types of transitions are associated with poorly written text and can be used to improve the accuracy of automatic grading of essays based on various non-discourse features (Miltsakaki and Kukich, 2000). In a more powerful generalization of centering, Barzilay and Lapata (2008) developed a novel approach which doesn't postulate a preference for any type of transition but rather computes a set of features that capture transitions of all kinds in the text and their relative proportion. Their entity coherence features prove to be very suitable for various tasks, notably for information ordering and reading difficulty level.

Form of reference is also important in well-written text and appropriate choices lead to improved readability. Use of pronouns for reference to highly salient entities is perceived as more desirable than the use of definite noun phrases (Gordon et al., 1993; Krahmer and Theune, 2002). The syntactic forms of first mention—when an entity is first introduced in a text—differ from those of subsequent mentions (Poesio and Vieira, 1998; Nenkova and McKeown, 2003) and can be exploited for improving and predicting text coherence (Siddharthan, 2003; Nenkova and McKeown, 2003; Elsner and Charniak, 2008).

3 Data

The objective of our study is to analyze various readability factors, including discourse relations, because few empirical studies exist that directly link discourse structure with text quality. In the past, subsections of the Penn Treebank (Marcus et al., 1994) have been annotated for discourse relations (Carlson et al., 2001; Wolf and Gibson, 2005). For our study we chose to work with the newly released Penn Discourse Treebank which is the largest annotated resource which focuses exclusively on implicit local relations between adjacent sentences and explicit discourse connectives.

3.1 Discourse annotation

The Penn Discourse Treebank (Prasad et al., 2008) is a new resource with annotations of discourse connectives and their senses in the Wall Street Journal

portion of the Penn Treebank (Marcus et al., 1994). All *explicit* relations (those marked with a discourse connective) are annotated. In addition, each adjacent pair of sentences within a paragraph is annotated. If there is a discourse relation, then it is marked *implicit* and annotated with one or more connectives. If there is a relation between the sentences but adding a connective would be inappropriate, it is marked *AltLex*. If the consecutive sentences are only related by entity-based coherence (Knott et al., 2001) they are annotated with *EntRel*. Otherwise, they are annotated with *NoRel*.

Besides labeling the connective, the PDTB also annotates the *sense* of each relation. The relations are organized into a hierarchy. The top level relations are Expansion, Comparison, Contingency, and Temporal. Briefly, an expansion relation means that the second clause continues the theme of the first clause, a comparison relation indicates that something in the two clauses is being compared, contingency means that there is a causal relation between the clauses, and temporal means they occur either at the same time or sequentially.

3.2 Readability ratings

We randomly selected thirty articles from the Wall Street Journal corpus that was used in both the Penn Treebank and the Penn Discourse Treebank.¹ Each article was read by at least three college students, each of whom was given unlimited time to read the texts and perform the ratings.² Subjects were asked the following questions:

- How well-written is this article?
- How well does the text fit together?
- How easy was it to understand?
- How interesting is this article?

For each question, they provided a rating between 1 and 5, with 5 being the best and 1 being the worst.

¹One of the selected articles was missing from the Penn Treebank. Thus, results that do not require syntactic information (Tables 1, 2, 4, and 6) are over all thirty articles, while Tables 3, 5, and 7 report results for the twenty-nine articles with Treebank parse trees.

²(Lapata, 2006) found that human ratings are significantly correlated with self-paced reading times, a more direct measure of processing effort which we plan to explore in future work.

After collecting the data, it turned out that most of the time subjects gave the same rating to all questions. For competent language users, we view text readability and text coherence as equivalent properties, measuring the extent to which a text is well written. Thus for all subsequent analysis, we will use only the first question (“On a scale of 1 to 5, how well written is this text?”). The score of an article was then the average of all the ratings it received. The article scores ranged from 1.5 to 4.33, with a mean of 3.2008 and a standard deviation of .7242. The median score was 3.286.

We define our task as predicting this average rating for each article. Note that this task may be more difficult than predicting reading level, as each of these articles appeared in the Wall Street Journal and thus is aimed at the same target audience. We suspected that in classifying adult text, more subtle features might be necessary.

4 Identifying correlates of text quality

4.1 Baseline measures

We first computed the Pearson correlation coefficients between the simple metrics that most traditional readability formulas use and the average human ratings. These results are shown in Table 1. We tested *the average number of characters per word*, *average number of words per sentence*, *maximum number of words per sentence*, and *article length* (F_7).³ Article length (F_7) was the only significant baseline factor, with correlation of -0.37. Longer articles are perceived as less well-written and harder to read than shorter ones. None of the other baseline metrics were close to being significant predictors of readability.

Average Characters/Word	$r = -.0859, p = .6519$
Average Words/Sentence	$r = .1637, p = .3874$
Max Words/Sentence	$r = .0866, p = .6489$
F_7 text length	$r = -.3713, p = .0434$

Table 1: Baseline readability features

³For ease of reference, we number each non-baseline feature in the text and tables.

4.2 Vocabulary

We use a unigram language model, where the probability of an article is:

$$\prod_w P(w|M)^{C(w)} \quad (1)$$

$P(w|M)$ is the probability of word-type w according to a background corpus M , and $C(w)$ is the number of times w appears in the article.

The log likelihood of an article is then:

$$\sum_w C(w) \log(P(w|M)) \quad (2)$$

Note that this model will be biased in favor of shorter articles. Since each word has probability less than 1, the log probability of each word is less than 0, and hence including additional words decreases the log likelihood. We compensate for this by performing linear regressions with the unigram log likelihood and with the number of words in the article as an additional variable.

The question then arises as to what to use as a background corpus. We chose to experiment with two corpora: the entire Wall Street Journal corpus and a collection of general AP news, which is generally more diverse than the financial news found in the WSJ. We predicted that the NEWS vocabulary would be more representative of the types of words our readers would be familiar with. In both cases we used Laplace smoothing over the word frequencies and a stoplist.

The vocabulary features we used are *article likelihood estimated from a language model from WSJ* (F_5), and *article likelihood according to a unigram language model from NEWS* (F_6). We also combine the two likelihood features with article length, in order to get a better estimate of the language model’s influence on readability independent of the length of the article.

F_5 Log likelihood, WSJ	$r = .3723, p = .0428$
F_6 Log likelihood, NEWS	$r = .4497, p = .0127$
LL with length, WSJ	$r = .3732, p = .0422$
LL with length, NEWS	$r = .6359, p = .0002$

Table 2: Vocabulary features

Both vocabulary-based features (F_5 and F_6) are significantly correlated with the readability judgments, with p -values smaller than 0.05 (see Table 2).

The correlations are positive: the more probable an article was based on its vocabulary, the higher it was generally rated. As expected, the NEWS model that included more general news stories had a higher correlation with people’s judgments. When combined with the length of the article, the unigram language model from the NEWS corpus becomes very predictive of readability, with the correlation between the two as high as 0.63.

4.3 Syntactic features

Syntactic constructions affect processing difficulty and so might also affect readability judgments. We examined the four syntactic features used in (Schwarm and Ostendorf, 2005): *average parse tree height* (F_1), *average number of noun phrases per sentence* (F_2), *average number of verb phrases per sentence* (F_3), and *average number of subordinate clauses per sentence (SBARs in the Penn Treebank tagset)* (F_4). The sentence “We’re talking about years ago [SBAR before anyone heard of asbestos having any questionable properties].” contains an example of an SBAR clause.

Having multiple noun phrases (entities) in each sentence requires the reader to remember more items, but may make the article more interesting. (Barzilay and Lapata, 2008) found that articles written for adults tended to contain many more entities than articles written for children. While including more verb phrases in each sentence increases the sentence complexity, adults might prefer to have related clauses explicitly grouped together.

F_1 Average Parse Tree Height	$r = -.0634, p = .7439$
F_2 Average Noun Phrases	$r = .2189, p = .2539$
F_3 Average Verb Phrases	$r = .4213, p = .0228$
F_4 Average SBARs	$r = .3405, p = .0707$

Table 3: Syntax-related features

The correlations between readability and syntactic features is shown in Table 3. The strongest correlation is that between readability and number of verb phrases (0.42). This finding is in line with prescriptive clear writing advice (Gunning, 1952; Spandel, 2004), but is to our knowledge novel in the computational linguistics literature. As (Bailin and Grafstein, 2001) point out, the sentences in (1) are easier to comprehend than the sentences in (2), even

though they are longer.

- (1) It was late at night, but it was clear. The stars were out and the moon was bright.
- (2) It was late at night. It was clear. The stars were out. The moon was bright.

Multiple verb phrases in one sentence may be indicative of explicit discourse relations, which we will discuss further in section 4.6.

Surprisingly, the use of clauses introduced by a (possibly empty) subordinating conjunction (SBAR), are actually positively correlated (and almost approaching significance) with readability. So while for children or less educated adults these constructions might pose difficulties, they were favored by our assessors. On the other hand, the average parse tree height negatively correlated with readability as expected, but surprisingly the correlation is very weak (-0.06).

4.4 Elements of lexical cohesion

In their classic study of cohesion in English, (Halliday and Hasan, 1976) discuss the various aspects of well written discourse, including the use of cohesive devices such as pronouns, definite descriptions and topic continuity from sentence to sentence.⁴ To measure the association between these features and readability rankings, we compute *the number of pronouns per sentence* (F_{11}) and *the number of definite articles per sentence* (F_{12}). In order to qualify topic continuity from sentence to sentence in the articles, we compute *average cosine similarity* (F_8), *word overlap* (F_9) and *word overlap over just nouns and pronouns* (F_{10}) between pairs of adjacent sentences⁵. Each sentence is turned into a vector of word-types, where each type’s value is its tf-idf (where document frequency is computed over all the articles in the WSJ corpus). The cosine similarity metric is then:

$$\cos(s, t) = \frac{s \cdot t}{|s| |t|} \quad (3)$$

⁴Other cohesion building devices discussed by Halliday and Hansan include lexical reiteration and discourse relations, which we address next.

⁵Similar features have been used for automatic essay grading as well (Higgins et al., 2004).

F_8 Avr. Cosine Overlap	$r = -.1012, p = .5947$
F_9 Avr. Word Overlap	$r = -.0531, p = .7806$
F_{10} Avr. Noun+Pronoun Overlap	$r = .0905, p = .6345$
F_{11} Avr. # Pronouns/Sent	$r = .2381, p = .2051$
F_{12} Avr # Definite Articles	$r = .2309, p = .2196$

Table 4: Superficial measures of topic continuity and pronoun and definite description use

None of these features correlate significantly with readability as can be seen from the results in Table 4. The overlap features are particularly bad predictors of readability, with average word/cosine overlap in fact being negatively correlated with readability. The form of reference—use of pronouns and definite descriptions—exhibit a higher correlation with readability (0.23), but these values are not significant for the size of our corpus.

4.5 Entity coherence

We use the Brown Coherence Toolkit⁶ to compute entity grids (Barzilay and Lapata, 2008) for each article. In each sentence, an entity is identified as the subject (S), object (O), other (X) (for example, part of a prepositional phrase), or not present (N). The probability of each transition type is computed. For example, an S-O transition occurs when an entity is the subject in one sentence then an object in the next; X-N transition occurs when an entity appears in non-subject or object position in one sentence and not present in the next, etc.⁷ The entity coherence features are the *probability of each of these pairs of transitions*, for a total of 16 features (F_{17-32} ; see complete results in Table 5).

None of the entity grid features are significantly correlated with the readability ratings. One very interesting result is that the proportion of S-S transitions in which the same entity was mentioned in subject position in two adjacent sentences, is negatively correlated with readability. In centering theory, this is considered the most coherent type of transition, keeping the same center of attention. Moreover, the feature most strongly correlated with readability is the S-N transition (0.31) in which the subject of one sentence does not appear at all in the following sen-

⁶<http://www.cs.brown.edu/~melsner/manual.html>

⁷The Brown Coherence Toolkit identifies NPs as the same entity if they have identical head nouns.

F_{17} Prob. of S-S transition	$r = -.1287, p = .5059$
F_{18} Prob. of S-O transition	$r = -.0427, p = .8261$
F_{19} Prob. of S-X transition	$r = -.1450, p = .4529$
F_{20} Prob. of S-N transition	$r = .3116, p = .0999$
F_{21} Prob. of O-S transition	$r = .1131, p = .5591$
F_{22} Prob. of O-O transition	$r = .0825, p = .6706$
F_{23} Prob. of O-X transition	$r = .0744, p = .7014$
F_{24} Prob. of O-N transition	$r = .2590, p = .1749$
F_{25} Prob. of X-S transition	$r = .1732, p = .3688$
F_{26} Prob. of X-O transition	$r = .0098, p = .9598$
F_{27} Prob. of X-X transition	$r = -.0655, p = .7357$
F_{28} Prob. of X-N transition	$r = .1319, p = .4953$
F_{29} Prob. of N-S transition	$r = .1898, p = .3242$
F_{30} Prob. of N-O transition	$r = .2577, p = .1772$
F_{31} Prob. of N-X transition	$r = .1854, p = .3355$
F_{32} Prob. of N-N transition	$r = -.2349, p = .2200$

Table 5: Linear correlation between human readability ratings and entity coherence.

tence. Of course, it is difficult to interpret the entity grid features one by one, since they are interdependent and probably it is the interaction of features (relative proportions of transitions) that capture overall readability patterns.

4.6 Discourse relations

Discourse relations are believed to be a major factor in text coherence. We computed another language model which is over discourse relations instead of words. We treat each text as a bag of relations rather than a bag of words. Each relation is annotated for both its sense and how it is realized (implicit or explicit). For example, one text might contain {Implicit Comparison, Explicit Temporal, NoRel}. We computed the probability of each of our articles according to a multinomial model, where the probability of a text with n relation tokens and k relation types is:

$$P(n) \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \quad (4)$$

$P(n)$ is the probability of an article having length n , x_i is the number of times relation i appeared, and p_i is the probability of relation i based on the Penn Discourse Treebank. $P(n)$ is the maximum likelihood estimation of an article having n discourse relations based on the entire Penn Discourse Treebank (the number of articles with exactly n discourse relations, divided by the total number of articles).

The *log likelihood of an article based on its discourse relations* (F_{13}) feature is defined as:

$$\log(P(n)) + \log(n!) + \sum_{i=1}^k (x_i \log(p_i) - \log(x_i!)) \quad (5)$$

The multinomial distribution is particularly suitable, because it directly incorporates length, which significantly affects readability as we discussed earlier. It also captures patterns of relative frequency of relations, unlike the simpler unigram model. Note also that this equation has an advantage over the unigram model that was not present for vocabulary. While every article contains at least one word, some articles do not contain any discourse relations. Since the PDTB annotated all explicit relations and relations between adjacent sentences in a paragraph, an article with no discourse connectives and only single sentence paragraphs would not contain any annotated discourse relations. Under the unigram model, these articles’ probabilities cannot be computed. Under the multinomial model, the probability of an article with zero relations is estimated as $Pr(N = 0)$, which can be calculated from the corpus.

As in the case of vocabulary features, the presence of more relations will lead to overall lower probabilities so we also consider the *number of discourse relations* (F_{14}) and *the log likelihood combined with the number of relations* as features. In order to isolate the effect of the type of discourse relation (explicitly expressed by a discourse connective such as “because” or “however” versus implicitly expressed by adjacency), we also compute multinomial model features for the *explicit discourse relations* (F_{15}) and over just the *implicit discourse relations* (F_{16}).

F_{13} LogL of discourse rels	r = .4835, p = .0068
F_{14} # of discourse relations	r = -.2729, p = .1445
LogL of rels with # of rels	r = .5409, p = .0020
# of relations with # of words	r = .3819, p = .0373
F_{15} Explicit relations only	r = .1528, p = .4203
F_{16} Implicit relations only	r = .2403, p = .2009

Table 6: Discourse features

The likelihood of discourse relations in the text under a multinomial model is very highly and significantly correlated with readability ratings, especially after text length is taken into account. Cor-

relations are 0.48 and 0.54 respectively. The probability of the explicit relations alone is not a sufficiently strong indicator of readability. This fact is disappointing as the explicit relations can be identified much more easily in unannotated text (Pitler et al., 2008). Note that the sequence of just the implicit relations is also not sufficient. This observation implies that the proportion of explicit and implicit relations may be meaningful but we leave the exploration of this issue for later work.

4.7 Summary of findings

So far, we introduced six classes of factors that have been discussed in the literature as readability correlates. Through statistical tests of associations we identified the individual factors significantly correlated with readability ratings. These are, in decreasing order of association strength:

- LogL of Discourse Relations (r = .4835)
- LogL, NEWS (r = .4497)
- Average Verb Phrases (.4213)
- LogL, WSJ (r = .3723)
- Number of words (r = -.3713)

Vocabulary and discourse relations are the strongest predictors of readability, followed by average number of verb phrases and length of the text. This empirical confirmation of the significance of discourse relations as a readability factor is novel for the computational linguistics literature. Note though that for our work we use oracle discourse annotations directly from the PDTB and no robust systems for automatic discourse annotation exist today.

The significance of the average number of verb phrases as a readability predictor is somewhat surprising but intriguing. It would lead to reexamination of the role of verbs/predicates in written text, which we also plan to address in future work. None of the other factors showed significant association with readability ratings, even though some correlations had relatively large positive values.

5 Combining readability factors

In this section, we turn to the question of how the combination of various factors improves the prediction of readability. We use the **leaps** package in R to find the best subset of features for linear regression, for subsets of size one to eight. We use the

squared multiple correlation coefficient (R^2) to assess the effectiveness of predictions. R^2 is the proportion of variance in readability ratings explained by the model. If the model predicts readability perfectly, $R^2 = 1$, and if the model has no predictive capability, $R^2 = 0$.

$$F_{13}, R^2 = 0.2662$$

$$F_6 + F_7, R^2 = 0.4351$$

$$F_6 + F_7 + F_{13}, R^2 = 0.5029$$

$$F_6 + F_7 + F_{13} + F_{14}, R^2 = 0.6308$$

$$F_1 + F_6 + F_7 + F_{10} + F_{13}, R^2 = 0.6939$$

$$F_1 + F_6 + F_7 + F_{10} + F_{13} + F_{23}, R^2 = 0.7316$$

$$F_1 + F_6 + F_7 + F_{10} + F_{13} + F_{22} + F_{23}, R^2 = 0.7557$$

$$F_1 + F_6 + F_7 + F_{10} + F_{11} + F_{13} + F_{19} + F_{30}, R^2 = 0.776.$$

The linear regression results confirm the expectation that the combination of different factors is a rather complex issue. As expected, discourse, vocabulary and length which were the significant individual factors appear in the best model for each feature set size. Their combination gives the best result for regression with three predictors, and they explain half of the variance in readability ratings, $R^2 = 0.5029$.

But the other individually significant feature, average number of verb phrases per sentence (F_3) never appears in the best models. Instead, F_1 —the depth of the parse tree—appears in the best model with more than four features.

Also unexpectedly, two of the superficial cohesion features appear in the larger models: F_{10} is the average word overlap over nouns and pronouns and F_{11} is the average number of pronouns per sentence. Entity grid features also make their way into the best models when more features are used for prediction: S-X, O-O, O-X, N-O transitions (F_{19} , F_{22} , F_{23} , F_{30}).

6 Readability as ranking

In this section we consider the problem of pairwise ranking of text readability. That is, rather than trying to predict the readability of a single document, we consider pairs of documents and predict which one is better. This task may in fact be the more natural one, since in most applications the main concern is with the relative quality of articles rather than their absolute scores. This setting is also beneficial in

terms of data use, because each pair of articles with different average readability scores now becomes a data point for the classification task.

We thus create a classification problem: given two articles, is article 1 more readable than article 2? For each pair of texts whose readability ratings on the 1 to 5 scale differed by at least 0.5, we form one data point for the ranking problem, resulting in 243 examples. The predictors are the differences between the two articles' features. For classification, we used WEKA's linear support vector implementation (SMO) and performance was evaluated using 10-fold cross-validation.

Features	Accuracy
None (Majority Class)	50.21%
ALL	88.88%
log_l_discourse_rels	77.77%
number_discourse_rels	74.07%
N-O transition	70.78%
O-N transition	69.95%
Avg_VPs_sen	69.54%
log_l_NEWS	66.25%
number_of_words	65.84%
Grid only	79.42%
Discourse only	77.36%
Syntax only	74.07%
Vocab only	66.66%
Length only	65.84%
Cohesion only	64.60%
no cohesion	89.30%
no vocab	88.88%
no length	88.47%
no discourse	88.06%
no grid	84.36%
no syntax	82.71%

Table 7: SVM prediction accuracy, linear kernel

The classification results are shown in Table 7. When all features are used for prediction, the accuracy is high, 88.88%. The length of the article can serve as a baseline feature—longer articles are ranked lower by the assessors, so this feature can be taken as baseline indicator of readability. Only six features used by themselves lead to accuracies higher than the length baseline. These results indicate that the most important individual factors in the readability ranking task, in decreasing order of importance, are log likelihood of discourse relations, number of discourse relations, N-O transitions, O-N

transitions, average number of VPs per sentence and text probability under a general language model.

In terms of classes of features, the 16 entity grid features perform the best, leading to an accuracy of 79.41%, followed by the combination of the four discourse features (77.36%), and syntax features (74.07%). This is evidence for the fact that there is a complex interplay between readability factors: the entity grid factors which individually have very weak correlation with readability combine well, while adding the three additional discourse features to the likelihood of discourses relations actually worsens performance slightly. Similar indication for interplay between features is provided by the class ablation classification results, in which classes of features are removed. Surprisingly, removing syntactic features causes the biggest deterioration in performance, a drop in accuracy from 88.88% to 82.71%. The removal of vocabulary, length, or discourse features has a minimal negative impact on performance, while removing the coherence features actually boosts performance.

7 Conclusion

We have investigated which linguistic features correlate best with readability judgments. While surface measures such as the average number of words per sentence or the average number of characters per word are not good predictors, there exist syntactic, semantic, and discourse features that do correlate highly. The average number of verb phrases in each sentence, the number of words in the article, the likelihood of the vocabulary, and the likelihood of the discourse relations all are highly correlated with humans' judgments of how well an article is written.

While using any one out of syntactic, lexical, coherence, or discourse features is substantially better than the baseline surface features on the discrimination task, using a combination of entity coherence and discourse relations produces the best performance.

8 Acknowledgments

This work was partially supported by an Integrative Graduate Education and Research Traineeship grant from National Science Foundation (NS-FIGERT 0504487) and by NSF Grant IIS-07-05671.

We thank Aravind Joshi, Bonnie Webber, and the anonymous reviewers for their many helpful comments.

References

- Y. Attali and J. Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).
- A. Bailin and A. Grafstein. 2001. The linguistic assumptions underlying readability formulae a critique. *Language and Communication*, 21(3):285–301.
- R. Barzilay and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop*, pages 1–10.
- M. Coleman and TL Liao. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.
- K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL'04*.
- Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.
- M. Elsner and E. Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-HLT'08, (short paper)*.
- E. Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.
- P. Gordon, B. Grosz, and L. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.
- B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.
- Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill; Fourth Printing edition.
- Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Coherence in English*. Longman Group Ltd, London, U.K.
- M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of NAACL HLT*, pages 460–467.
- D. Higgins, J. Burstein, D. Marcu, and C. Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of HLT/NAACL'04*.

- N. Karamanis, M. Poesio, C. Mellish, and J. Oberlander. (to appear). Evaluating centering for information ordering using corpora. *Computational Linguistics*.
- JP Kincaid. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.
- A. Knott, J. Oberlander, M. O'Donnell, and C. Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. *Text representation: linguistic and psycholinguistic aspects*, pages 181–196.
- E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications.
- M. Lapata. 2006. Automatic evaluation of information ordering: Kendalls tau. *Computational Linguistics*, 32(4):471–484.
- W. Mann and S. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8.
- M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- G.H. McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646.
- E. Miltsakaki and K. Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of ACL'00*, pages 408–415.
- A. Nenkova and K. McKeown. 2003. References to named entities: a corpus study. In *Proceedings of HLT/NAACL 2003 (short paper)*.
- E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters and Demonstrations*, pages 85–88, Manchester, UK, August.
- M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC'08*.
- KA Schriver. 1989. Evaluating text quality: the continuum from text-focused to reader-focused methods. *Professional Communication, IEEE Transactions on*, 32(4):238–255.
- S. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL'05*, pages 523–530.
- L. Si and J. Callan. 2001. A statistical model for scientific readability. *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.
- A. Siddharthan. 2003. *Syntactic simplification and Text Cohesion*. Ph.D. thesis, University of Cambridge, UK.
- V. Spandel. 2004. *Creating writers through 6-trait writing assessment and instruction*. Allyn & Bacon.
- F. Wolf and E. Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288.