# Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts

**Michael J. Heilman** **Kevyn Collins-Thompson** **Jamie Callan** **Maxine Eskenazi**

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
4502 Newell Simon Hall
Pittsburgh, PA 15213-8213

{mheilman,kct,callan,max}@cs.cmu.edu

## Abstract

This work evaluates a system that uses interpolated predictions of reading difficulty that are based on both vocabulary and grammatical features. The combined approach is compared to individual grammar- and language modeling-based approaches. While the vocabulary-based language modeling approach outperformed the grammar-based approach, grammar-based predictions can be combined using confidence scores with the vocabulary-based predictions to produce more accurate predictions of reading difficulty for both first and second language texts. The results also indicate that grammatical features may play a more important role in second language readability than in first language readability.

## 1 Introduction

The REAP tutoring system (Heilman, et al. 2006), aims to provide authentic reading materials of the appropriate difficulty level, in terms of both vocabulary and grammar, for English as a Second Language students. An automatic measure of readability that incorporated both lexical and grammatical features was thus needed.

For first language (L1) learners (i.e., children learning their native tongue), reading level has been predicted using a variety of techniques, based on models of a student's lexicon, grammatical surface features such as sentence length (Flesch, 1948), or combinations of such features (Schwarm and Ostendorf, 2005). It was shown by Collins-Thompson and Callan (2004) that a vocabulary-based language modeling approach was effective at predicting the readability of grades 1 to 12 of Web documents of varying length, even with high levels of noise.

Prior work on first language readability by Schwarm and Ostendorf (2005) incorporated grammatical surface features such as parse tree depth and average number of verb phrases. This work combining grammatical and lexical features was promising, but it was not clear to what extent the grammatical features improved predictions.

Also, discussions with L2 instructors suggest that a more detailed grammatical analysis of texts that examines features such as passive voice and various verb tenses can provide better features with which to predict reading difficulty. One goal of this work is to show that the use of pedagogically motivated grammatical features (e.g., passive voice, rather than the number of words per sentence) can improve readability measures based on lexical features alone.

One of the differences between L1 and L2 readability is the timeline and processes by which first and second languages are acquired. First language acquisition begins at infancy, and the primary grammatical structures of the target language are acquired by age four in typically developing chil-

dren (Bates, 2003). That is, most grammar is acquired prior to the beginning of a child's formal education. Therefore, most grammatical features seen at high reading levels such as high school are present with similar frequencies at low reading levels such as grades 1-3 that correspond to elementary school-age children. It should be noted that sentence length is one grammar-related difference that can be observed as L1 reading level increases. Sentences are kept short in texts for low L1 reading levels in order to reduce the cognitive load on child readers. The average sentence length of texts increases with the age and reading level of the intended audience. This phenomenon has been utilized in early readability measures (Flesch, 1948). Vocabulary change, however, continues even into adulthood, and has been shown to be a more effective predictor of L1 readability than simpler measures such as sentence length (Collins-Thompson and Callan, 2005).

Second language learners, unlike their L1 counterparts, are still very much in the process of acquiring the grammar of their target language. In fact, even intermediate and advanced students of second languages, who correspond to higher L2 reading levels, often struggle with the grammatical structures of their target language. This phenomenon suggests that grammatical features may play a more important role in predicting and measuring L2 readability. That is not to say, however, that vocabulary cannot be used to predict L2 reading levels. Second language learners are learning both vocabulary and grammar concurrently, and reading materials for this population are chosen or authored according to both lexical and grammatical complexity. Therefore, the authors predict that a readability measure for texts intended for second language learners that incorporates both grammatical and lexical features could clearly outperform a measure based on only one of these two types of features.

This paper begins with descriptions of the language modeling and grammar-based prediction systems. A description of the experiments follows that covers both the evaluation metrics and corpora used. Experimental results are presented, followed by a discussion of these results, and a summary of the conclusions of this work.

## 2 Language Model Readability Prediction for First Language Texts

Statistical language modeling exploits patterns of use in language. To build a statistical model of text, training examples are used to collect statistics such as word frequency and order. Each training example has a label that tells the model the 'true' category of the example. In this approach, one statistical model is built for each grade level to be predicted.

The statistical language modeling approach has several advantages over traditional readability formulas, which are usually based on linear regression with two or three variables. First, a language modeling approach generally gives much better accuracy for Web documents and short passages (Collins-Thompson and Callan, 2004). Second, language modeling provides a probability distribution across *all* grade models, not just a single prediction. Third, language modeling provides more data on the relative difficulty of each word in the document. This might allow an application, for example, to provide more accurate vocabulary assistance.

The statistical model used for this study is based on a variation of the multinomial Naïve Bayes classifier. For a given text passage $T$, the semantic difficulty of $T$ relative to a specific grade level $G_i$ is predicted by calculating the likelihood that the words of $T$ were generated from a representative language model of $G_i$. This likelihood is calculated for each of a number of language models, corresponding to reading difficulty levels. The reading difficulty of the passage is then estimated as the grade level of the language model most likely to have generated the passage $T$.

The language models employed in this work are simple: they are based on unigrams and assume that the probability of a token is independent of the surrounding tokens. A unigram language model is simply defined by a list of types (words) and their individual probabilities. Although this is a weak model, it can be effectively trained from less labeled data than more complex models, such as bigram or trigram models. Additionally, higher order n-gram models might capture grammatical as well as lexical differences. The relative contributions of grammatical and lexical features were thus better distinguished by using unigram language

models that more exclusively focus on lexical differences.

In this language modeling approach, a generative model is assumed for a passage $T$, in which a hypothetical author generates the tokens of $T$ by:

1. Choosing a grade language model, $G_i$, from the set $G = \{G_i\}$ of 12 unigram language models, according to a prior probability distribution $P(G_i)$.

2. Choosing a passage length $|T|$ in tokens according to a probability distribution $P(|T|)$.

3. Sampling $|T|$ tokens from $G_i$'s multinomial word distribution according to the 'naïve' assumption that each token is independent of all other tokens in the passage, given the language model $G_i$.

These assumptions lead to the following expression for the probability of $T$ being generated by language model $G_i$ according to a multinomial distribution:

$$P(T \mid G_i) = P(|T|)\,|T|!\prod_{w \in V} \frac{P(w \mid G_i)^{C(w)}}{C(w)!}$$

Next, according to Bayes' Theorem:

$$P(G_i \mid T) = \frac{P(G_i)P(T \mid G_i)}{P(T)}.$$

Substituting (1) into (2), taking logarithms, and simplifying produces:

$$\log P(G_i \mid T) = \sum_{w \in V} C(w) \log P(w \mid G_i) - \sum_{w \in V} \log C(w)! + \log R + \log S \quad ,$$

where $V$ is the list of all types in the passage $T$, $w$ is a type in $V$, and $C(w)$ is the number of tokens with type $w$ in $T$. For simplicity, the factor $R$ represents the contribution of the prior $P(G_i)$, and $S$ represents the contribution of the passage length $|T|$, given the grade level.

Two further assumptions are made to simplify the illustration:

1. That all grades are equally likely *a priori*. That is, $P(G_i) = \frac{1}{N_G}$ where $N_G$ is the number of grade levels. For example, if there are 12 grade levels, then $N_G = 12$. This allows $\log R$ to be ignored.

2. That all passage lengths (up to a maximum length $M$) are equally likely. This allows $\log S$ to be ignored.

These may be poor assumptions in a real application, but they can be easily included or excluded in the model as desired. The $\log C(w)!$ term can also be ignored because it is constant across levels. Under these conditions, an extremely simple form for the grade likelihood remains. In order to find which model $G^i$ maximizes Equation (3), the model which $G^i$ that maximizes the following equation must be found:

$$L(T \mid G_i) = \sum_{w \in V} C(w) \log P(w \mid G_i)$$

This is straightforward to compute: for each token in the passage $T$, the log probability of the token according to the language model of $G^i$ is calculated. Summing the log probabilities of all tokens produces the overall likelihood of the passage, given the grade. The grade level with the maximum likelihood is then chosen as the final readability level prediction.

This study employs a slightly more sophisticated extension of this model, in which a sliding window is moved across the text, with a grade prediction being made for each window. This results in a distribution of grade predictions. The grade level corresponding to a given percentile of this distribution is chosen as the prediction for the entire document. The values used in these experiments for the percentile thresholds for L1 and L2 were chosen by accuracy on held-out data.

## 3 Grammatical Construction Readability Prediction for Second Language Texts

The following sections describe the approach to predicting readability based on grammatical features. As with any classifier, two components are required to classify texts by their reading level: first, a definition for and method of identifying features; second, an algorithm for using these features to classify a given text. A third component, training data, is also necessary in this classification

task. The corpus of materials used for training and testing is discussed in a subsequent section.

### 3.1 Features for Grammar-based Prediction

L2 learners usually learn grammatical patterns explicitly from grammar explanations in L2 textbooks, unlike their L1 counterparts who learn them implicitly through natural interactions. Grammatical features would therefore seem to be an essential component of an automatic readability measure for L2 learners, who must actively acquire both the lexicon and grammar of their target language.

The grammar-based readability measure relies on being able to automatically identify grammatical constructions in text. Doing so is a multi-step process that begins by syntactically parsing the document. The Stanford Parser (Klein and Manning, 2002) was used to produce constituent structure trees. The choice of parser is not essential to the approach, although the accuracy of parsing does play a role in successful identification of certain grammatical patterns. PCFG scores from the parser were also used to filter out some of the ill-formed text present in the test corpora. The default training set of Penn Treebank (Marcus et al. 1993) was used for the parser because the domain and style of those texts actually matches fairly well with the domain and style of the texts on which a reading level predictor for second language learners might be used.

Once a document is parsed, the predictor uses Tgrep2 (Rohde, 2005), a tree structure searching tool, to identify instances of the target patterns. A Tgrep2 pattern defines dominance, sisterhood, precedence, and other relationships between nodes in the parse tree for a sentence. A pattern can also place constraints on the terminal symbols (e.g., words and punctuation), such that a pattern might require a form of the copula "be" to exist in a certain position in the construction. An example of a TGrep2 search pattern for the progressive verb tense is the following:

"VP < /^VB/ < (VP < VBG)"

Searching for this pattern returns sentences in which a verb phrase (VP) dominates an auxiliary verb (whose symbol begins with VB) as well as another verb phrase, which in turn dominates a verb in gerund form (VBG). An example of a matching sentence is, "The student was reading a book," shown in Figure 2.

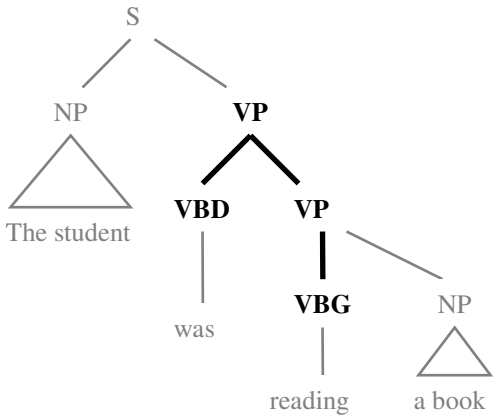


Figure 2: The parse tree for an example sentence that matches a pattern for progressive verb tense.

A set of 22 relevant grammatical constructions were identified from grammar textbooks for three different ESL levels (Fuchs et al., 2005). These grammar textbooks had different authors and publishers than the ones used in the evaluation corpora in order to minimize the chance of experimental results not generalizing beyond the specific materials employed in this study. The ESL levels correspond to the low-intermediate (hereafter, level 3), high-intermediate (level 4), and advanced (level 5) courses at the University of Pittsburgh's English Language Institute. The constructions identified in these grammar textbooks were then implemented in the form of Tgrep2 patterns.

| **Feature** | Lowest Level | Highest Level |
|---|---|---|
| Passive Voice | 0.11 | 0.71 |
| Past Participle | 0.28 | 1.63 |
| Perfect Tense | 0.01 | 0.33 |
| Relative Clause | 0.54 | 0.60 |
| Continuous Tense | 0.19 | 0.27 |
| Modal | 0.80 | 1.44 |

Table 1: The rates of occurrence per 100 words of a few of the features used by the grammar-based predictor. Rates are shown for the lowest (2) and highest (5) levels in the L2 corpus.

The rate of occurrence of constructions was calculated on a per word basis. A per-word rather

than a per-sentence measure was chosen because a per-sentence measure would depend too greatly on sentence length, which also varies by level. It was also desirable to avoid having sentence length confounded with other features. Table 1 shows that the rates of occurrence of certain constructions become more frequent as level increases. This systematic variation across levels is the basis for the grammar-based readability predictions.

A second feature set was defined that consisted of 12 grammatical features that could easily be identified without computationally intensive syntactic parsing. These features included sentence length, the various verb forms in English, including the present, progressive, past, perfect, continuous tenses, as well as part of speech labels for words. The goal of using a second feature set was to examine how dependent prediction quality was on a specific set of features, as well as to test the extent to which the output of syntactic parsing might improve prediction accuracy.

### 3.2 Algorithm for Grammatical Feature-based Classification

A k-Nearest Neighbor (kNN) algorithm is used for classification based on the grammatical features described above. The kNN algorithm is an instance-based learning technique originally developed by Cover and Hart (1967) by which a test instance is classified according to the classifications of a given number (k) of training instances closest to it. Distance is defined in this work as the Euclidean distance of feature vectors. Mitchell (1997) provides more details on the kNN algorithm. This algorithm was chosen because it has been shown to be effective in text classification tasks when compared to other popular methods (Yang 1999). A k value of 12 was chosen because it provided the best performance on held-out data.

Additionally, it is straightforward to calculate a confidence measure with which kNN predictions can be combined with predictions from other classifiers—in this case with predictions from the unigram language modeling-based approach described above. A confidence measure was important in this task because it provided a means with which to combine the grammar-based predictions with the predictions from the language modeling-based predictor while maintaining separate models for each type of feature. These separate models were maintained to better determine the relative contributions of grammatical and lexical features.

A static linear interpolation of predictions using the two approaches led to only minimal reductions of prediction error, likely because predictions from the poorer performing grammar-based classifier were always given the same weight. However, with the confidence measures, predictions from the grammar-based classifier could be given more weight when the confidence measure was high, and less weight when the measure was low and the predictions were likely to be inaccurate. The case-dependent interpolation of prediction values allowed for the effective combination of language modeling- and grammar-based predictions.

The confidence measure employed is the proportion of the $k$ most similar training examples, or nearest neighbors, that agree with the final label chosen for a given test document. For example, if seven of ten neighbors have the same label, then the confidence score will be 0.6. The interpolated readability prediction value is calculated as follows:

$$L_I = L_{LM} + C_{kNN} * L_{GR},$$

where $L_{LM}$ is the language model-based prediction, $L_{GR}$ is the grammar-based prediction from the kNN algorithm, and $C_{kNN}$ is the confidence value for the kNN prediction. The language modeling approach is treated as a black box, but it would likely be beneficial to have confidence measures for it as well.

## 4 Descriptions of Experiments

This section describes the experiments used to test the hypothesis that grammar-based features can improve readability measures for English, especially for second language texts. The measures and cross-validation setup are described. A description of the evaluation corpora of labeled first and second language texts follows.

### 4.1 Experimental Setup

Two measurements were used in evaluating the effectiveness of the reading level predictions. First, the correlation coefficient evaluated whether the trends of prediction values matched the trends for human-labeled texts. Second, the mean squared error of prediction values provided a

measure of how correct each of the predictors was on average, penalizing more severe errors more heavily. Mean square error was used rather than simple accuracy (i.e., number correct divided by sample size) because the task of readability prediction is more akin to regression than classification. Evaluation measures such as accuracy, precision, and recall are thus less meaningful for readability prediction tasks because they do not capture the fact that an error of 4 levels is more costly than an error of a single level.

A nine-fold cross-validation was employed. The data was first split into ten sets. One set was used as held-out data for selecting the parameter *k* for the kNN algorithm and the percentile value for the language modeling predictor, and then the remaining nine were used to evaluate the quality of predictions. Each of these nine was in turn selected as the test set, and the other eight were used as training data.

### 4.2 Corpora of Labeled Texts

Two corpora of labeled texts were used in the evaluation. The first corpus was from a set of texts gathered from the Web for a prior evaluation of the language modeling approach. The 362 texts had been assigned L1 levels (1-12) by grade school teachers, and consisted of approximately 250,000 words. For more details on the L1 corpus, see (Collins-Thompson and Callan, 2005).

The second corpora consisted of textbook materials (Adelson-Goldstein and Howard, 2004, for level 2; Ediger and Pavlik, 2000, for levels 3 and 4; Silberstein, 2002, for level 5) from a series of English as a Second Language reading courses at the English Language Institute at the University of Pittsburgh. The four reading practice textbooks that constitute this corpus were from separate authors and publishers than the grammar textbooks used to select and define grammatical features. The reading textbooks in the corpus are used in courses intended for beginning (level 2) through advanced (level 5) students. The textbooks were scanned into electronic format, and divided into fifty roughly equally sized files. This second language corpus consisted of approximately 200,000 words.

Although the sources and formats of the two corpora were different, they share a number of characteristics. Their size was roughly equal. The documents in both were also fairly but not perfectly evenly distributed across the levels. Both corpora also contained a significant amount of noise which made accurate prediction of reading level more challenging. The L1 corpus was from the Web, and therefore contained navigation menus, links, and the like. The texts in the L2 corpus also contained significant levels of noise due to the inclusion of directions preceding readings, exercises and questions following readings, as well as labels on figures and charts. The scanned files were not hand-corrected in this study, in part to test that the measures are robust to noise, which is present in the Web documents for which the readability measures are employed in the REAP tutoring system.

The grammar-based prediction seems to be more significantly negatively affected by the noise in the two corpora because the features rely more on dependencies between different words in the text. For example, if a word happened to be part of an image caption rather than a well-formed sentence, the unigram language modeling approach would only be affected for that word, but the grammar-based approach might be affected for features spanning an entire clause or sentence.

## 5 Results of Experiments

The results show that for both the first and second language corpora, the language modeling (LM) approach alone produced more accurate predictions than the grammar-based approach alone. The mean squared error values (Table 2) were lower, and the correlation coefficients (Table 3) were higher for the LM predictor than the grammar-based predictor.

The results also indicate that while grammar-based predictions are not as accurate as the vocabulary-based scores, they can be combined with vocabulary-based scores to produce more accurate interpolated scores. The interpolated predictions combined by using the kNN confidence measure were slightly and in most tests significantly more accurate in terms of mean squared error than the predictions from either single measure. Interpolation using the first set of grammatical features led to 7% and 22% reductions in mean squared error on the L1 and L2 corpora, respectively. These results were verified using a one-tailed paired t-test

of the squared error values of the predictions, and significance levels are indicated in Table 2.

| Mean Squared Error Values | | |
|---|---|---|
| Test Set (Num. Levels) | **L1(12)** | **L2(4)** |
| Language Modeling | 5.02 | 0.51 |
| Grammar | 10.27 | 1.08 |
| Interpolation | 4.65* | 0.40** |
| Grammar2 (feature set #2) | 12.77 | 1.26 |
| Interp2. (feature set #2) | 4.73 | 0.43* |

Table 2. Comparison of Mean Squared Error of predictions compared to human labels for different methods. Interpolated values are significantly better compared to language modeling predictions where indicated (* = $p<0.05$, ** = $p<0.01$).

| Correlation Coefficients | | |
|---|---|---|
| Test Set (Num. Levels) | **L1(12)** | **L2(4)** |
| Language Modeling | 0.71 | 0.80 |
| Grammar | 0.46 | 0.55 |
| Interpolation | 0.72 | 0.83 |
| Grammar2 (feature set #2) | 0.34 | 0.48 |
| Interp2. (feature set #2) | 0.72 | 0.81 |

Table 3. Comparison of Correlation Coefficients of prediction values to human labels for different prediction methods.

The trends were similar for both sets of grammatical features. However, the first set of features that included complex syntactic constructs led to better performance than the second set, which included only verb tenses, part of speech labels, and sentence length. Therefore, when syntactic parsing is not feasible because of corpora size, it seems that grammatical features requiring only part-of-speech tagging and word counts may still improve readability predictions. This is practically important because parsing can be too computationally intensive for large corpora.

All prediction methods performed better, in terms of correlations, on the L2 corpus than on the L1 corpus. The L2 corpus is somewhat smaller in size and should, if only on the basis of training material available to the prediction algorithms, actually be more difficult to predict than the L1 corpus. To ensure that the range of levels was not causing the four-level L2 corpus to have higher predictions than the twelve-level L1 corpus, the L1 corpus was also divided into four bins (grades 1-3, 4-6, 7-9, 10-12). The accuracy of predictions for the binned version of the L1 corpus was not substantially different than for the 12-level version.

## 6 Discussion

In the experimental tests, the LM approach was more effective for measuring both L1 and L2 readability. There are several potential causes of this effect. First, the language modeling approach can utilize all the words as they appear in the text as features, while the grammatical features were chosen and defined manually. As a result, the LM approach can make measurements on a text for as many features as there are words in its lexicon. Additionally, the noise present in the corpora likely affected the grammar-based approach disproportionately more because that method relies on accurate parsing of relationships between words.

Additionally, English is a morphologically impoverished language compared to most languages. Text classification, information retrieval, and many other human language technology tasks can be accomplished for English without accounting for grammatical features such as morphological inflections. For example, an information retrieval system can perform reasonably well in English without performing stemming, which does not greatly increase performance except when queries and documents are short (Krovetz, 1993).

However, most languages have a rich morphology by which a single root form may have thousands or perhaps millions of inflected or derived forms. Language technologies must account for morphological features in such languages or the vocabulary grows so large that it becomes unmanageable. Lee (2004), for example, showed that morphological analysis can improve the quality of statistical machine translation for Arabic. Thus it seems that grammatical features could contribute even more to measures of readability for texts in other languages.

That said, the use of grammatical features appears to play a more important role in readability measures for L2 than for L1. When interpolated with grammar-based scores, the reduction of mean squared error over the language modeling approach for L1 was only 7%, while for L2 the reduction or squared error was 22%. An evaluation on corpora with less noise would likely bring out these differ-

ences further and show grammar to be an even more important factor in second language readability. This result is consistent with the fact that second language learners are still in the process of acquiring the basic grammatical constructs of their target language.

## 7 Conclusion

The results of this work suggest that grammatical features can play a role in predicting reading difficulty levels for both first and second language texts in English. Although a vocabulary-based language modeling approach outperformed the grammar-based predictor, an interpolated measure using confidence scores for the grammar-based predictions showed improvement over both individual measures. Also, grammar appears to play a more important role in second language readability than in first language readability. Ongoing work aims to improve grammar-based readability by reducing noise in training data, automatically creating larger grammar feature sets, and applying more sophisticated modeling techniques.

## 8 Acknowledgements

We would like to acknowledge Lori Levin for useful advice regarding grammatical constructions, as well as the anonymous reviewers for their suggestions.

This material is based on work supported by NSF grant IIS-0096139 and Dept. of Education grant R305G03123. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsors.

## References

J. Adelson-Goldstein and L. Howard. 2004. *Read and Reflect 1*. Oxford University Press, USA.

E. Bates. 2003. On the nature and nurture of language. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco, F. Jacob, E. Bizzi, P. Calissano, & V. Volterra (Eds.), *Frontiers of biology: The brain of Homo sapiens* (pp. 241–265). Rome: Istituto della Enciclopedia Italiana fondata da Giovanni Trecanni.

M. Fuchs, M. Bonner, M. Westheimer. 2005. *Focus on Grammar,* 3rd Edition. Pearson ESL.

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. *Proceedings of the HLT/NAACL Annual Conference*.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21-27.

A. Ediger and C. Pavlik. 2000. *Reading Connections Intermediate*. Oxford University Press, USA.

A. Ediger and C. Pavlik. 2000. *Reading Connections High Intermediate*. Oxford University Press, USA.

M. Heilman, K. Collins-Thompson, J. Callan & M. Eskenazi. 2006. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. *Proceedings of the Ninth International Conference on Spoken Language Processing.*

D. Klein and C. D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems* 15 (NIPS 2002), December 2002.

R. Krovetz. 1993. Viewing morphology as an inference process. *SIGIR-93*, 191–202.

Y. Lee. 2004. Morphological Analysis for Statistical Machine Translation. *Proceedings of the HLT/NAACL Annual Conference*.

M. Marcus, B. Santorini and M. Marcinkiewicz. 1993. "Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics*, 19(2).

T. Mitchell. 1997. *Machine Learning*. The McGraw-Hill Companies, Inc. pp. 231-236.

D. Rohde. 2005. *Tgrep2 User Manual*. http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf.

S. Schwarm, and M. Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

S. Silberstein, B. K. Dobson, and M. A. Clarke. 2002. *Reader's Choice,* 4th edition. University of Michigan Press/ESL.

Y. Yang. 1999. A re-examination of text categorization methods. *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'99, pp 42--49).