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

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

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. Correlations 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 cohesion 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 substantally 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 (NSFIGERT 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 Liau. 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; Fouth Printing edition.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion 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*.