An Unsupervised Method for Detecting Grammatical Errors

Martin Chodorow
Hunter College of CUNY
695 Park Avenue
New York, NY
martin.chodorow @.hunter.cuny.edu

Claudia Leacock
Educational Testing Service
Rosedale Road
Princeton, NJ
cleacock@ets.org

Abstract

We present an unsupervised method for detecting grammatical errors by inferring negative evidence from edited textual corpora. The system was developed and tested using essay-length responses to prompts on the Test of English as a Foreign Language (TOEFL). The error-recognition system, ALEK, performs with about 80% precision and 20% recall.

Introduction

A good indicator of whether a person knows the meaning of a word is the ability to use it appropriately in a sentence (Miller and Gildea, 1987). Much information about usage can be obtained from quite a limited context: Choueka and Lusignan (1985) found that people can typically recognize the intended sense of a polysemous word by looking at a narrow window of one or two words around it. Statistically-based computer programs have been able to do the same with a high level of accuracy (Kilgarriff and Palmer, 2000). The goal of our work is to automatically identify *inappropriate usage* of specific vocabulary words in essays by looking at the local contextual cues around a target word. We have developed a statistical system, ALEK (Assessing Lexical Knowledge), that uses statistical analysis for this purpose.

A major objective of this research is to avoid the laborious and costly process of collecting errors (or negative evidence) for each word that we wish to evaluate. Instead, we train ALEK on a general corpus of English and on edited text containing example uses of the target word. The system identifies inappropriate usage based on differences between the word's local context cues in an essay and the models of context it has derived from the corpora of well-formed sentences.

A requirement for ALEK has been that all steps in the process be automated, beyond choosing the words to be tested and assessing the results. Once a target word is chosen, preprocessing, building a model of the word's appropriate usage, and identifying usage errors in essays is performed without manual intervention.

ALEK has been developed using the Test of English as a Foreign Language (TOEFL) administered by the Educational Testing Service. TOEFL is taken by foreign students who are applying to US undergraduate and graduate-level programs.

1 Background

Approaches to detecting errors by non-native writers typically produce grammars that look for specific expected error types (Schneider and McCoy, 1998; Park, Palmer and Washburn, 1997). Under this approach, essays written by ESL students are collected and examined for errors. Parsers are then adapted to identify those error types that were found in the essay collection.

We take a different approach, initially viewing error detection as an extension of the word sense disambiguation (WSD) problem. Corpus-based WSD systems identify the intended sense of a polysemous word by (1) collecting a set of example sentences for each of its various senses and (2) extracting salient contextual cues from these sets to (3) build a statistical model for each sense. They identify the intended sense of a word in a novel sentence by extracting its contextual cues and selecting the most similar word sense model (e.g., Leacock, Chodorow and Miller (1998), Yarowsky (1993)).

Golding (1995) showed how methods used for WSD (decision lists and Bayesian classifiers) could be adapted to detect errors resulting from common spelling confusions among sets such as *there*, *their*, and *they're*. He extracted contexts from correct usage of each confusable word in a training corpus and then identified a new occurrence as an error when it matched the wrong context.

However, most grammatical errors are not the result of simple word confusions. This complicates the task of building a model of incorrect usage. One approach we considered was to proceed without such a model: represent appropriate word usage (across senses) in a single model and compare a novel example to that model. The most appealing part of this formulation was that we could bypass the knowledge acquisition bottleneck. All occurrences of the word in a collection of edited text could be automatically assigned to a single training set representing appropriate usage. Inappropriate usage would be signaled by contextual cues that do not occur in training.

Unfortunately, this approach was not effective for error detection. An example of a word usage error is often very similar to the model of appropriate usage. An incorrect usage can contain two or three salient contextual elements as well as a single anomalous element. The problem of error detection does not entail finding similarities to appropriate usage, rather it requires identifying one element among the contextual cues that simply does not fit.

2 ALEK Architecture

What kinds of anomalous elements does ALEK identify? Writers sometimes produce errors that violate basic principles of English syntax (e.g., *a desks*), while other mistakes show a lack of information about a specific vocabulary item (e.g., *a knowledge*). In order to detect these two types of problems, ALEK uses a 30-million word general corpus of English from the San Jose Mercury News (hereafter referred to as the *general corpus*) and, for each target word, a set of 10,000 example sentences from North American newspaper text[1] (hereafter referred to as the *word-specific corpus*).

ALEK infers negative evidence from the contextual cues that *do not* co-occur with the target word – either in the word specific corpus or in the general English one. It uses two kinds of contextual cues in a ±2 word window around the target word: function words (closed-class items) and part-of-speech tags (Brill, 1994). The Brill tagger output is post-processed to "enrich" some closed class categories of its tag set, such as subject versus object pronoun and definite versus indefinite determiner. The enriched tags were adapted from Francis and Kučera (1982).

After the sentences have been preprocessed, ALEK counts sequences of adjacent part-of-speech tags and function words (such as determiners, prepositions, and conjunctions). For example, the sequence *a/AT full-time/JJ job/NN* contributes one occurrence each to the bigrams AT+JJ, JJ+NN, a+JJ, and to the part-of-speech tag trigram AT+JJ+NN. Each individual tag and function word also contributes to its own unigram count. These frequencies form the basis for the error detection measures.

From the general corpus, ALEK computes a mutual information measure to determine which sequences of part-of-speech tags and function words are unusually rare and are, therefore, likely to be ungrammatical in English (e.g., singular determiner preceding plural noun, as in **a desks*). Mutual information has often been used to detect combinations of words that occur more frequently than we would expect based on the assumption that the words are independent. Here we use this measure for the opposite purpose – to find combinations that occur less often than expected. ALEK also looks for sequences that are common in general but unusual in the word specific corpus (e.g., the singular determiner *a* preceding a singular noun is common in English but rare when the noun is

[1] The corpora are extracted from the ACL-DCI corpora. In selecting the sentences for the word specific corpora, we tried to minimize the mismatch between the domains of newspapers and TOEFL essays. For example, in the newspaper domain, *concentrate* is usually used as a noun, as in *orange juice concentrate* but in TOEFL essays it is a verb 91% of the time. Sentence selection for the word specific corpora was constrained to reflect the distribution of part-of-speech tags for the target word in a random sample of TOEFL essays.

knowledge). These divergences between the two corpora reflect syntactic properties that are peculiar to the target word.

2.1 Measures based on the general corpus:

The system computes mutual information comparing the proportion of observed occurrences of bigrams in the general corpus to the proportion expected based on the assumption of independence, as shown below:

$$MI = \log_2\left(\frac{P(AB)}{P(A) \times P(B)}\right)$$

Here, P(AB) is the probability of the occurrence of the AB bigram, estimated from its frequency in the general corpus, and P(A) and P(B) are the probabilities of the first and second elements of the bigram, also estimated from the general corpus. Ungrammatical sequences should produce bigram probabilities that are much smaller than the product of the unigram probabilities (the value of MI will be negative). Trigram sequences are also used, but in this case the mutual information computation compares the co-occurrence of ABC to a model in which A and C are assumed to be conditionally independent given B (see Lin, 1998).

$$MI = \log_2\left(\frac{P(ABC)}{P(B) \times P(A \mid B) \times P(C \mid B)}\right)$$

Once again, a negative value is often indicative of a sequence that violates a rule of English.

2.2 Comparing the word-specific corpus to the general corpus:

ALEK also uses mutual information to compare the distributions of tags and function words in the word-specific corpus to the distributions that are expected based on the general corpus. The measures for bigrams and trigrams are similar to those given above except that the probability in the numerator is estimated from the word-specific corpus and the probabilities in the denominator come from the general corpus. To return to a previous example, the phrase *a knowledge* contains the tag bigram for singular determiner followed by singular noun (AT NN). This sequence is much less common in the word-specific corpus for *knowledge* than would be expected from the general corpus unigram probabilities of AT and NN.

In addition to bigram and trigram measures, ALEK compares the target word's part-of-speech tag in the word-specific corpus and in the general corpus. Specifically, it looks at the conditional probability of the part-of-speech tag given the major syntactic category (e.g., plural noun given noun) in both distributions, by computing the following value.

$$\log_2\left(\frac{P_{specific\_corpus}(tag \mid category)}{P_{general\_corpus}(tag \mid category)}\right)$$

For example, in the general corpus, about half of all noun tokens are plural, but in the training set for the noun *knowledge*, the plural *knowledges* occurs rarely, if at all.

The mutual information measures provide candidate errors, but this approach overgenerates – it finds rare, but still quite grammatical, sequences. To reduce the number of false positives, no candidate found by the MI measures is considered an error if it appears in the word-specific corpus at least two times. This increases ALEK's precision at the price of reduced recall. For example, *a knowledge* will not be treated as an error because it appears in the training corpus as part of the longer *a knowledge of* sequence (as in *a knowledge of mathematics*).

ALEK also uses another statistical technique for finding rare and possibly ungrammatical tag and function word bigrams by computing the χ^2 (*chi* square) statistic for the difference between the bigram proportions found in the word-specific and in the general corpus:

$$\chi^2 = \left(\frac{(P_{word\_specific} - P_{general\_corpus})^2}{P_{general\_corpus}(1 - P_{gerneral\_corpus}) / N_{word\_specific}}\right)$$

The χ^2 measure faces the same problem of overgenerating errors. Due to the large sample sizes, extreme values can be obtained even though effect size may be minuscule. To reduce false positives, ALEK requires that effect sizes be at least in the moderate-to-small range (Cohen and Cohen, 1983).

Direct evidence from the word specific corpus can also be used to control the overgeneration of errors. For each candidate error, ALEK compares the larger context in which the bigram appears to the contexts that have been analyzed in the word-specific corpus. From the word-specific corpus, ALEK forms *templates*, sequences of words and tags that represent the local context of the target. If a test sentence contains a low probability bigram (as measured by the χ^2 test), the local context of the target is compared to all the templates of which it is a part. Exceptions to the error, that is longer grammatical sequences that contain rare sub-sequences, are found by examining conditional probabilities. To illustrate this, consider the example of *a knowledge* and *a knowledge of.* The conditional probability of *of* given *a knowledge* is high, as it accounts for almost all of the occurrences of *a knowledge* in the word-specific corpus. Based on this high conditional probability, the system will use the template for *a knowledge of* to keep it from being marked as an error. Other function words and tags in the +1 position have much lower conditional probability, so for example, *a knowledge is* will not be treated as an exception to the error.

2.3 Validity of the n-gram measures

TOEFL essays are graded on a 6 point scale, where 6 demonstrates "clear competence" in writing on rhetorical and syntactic levels and 1 demonstrates "incompetence in writing". If low probability n-grams signal grammatical errors, then we would expect TOEFL essays that received lower scores to have more of these n-grams. To test this prediction, we randomly selected from the TOEFL pool 50 essays for each of the 6 score values from 1.0 to 6.0. For each score value, all 50 essays were concatenated to form a super-essay. In every super-essay, for each adjacent pair and triple of tags containing a noun, verb, or adjective, the bigram and trigram mutual information values were computed based on the general corpus.

Score	% of bigrams	% of trigrams
1.0	3.6	1.4
2.0	3.4	0.8
3.0	2.6	0.6
4.0	1.9	0.3
5.0	1.3	0.4
6.0	1.5	0.3

Table 1: Percent of n-grams with mutual information < -3.60, by score point

Table 1 shows the proportions of bigrams and trigrams with mutual information less than –3.60. As predicted, there is a significant negative correlation between the score and the proportion of low probability bigrams (r_s= -.94, n=6, p<.01, two-tailed) and trigrams (r_s= -.84, n=6, p<.05, two-tailed).

2.4 System development

ALEK was developed using three target words that were extracted from TOEFL essays: *concentrate*, *interest*, and *knowledge*. These words were chosen because they represent different parts of speech and varying degrees of polysemy. Each also occurred in at least 150 sentences in what was then a small pool of TOEFL essays. Before development began, each occurrence of these words was manually labeled as an appropriate or inappropriate usage – without taking into account grammatical errors that might have been present elsewhere in the sentence but which were not within the target word's scope.

Critical values for the statistical measures were set during this development phase. The settings were based empirically on ALEK's performance so as to optimize precision and recall on the three development words. Candidate errors were those local context sequences that produced a mutual information value of less than –3.60 based on the general corpus; mutual information of less than –5.00 for the specific/general comparisons; or a χ^2 value greater than 12.82 with an effect size greater than 0.30. Precision and recall for the three words are shown below.

Target word	n	Precision	Recall
Concentrate	169	.875	.280
Interest	416	.840	.330
Knowledge	761	.918	.570

Table 2: Development Words

Test Word	Precision	Recall	Total Recall (estimated)	Test Word	Precision	Recall	Total Recall (estimated)
Affect	.848	.762	.343	Energy	.768	.666	.104
Area	.752	.846	.205	Function	.800	.714	.168
Aspect	.792	.717	.217	Individual	.576	.742	.302
Benefit	.744	.709	.276	Job	.728	.679	.103
Career	.736	.671	.110	Period	.832	.670	.102
Communicate	.784	.867	.274	Pollution	.912	.780	.310
Concentrate	.848	.791	.415	Positive	.784	.700	.091
Conclusion	.944	.756	.119	Role	.728	.674	.098
Culture	.704	.656	.083	Stress	.768	.578	.162
Economy	.816	.666	.235	Technology	.728	.674	.093
				Mean	**.779**	**.716**	**.190**

Table 3: Precision and recall for 20 test words

3 Experimental Design and Results

ALEK was tested on 20 words. These words were randomly selected from those which met two criteria: (1) They appear in a university word list (Nation, 1990) as words that a student in a US university will be expected to encounter and (2) there were at least 1,000 sentences containing the word in the TOEFL essay pool.

To build the usage model for each target word, 10,000 sentences containing it were extracted from the North American News Corpus. Preprocessing included detecting sentence boundaries and part-of-speech tagging. As in the development system, the model of general English was based on bigram and trigram frequencies of function words and part-of-speech tags from 30-million words of the San Jose Mercury News.

For each test word, all of the test sentences were marked by ALEK as either containing an error or not containing an error. The size of the test set for each word ranged from 1,400 to 20,000 with a mean of 8,000 sentences.

3.1 Results

To evaluate the system, for each test word we randomly extracted 125 sentences that ALEK classified as containing no error (C-set) and 125 sentences which it labeled as containing an error (E-set). These 250 sentences were presented to a linguist in a random order for blind evaluation. The linguist, who had no part in ALEK's development, marked each usage of the target word as incorrect or correct and in the case of incorrect usage indicated how far from the target one would have to look in order to recognise that there was an error. For example, in the case of "an period" the error occurs at a distance of one word from *period.* When the error is an omission, as in "lived in Victorian period", the distance is where the missing word should have appeared. In this case, the missing determiner is 2 positions away from the target. When more than one error occurred, the distance of the one closest to the target was marked.

Table 3 lists the precision and recall for the 20 test words. The column labelled "Recall" is the proportion of human-judged errors in the 250-sentence sample that were detected by ALEK. "Total Recall" is an estimate that extrapolates from the human judgements of the sample to the entire test set. We illustrate this with the results for *pollution.* The human judge marked as incorrect usage 91.2% of the sample from ALEK's E-set and 18.4% of the sample from its C-set. To estimate overall incorrect usage, we computed a weighted mean of these two rates, where the weights reflected the proportion of sentences that were in the E-set and C-set. The E-set contained 8.3% of the *pollution* sentences and the C-set had the remaining 91.7%. With the human judgements as the gold standard, the estimated overall rate of incorrect usage is (.083 × .912 + .917 × .184) = .245. ALEK's estimated recall is the proportion of sentences in the E-set times its precision, divided by the overall estimated error rate (.083 × .912) / .245 = .310.

The precision results vary from word to word. *Conclusion* and *pollution* have precision in the low to middle 90's while *individual*'s precision is 57%. Overall, ALEK's predictions are about 78% accurate. The recall is limited in part by the fact that the system only looks at syntactic information, while many of the errors are semantic.

3.2 Analysis of Hits and Misses

Nicholls (1999) identifies four error types: an unnecessary word (*affect *to* their emotions), a missing word (*opportunity of job.), a word or phrase that needs replacing (**every jobs*), a word used in the wrong form (**pollutions*). ALEK recognizes all of these types of errors. For closed class words, ALEK identified whether a word was *missing*, the wrong word was used (*choice*), and when an *extra* word was used. Open class words have a fourth error category, *form*, including inappropriate compounding and verb agreement. During the development stage, we found it useful to add additional error categories. Since TEOFL graders are not supposed to take punctuation into account, *punctuation* errors were only marked when they caused the judge to "garden path" or initially misinterpret the sentence. *Spelling* was marked either when a function word was misspelled, causing part-of-speech tagging errors, or when the writer's intent was unclear.

The distributions of categories for hits and misses, shown in Table 4, are not strikingly different. However, the hits are primarily syntactic in nature while the misses are both semantic (as in open-class:choice) and syntactic (as in closed-class:missing).

ALEK is sensitive to open-class word confusions (*affect* vs *effect*) where the part of speech differs or where the target word is confused with another word (**In this aspect,...* instead of *In this respect,...*). In both cases, the system recognizes that the target is in the wrong syntactic environment. Misses can also be syntactic – when the target word is confused with another word but the syntactic environment fails to trigger an error. In addition, ALEK does not recognize semantic errors when the error involves the misuse of an open-class word in combination with the target (for example, *make* in "*they *make* benefits").

Category	% Hits	% Misses
Closed-class – choice	22.5	15.5
–extra	15.5	13.0
–missing	8.0	8.5
Open-class – choice	12.0	19.0
– extra	.5	1.0
– missing	.5	1.5
– form	28.0	28.5
Punctuation	5.5	1.5
Sentence fragment	1.5	2.0
Spelling/typing error	5.5	8.5
Word order	.5	1.0

Table 4: Hits and misses based on a random sample of 200 hits and 200 misses

Closed class words typically are either selected by or agree with a head word. So why are there so many misses, especially with prepositions? The problem is caused in part by polysemy – when one sense of the word selects a preposition that another sense does not. When *concentrate* is used spatially, it selects the preposition *in*, as "the stores were concentrated *in* the downtown area". When it denotes mental activity, it selects the preposition *on*, as in "Susan concentrated *on* her studies". Since ALEK trains on all senses of *concentrate*, it does not detect the error in "*Susan concentrated *in* her studies". Another cause is that adjuncts, especially temporal and locative adverbials, distribute freely in the word-specific corpora, as in "Susan concentrated *in* her room." This second problem is more tractable than the polysemy problem – and would involve training the system to recognize certain types of adjuncts.

3.3 Analysis of False Positives

False positives, when ALEK "identifies" an error where none exists, fall into six major categories. The percentage of each false positive type in a random sample of 200 false positives is shown in Table 5.

Domain mismatch: Mismatch of the newspaper-domain word-specific corpora and essay-domain test corpus. One notable difference is that some TOEFL essay prompts call for the writer's opinion. Consequently,

Error Type	% Occurrence
Domain mismatch	12.5
Tagger	17.0
Syntactic	14.5
Free distribution	16.5
Punctuation	12.0
Infrequent tags	9.0
Other	18.5

Table 5. Distribution of false positive types

TOEFL essays often contain first person references, whereas newspaper articles are written in the third person. We need to supplement the word-specific corpora with material that more closely resembles the test corpus.

Tagger: Incorrect analysis by the part-of-speech tagger. When the part-of-speech tag is wrong, ALEK often recognizes the resulting *n*-gram as anomalous. Many of these errors are caused by training on the Brown corpus instead of a corpus of essays.

Syntactic analysis: Errors resulting from using part-of-speech tags instead of supertags or a full parse, which would give syntactic relations between constituents. For example, ALEK false alarms on arguments of ditransitive verbs such as *offer* and flags as an error "*you benefits*" in "*offers you benefits*".

Free distribution: Elements that distribute freely, such as adverbs and conjunctions, as well as temporal and locative adverbial phrases, tend to be identified as errors when they occur in some positions.

Punctuation: Most notably omission of periods and commas. Since these errors are not indicative of one's ability to use the target word, they were not considered as errors unless they caused the judge to misanalyze the sentence.

Infrequent tags. An undesirable result of our "enriched" tag set is that some tags, e.g., the post-determiner *last*, occur too infrequently in the corpora to provide reliable statistics.

Solutions to some of these problems will clearly be more tractable than to others.

4 Comparison of Results

Comparison of these results to those of other systems is difficult because there is no generally accepted test set or performance baseline. Given this limitation, we compared ALEK's performance to a widely used grammar checker, the one incorporated in Microsoft's Word97. We created files of sentences used for the three development words *concentrate*, *interest*, and *knowledge*, and manually corrected any errors outside the local context around the target before checking them with Word97. The performance for *concentrate* showed overall precision of 0.89 and recall of 0.07. For *interest*, precision was 0.85 with recall of 0.11. In sentences containing *knowledge*, precision was 0.99 and recall was 0.30. Word97 correctly detected the ungrammaticality of *knowledges* as well as *a knowledge*, while it avoided flagging *a knowledge of.*

In summary, Word97's precision in error detection is impressive, but the lower recall values indicate that it is responding to fewer error types than does ALEK. In particular, Word97 is not sensitive to inappropriate selection of prepositions for these three words (e.g., **have knowledge on history*, **to concentrate at science*). Of course, Word97 detects many kinds of errors that ALEK does not.

Research has been reported on grammar checkers specifically designed for an ESL population. These have been developed by hand, based on small training and test sets. Schneider and McCoy (1998) developed a system tailored to the error productions of American Sign Language signers. This system was tested on 79 sentences containing determiner and agreement errors, and 101 grammatical sentences. We calculate that their precision was 78% with 54% recall. Park, Palmer and Washburn (1997) adapted a categorial grammar to recognize "classes of errors [that] dominate" in the nine essays they inspected. This system was tested on eight essays, but precision and recall figures are not reported.

5 Conclusion

The unsupervised techniques that we have presented for inferring negative evidence are effective in recognizing grammatical errors in written text.

Preliminary results indicate that ALEK's error detection is predictive of TOEFL scores. If ALEK accurately detects usage errors, then it should report more errors in essays with lower scores than in those with higher scores. We have already seen in Table 1 that there is a negative correlation between essay score and two of ALEK's component measures, the general corpus n-grams. However, the data in Table 1 were not based on specific vocabulary items and do not reflect overall system performance, which includes the other measures as well.

Table 6 shows the proportion of test word occurrences that were classified by ALEK as containing errors within two positions of the target at each of 6 TOEFL score points. As predicted, the correlation is negative ($r_s = -1.00$, $n = 6, p < .001$, two-tailed). These data support the validity of the system as a detector of inappropriate usage, even when only a limited number of words are targeted and only the immediate context of each target is examined.

Score	ALEK	Human
1	.091	---
2	.085	.375
3	.067	.268
4	.057	.293
5	.048	.232
6	.041	.164

Table 6: Proportion of test word occurrences, by score point, classified as containing an error by ALEK and by a human judge

For comparison, Table 6 also gives the estimated proportions of inappropriate usage by score point based on the human judge's classification. Here, too, there is a negative correlation: $r_s = -.90$, $n = 5$, p < .05, two-tailed.

Although the system recognizes a wide range of error types, as Table 6 shows, it detects only about one-fifth as many errors as a human judge does. To improve recall, research needs to focus on the areas identified in section 3.2 and, to improve precision, efforts should be directed at reducing the false positives described in 3.3.

ALEK is being developed as a diagnostic tool for students who are learning English as a foreign language. However, its techniques could be incorporated into a grammar checker for native speakers.

Acknowledgments

We thank Susanne Wolff for evaluating the test sentences, and Robert Kantor, Ken Sheppard and 3 anonymous reviewers for their helpful suggestions.

References

Brill, E. 1994. Some advances in rule-based part-of-speech tagging. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle, AAAI.

Choueka, Y. and S. Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19:147-158.

Cohen, J. and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Francis, W. and H. Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, Houghton Mifflin.

Golding, A. 1995. A Bayesian hybrid for context-sensitive spelling correction. *Proceedings of the 3rd Workshop on Very Large Corpora*. Cambridge, MA. 39—53.

Kilgarriff, A. and M. Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computers and the Humanities*, 34:1—2.

Leacock, C., M. Chodorow and G.A. Miller. 1998. 1998. Using corpus statistics and WordNet's lexical relations for sense identification. *Computational Linguistics*, 24:1.

Lin, D. 1998. Extracting collocations from text corpora. *First Workshop on Computational Terminology*. Montreal, Canada.

Miller, G.A. and P. Gildea. 1987. How children learn words. *Scientific American*, 257.

Nation, I.S.P. 1990. *Teaching and learning vocabulary*. New York: Newbury House.

Nicholls, D. 1999. The Cambridge Learner Corpus – Error coding and analysis. Summer Workshop on Learner Corpora. Tokyo

Park, J.C., M. Palmer and G. Washburn. 1997. Checking grammatical mistakes for English-as-a-second-language (ESL) students. *Proceedings of KSEA-NERC*. New Brunswick, NJ.

Schneider, D.A. and K.F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. *Proceedings of Coling-ACL-98*, Montréal.

Yarowsky, D. 1993. One sense per collocation. Proceedings of the ARPA Workshop on Human Language Technology. San Francisco. Morgan Kaufman.