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

4.2 Vocabulary

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

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

$P(w|M)$ 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:

$$\sum_w C(w)\log(P(w|M)) \tag{2}$$

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).

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

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 sentences[5]. 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:

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

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

[5] Similar 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 Toolkit[6] 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 sentence. 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.

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.

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:

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

$P(n)$ 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).

[6] http://www.cs.brown.edu/ melsner/manual.html

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