

# Impact of Word Sense Disambiguation on Ordering Dictionary Definitions in Vocabulary Learning Tutors

Kevin Dela Rosa, Maxine Eskenazi

Language Technologies Institute

Carnegie Mellon University

Pittsburgh, PA 15213

kdelaros@cs.cmu.edu, max@cs.cmu.edu

## Abstract

Past research has shown that dictionaries and glosses can be beneficial in computer assisted language learning, particularly in vocabulary learning. We propose that L2 vocabulary learners can benefit from the use of a dictionary whose definitions are sensitive to the provided reading context, and that advances in the natural language processing task of word sense disambiguation can be used to automatically order the definitions of such a dictionary. An *in-vivo* study was conducted with ESL students to investigate the effect that the order of definitions has on vocabulary learning using REAP, a computer based vocabulary tutor. Our results showed that students benefited from having the algorithmically determined best definitions listed at the top of the definition list. Furthermore, our results suggest that word sense disambiguation may currently be good enough for use in intelligent language tutoring environments.

## Introduction

Effective use of dictionaries and glosses has been an area of interest in computer assisted language learning (CALL), especially in vocabulary learning. One issue with dictionary usage in CALL that has not been investigated thoroughly is how the order and amount of dictionary definitions provided to students affect L2 (second language) vocabulary learning. We propose that providing a dictionary whose definition ordering is sensitive to the meaning of the word in the relevant reading context can help students learn vocabulary more effectively than the usual habit of lookup in a static dictionary where definitions are ordered by frequency rather than by relevance. Furthermore, since producing manually ordered definitions for every reading in a language tutor can be costly and un-scalable for large corpora, we propose that word sense disambiguation

(WSD) techniques can be used to automatically order dictionary definitions.

In this paper we first discuss past work that investigates the role of dictionaries in vocabulary learning and the use of word sense disambiguation in CALL. Next we describe the WSD methodology employed in this study and a classroom study that evaluates the usefulness of automatically ordered dictionary definitions in L2 vocabulary learning. Finally we offer a discussion of the results and suggest future research directions.

## Background

The role and effectiveness of glosses and dictionaries in language education (both in L1 and L2 learning) is often a contentious issue among researchers and educators. Some educators argue that the use of dictionaries while reading can lead to inefficient learning, particularly in reading comprehension, claiming that students sometimes fail to find the correct dictionary entry (Bogaards 1998) or that the time taken to look up words interferes with students' short-term memory, preventing them from fully focusing on the text (Knight 1994). However, many studies have also shown that students are more likely to find the correct definition of an unknown word from a dictionary than by guessing from the context (Bogaards 1998), and with the advent of electronic & online dictionaries performing word look ups take less time, thus resulting in less distraction from the text (Koyama and Takeuchi 2004). Additionally, many studies have shown that dictionary use, particularly in L2 language learning, can improve reading comprehension and vocabulary acquisition (Prichard 2008; Knight 1994; Luppescu and Day 1993; Summers 1988).

In CALL, a variety of dictionaries and glosses have been shown to be effective in L2 learning, and the focus has shifted from determining whether glosses are effective to finding the types of glosses that are effective. For example, a study by Lomicka (1998) involving college students in a second semester French course found that students pro-

vided with a "full gloss" (L1 translation, L2 definitions, pronunciation, and pictures) in a computerized reading task lead to better text comprehension than those provided with a "traditional" gloss (L1 translation and L2 definitions) or no gloss. Additionally, a study by Yoshii (2006) involving Japanese college students learning English showed that both L1 and L2 glosses with text and pictures are effective for incidental vocabulary learning. Lastly, Laufer and Hill (2000) found L2 English learners have different lookup preferences and use different kinds of information for a word provided by an electronic dictionary, such as a digitized voice recording, English meaning, L1 meaning, root, other forms of word, phonemic transcription, and other semantic and syntactic details, reinforce each other in reading comprehension and vocabulary retention.

An issue that teachers have with the use of dictionaries, as mentioned previously, is that sometimes students select the wrong meaning for an unknown word while performing a word look up, which can lead to miscomprehensions of the word as well as of the reading material. We propose that advances in natural language processing specifically on the task of word sense disambiguation can help alleviate this problem by ordering dictionary definitions in an intelligent tutoring environment. WSD is a well studied area in natural language processing, especially in the supervised setting (Ide and Jean, 1998; Schütze 1998; Pederson and Bruce 1997). A study by Kulkarni et al. (2008), showed that WSD-ordered definitions can be helpful in vocabulary learning and we extend their methodology by making use of crowd-sourcing to build our WSD training data set, which is much less expensive than paying experts to label the data and results in similar quality, and examining whether seeing a single definition is preferable to a group of definitions.

## Word Sense Disambiguation Methodology

For this study we constructed WSD classifiers to order dictionary definitions by making use of a training data set consisting of instances of target words found in documents discovered by REAP (Heilman et al. 2006). In the following sections we discuss the methodology we used for word sense disambiguation by describing the data set that we constructed to train our WSD classifiers, the various algorithms we experimented with, and the feature set used in our classifiers.

### Overview of REAP

REAP, which stands for **RE**ADER-specific **P**ractice, is a web based language tutor developed at Carnegie Mellon University that makes use of documents harvested from the internet for L2 vocabulary learning and reading comprehension (Heilman et al. 2006). REAP's interface has a number of features that help to enhance a student's learning experience, such as the ability to provide reader-specific passages and focus word highlighting, and the

ability to generate synthesized versions of every word that appear in the passages.

The feature most relevant to this study is the dictionary word lookup system that is embedded in the interface which allows students to look up the definition of any of the words they encounter during readings.

## Word Sense Disambiguation Data Set

In order to train our WSD classifiers, we constructed a data set which made use of documents from REAP, target words from the Academic Word List (Coxhead 2000), and definitions from the Cambridge Advanced Learning Dictionary (CALD) (Walter 2005). The contents and construction of our data set are described in the following sections. In total, our WSD training data set has 18,250 labeled word sense instances for a set of 192 target words.

**Words and Definitions.** The WSD training data set contains a set of target words to be disambiguated which comes from the Academic Word List. The definitions for each target word came from the Cambridge Advanced Learning Dictionary, which were then grouped into word senses. A total of 192 target words that had more than one sense for the same part-of-speech were selected. For the classroom study, we looked at a subset of these words (19 words), to analyze how the algorithmically determined dictionary definitions affected L2 vocabulary learning.

**Crowd-sourcing and Word Sense Induction.** Amazon Mechanical Turk (AMT) was used to cluster the dictionary definitions, in a task called Word Sense Induction, since paying experts to perform this task would be too expensive and would result in similar quality. The crowd-sourcing methodology used to produce our word senses is described in (Parent and Eskenazi 2010). Parent and Eskenazi showed that AMT can reliably be used to cluster dictionary definitions, and yields inter-annotator agreement with experts that corresponds to agreement between experts.

**Documents and Word Instances.** The WSD training data set contains a set of documents taken from REAP, each of which contain one or more instances of a target word whose correct sense was labeled during a crowd-sourcing task. In total there were 14,613 total documents in the WSD training data set, and 18,250 target word instances.

## Features

We used two types of features in our WSD classifiers; unigrams (UNI) and part-of-speech (POS) of surrounding words within some window of the word being categorized, which have been shown to be effective on the task of supervised word sense disambiguation (Mohammad and Pederson 2004). For unigrams we experimented with word windows ( $-w$ ,  $+w$ ) of size 5 to 100 in steps of 5, and for the part-of-speech features we considered the part-of-speech of the words within a  $(-2, +2)$  window of the target word. The Natural Language Toolkit (NLTK) was used for part-of-speech tagging (Bird, Klein, and Loper 2009).

In our WSD corpus there are 209,788 unique words and 37 possible parts-of speech for each word, which lead to a

total of 209,936 sparse features for use in our classifiers (1 for each possible unigram in the context window, and 1 for the part-of-speech in each of the 4 surrounding positions). Additionally, we discovered that eliminating rare words, namely those that occur in less than 5 documents in the WSD corpus, resulted in a vocabulary of 36,913 words and produces nearly identically performing classifiers with a much smaller number of total features (37,061). Therefore this smaller feature set was used in our final experiments. Furthermore, we experimented with features based on the co-occurrence of terms in the context window around a word and the definition text available in the gloss, but these features did not yield significant improvements and their results are not mentioned below.

## Classifiers

For this study we selected three supervised learning classifiers:

- AdaBoost (AB)
- Multinomial Naïve Bayes (MNB)
- Support Vector Machines (SVM)

These three algorithms were selected because they are commonly used on the word sense disambiguation task (Marquez et al. 2006). For each of these machine learning classifiers we made use of implementations that are part of WEKA, a freely available machine learning software suite (Hall et al. 2009). For the AdaBoost classifier, we used an implementation of the AdaBoost.M1 algorithm (Freund and Schapire 1996), which is a simple generalization of AdaBoost for multi-class classification, with decision stumps as the weak classifiers. For the Support Vector Machine classifier, we used an implementation based on John Platt’s sequential minimal optimization for training

the support vectors (Platt 1998).

## Word Sense Disambiguation Evaluation

For this study, we trained one multi-class classifier of each of the three types described above per word. The following sections describe how we trained the classifiers, and how each one performed.

### Experimental Setup

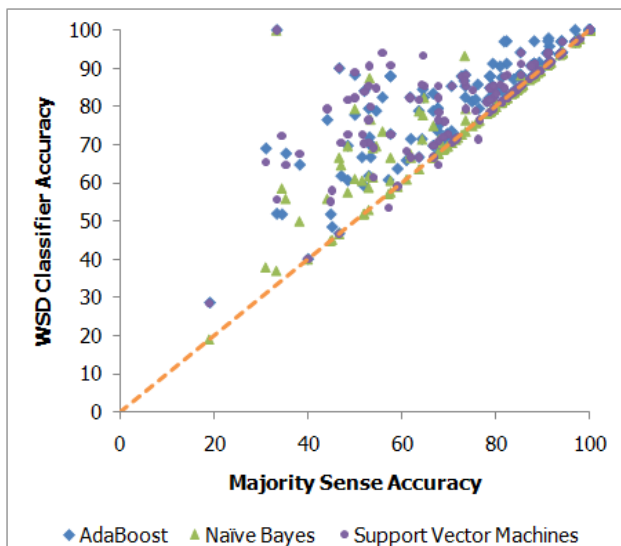
In order to evaluate how each of the classifiers performed, we ran an experiment on all 192 words in our data set. For each word, we trained three different multi-class classifiers, one for each of the algorithms described in the last section: AdaBoost, Multinomial Naïve Bayes, and Support Vector Machines. For each classifier, we trained on 66% of the instances available in the data set for the given word, and tested on the remaining instances. On average, each word had 89.88 instances available. Furthermore, for each classifier we tried, we experimented with unigram features with word windows (-  $w$ , +  $w$ ) of size 5 to 100 in steps of 5, in order to determine the best window size for the given classifier.

The baseline classifier that we chose to evaluate against was one that selects the majority sense, or, in other words, the sense of the word that occurs most often in the training corpus. The majority sense classifier had an average classification accuracy of 78.83% across all 192 words, and 69.18% on the 19 words used in the classroom study.

### Results

On the 192 words in our experiment, the AdaBoost classifier most consistently had accuracy equal to the best achieved per word, doing so for 72.92% of the words, while Multinomial Naïve Bayes, Support Vector Machines, and the baseline achieved this for 41.15%, 63.54%, and 39.53% of the words respectively. The fact that the baseline has the accuracy equal to the best achieved for about forty percent of the words illustrates how difficult it is to beat the majority sense baseline. Figure 1 compares the accuracy of the different classifiers against the baseline for each word. Most points in this figure lie above the dashed-line or at the line, signifying classification better or equal to the majority sense classifier.

The best average window sizes were 47.24 ( $\pm$  2.137), 47.40 ( $\pm$  2.074), and 47.92 ( $\pm$  2.061) the average classification accuracies were 85.22% ( $\pm$  0.9391), 81.54% ( $\pm$  1.112), and 85.24% ( $\pm$  0.882) for AB, MNB, and SVM respectively. The AB and SVM classifiers’ performances were clearly the best, both performed significantly better than MNB ( $p < 0.005$ ), but there wasn’t a clear statistically significant difference between the performance of the two algorithms. Furthermore, as a simple composite classifier, if one uses the strategy of selecting the classifier for each word that performed the best among the three types on the training instances, the test accuracy achieved is 88.03% ( $\pm$  0.7264), which is significantly better than using any one of the clas-



**Figure 1.** Comparison of WSD classifier performance on all words in data set against majority sense baseline, where dashed line is  $x$ -axis =  $y$ -axis and points above line signify classifier performance better than the baseline.

sifiers alone ( $p < 0.001$ ). The WSD classifier performances on the 19 words used in the classroom study are shown in Table 1, and show similar trends to the classifier performance on all 192 words in the data set.

## Classroom Study

In order to determine what effect automatic dictionary definition ordering has on L2 language learners, we conducted a user study with ESL students. We conducted an eight week *in-vivo* study at the University of Pittsburgh’s English Language Institute that contrasted different definition orderings, using a vocabulary tutor called REAP.

### Study Details

For the classroom study we had a population of 24 ESL college students, whose native languages included Arabic, Chinese, Korean, and Spanish. Group readings were given as class activities, centered on 19 focus words from the Academic Word List, presented only once in the reading passages, followed by practice *closed cloze questions*, also known as multiple choice fill-in the blank questions. Aside from having multiple senses, there was no particular reason

Word	# of Senses	Baseline Accuracy	Best Accuracy	Algorithm (Window)
abstract	2	71.43	71.43	Any (10)
aspect	4	100.00	100.00	Any (10)
brief	2	50.00	88.24	AB (45)
channel	3	46.67	90.00	AB (75), SVM (10)
civil	3	67.74	70.97	SVM (50)
commodity	2	88.24	91.18	AB (10), SVM (25)
confer	2	81.25	87.50	AB (50)
liberal	2	89.29	89.29	Any (10)
major	2	76.19	85.71	AB (65)
project	3	64.29	85.71	SVM (95)
quote	2	96.97	96.97	Any (10)
range	3	69.70	72.73	AB (10), SVM (50)
register	4	34.48	72.41	SVM (15)
secure	3	83.87	83.87	Any (10)
sole	2	85.71	85.71	Any (10)
structure	2	38.24	67.65	SVM (60)
trace	3	31.03	68.97	AB (10)
volume	2	81.82	96.97	AB (10)
welfare	2	57.58	90.91	SVM (50)

**Table 1.** Best WSD classifier performance on 19 words used in classroom study. Any signifies all 3 algorithms performing same as baseline “maioritv sense” classifier.

for selecting any of the focus words. During the reading activities, the definitions for the focus words were presented in one of three conditions, each of which was seen by each student at some point during the study:

- **Out of Order:** List of definitions presented where the best definition for the given context is not presented first, based on the original definition order in CALD, but with the first definition moved out of order if it was also the best definition for the given context.
- **Algorithm Best First:** List of definitions presented where the best definition for the given context as determined by WSD is presented first.
- **Algorithm Best:** Single best definition for the given context as determined by WSD.

The Out of Order group was selected as a control group to contrast with our other definition ordering schemes. An alternative ordering could have been “most frequent” ordering, but this was not used as it is unclear what corpus to base our frequency on, and it may be the case that the original dictionary definitions were based on this concept already. The Algorithm Best First was selected to contrast with the Out of Order group to determine whether ordering definitions mattered or if self-discovery of definitions, which we assume would be the students’ strategy for the case of Out of Order, is more important. The Algorithm Best group was selected to contrast against Algorithm Best First and Out of Order to see if one definition is sufficient for learning or if there is an advantage to providing multiple definitions.

A pre-test was administered at the beginning of the study, consisting of closed cloze questions centered on each focus word. A similar set of questions was presented to the students during the post-test which was administered one week after the final group reading. One week after the pre-test, six weekly reading activities were administered, each of which were focused on a single document that took approximately 20-30 minutes for our students to complete.

## Study Results

The data shows that the use of the REAP system significantly helped students improve their performance, as made evident by the average overall gains between the pre-test and post-test ( $p < 0.001$ , as measured by a paired t-test). The overall average normalized gain between pre-test and post-test was  $0.3341 (\pm 0.0568)$ .

Figure 2 compares the average post-reading practice question performance for words under different definition ordering schemes, where the average practice question accuracy was  $0.8596 (\pm 0.3041)$ ,  $0.9299 (\pm 0.0278)$ , and  $0.9048 (\pm 0.0354)$ , for Out of Order (OO), Algorithm Best First (ABF), and Algorithm Best (AB) respectively. On average, students performed better on practice questions when the WSD ordered definitions were provided (AB and ABF) during the readings, as opposed to the OO control group ( $p < 0.04$ ). This shows that the use of WSD to order the dictionary definitions shown to students result in better

practice question accuracy than showing an unordered list of definitions for questions administered directly after a reading.

When the students were asked if they found the definition order helpful after the readings, people preferred the ABF ordering over AB ( $p < 0.04$ ) and OO ( $p < 0.11$ ). This shows that students definitely preferred being showed multiple definitions over seeing a single one, and in particular they tended to like the ordering that had the correct definition (as determined by the WSD classifiers) for the given context at the top of the list, although this was not explicitly revealed to them. Additionally, for the practice questions on words whose definition list had more than one definition (ABF and OO word groups), the average time spent per question was significantly longer ( $p < 0.01$ ) than the condition when we provided only one definition (AB), with an average time per question of 174.23 ( $\pm 22.94$ ) seconds vs. 124.95 ( $\pm 23.50$ ) seconds. One possible explanation for this phenomenon is that in the AB condition students were focused on a single definition and when recalling this information during the practice questions they did not have to think about other possible definitions which

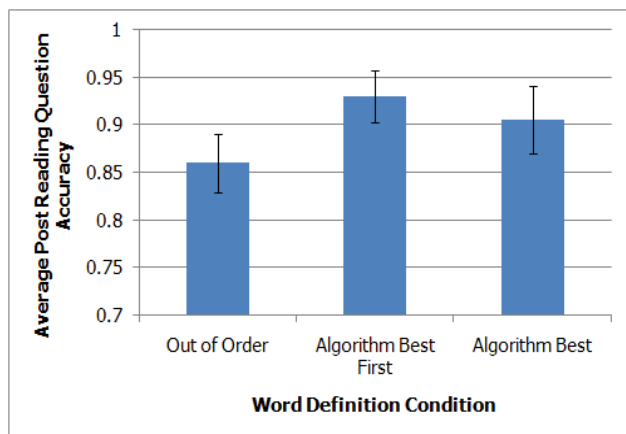
were provided in the ABF and OO conditions.

Figure 3 compares the average normalized gain between the pre-test and post-test for the word definition conditions, where the average gains were 0.3404 ( $\pm 0.0629$ ), 0.3635 ( $\pm 0.0599$ ), and 0.2871 ( $\pm 0.0879$ ), for OO, AB, and ABF respectively. While the differences in gains between each of the three groups were not statistically significant, the gains for the composite group (OO and ABF) for definition orderings consisting of more than one definition were generally higher than the condition when only one definition was provided (AB) ( $p < 0.15$ ). This shows that while showing a single definition during practice questions that are seen directly after the student first encounters the word during a reading may seem helpful, it may not be as good as showing multiple definitions when considering long term retention of the word, which seems to favor the conditions where all definitions are presented. Furthermore, the average pre-test to post-test gains on ABF words were slightly better than the OO control group, which implies that providing the WSD determined best definition as the first entry in the dictionary can be helpful to students, though this effect is much less noticeable on the long term retention task than on the questions given directly after the readings that first exposed the students to the words.

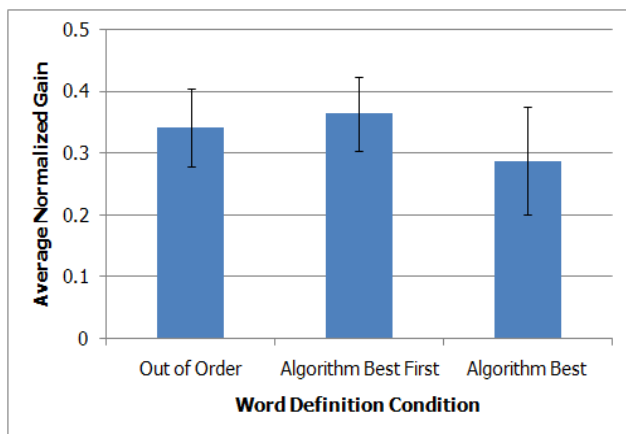
## Discussion

The results of our classroom study suggest that the ordering of dictionary definitions can make a positive difference for L2 vocabulary learners. While providing the single best definition can help vocabulary learners on assessment tasks administered shortly after the first exposure to the target words, it is much more advantageous to provide multiple definitions when longer term retention is considered, as made evident by the average pre-test to post-test gains. Furthermore, in addition to resulting in the highest average gains and post-reading accuracy, the Algorithm Best First ordering was the one most preferred by the students, as made evident by the results of surveys administered after each reading activity. This result seems to be related to the observation that students tend to stop reading the definition entries for a word after the top few, regardless of whether they provided the student with the correct definition for the given context (Kulkarni et al. 2008) which may explain why the students preferred the Algorithm Best First ordering over the Out of Order control group, as they may tend to read only the first or the first few definitions for a word, which would often give them the wrong definition for the Out of Order definition ordering scheme and the correct one for the Algorithm Best First definition ordering scheme.

The word sense disambiguation methodology employed in this study seemed to be effective in both predicting the most correct sense of a word and in providing a useful dictionary definition ordering to the students. This result is encouraging as it suggests that WSD may currently be good enough for use in intelligent language tutoring environments.



**Figure 2.** Effect of dictionary definition ordering on accuracy on post-reading practice questions.



**Figure 3.** Effect of dictionary definition ordering on normalized gain between pre-test and post-test cloze questions.

## Conclusion

We proposed that providing a dictionary whose definition ordering is sensitive to the provided reading context can help L2 language learning students learn vocabulary more effectively, and that word sense disambiguation techniques can be used to automatically order dictionary definitions. This was motivated by observations made by language educators that students sometimes select the wrong definition for a word when using a dictionary, which can lead to misunderstandings of the reading material.

We described a WSD methodology which makes use of well known algorithms and features, and leverages the power of crowd-sourcing to collect training data. Our WSD algorithms were able to achieve high accuracy on our data set and were effective in providing a useful dictionary definition ordering to the students.

Our classroom study showed that students benefited from having the best definitions (as determined by our algorithms) listed at the top of the definition list; students also preferred this ordering scheme more than the other ones tested. Furthermore, students retain vocabulary better when a list of definitions was provided, as opposed to being given the single best definition, as indicated by our post-test results.

One related future research problem is to determine whether there is a noticeable learning difference between using our WSD powered electronic dictionary and a physical dictionary. Furthermore, it may be interesting to see how much vocabulary learning would be affected if a dictionary was not provided to the students.

## Acknowledgements

This project is supported through the Cognitive Factors Thrust of the Pittsburgh Science of Learning Center which is funded by the US National Science Foundation under grant number SBE-0836012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

The authors would like to thank Betsy Davis, Sherice Clarke, Carol Harmatz, Anna Venishnick, Rebecca Wojcik, Chris Ortiz, and all teachers and administrators at the University of Pittsburgh's English Language Institute for their participation and input in the studies. Additionally, we would like to thank Gabriel Parent for his help in assembling the training data set for the WSD classifiers.

## References

Bogaards, P. 1998. Using dictionaries: Which words are looked up by foreign language learners?. In Atkins B., and Varantola, K. eds., *Studies of dictionary use by language learners and translators*, 151–157. Tübingen, Germany: Niemeyer.

Knight, S. 1994. Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal* 78: 285–299.

Koyama, T., and Takeuchi, O. 2004. How look up frequency affects EFL learning: An empirical study on the use of handheld-electronic dictionaries. In *Proceedings of CLaSIC 2004*, 1018–1024.

Prichard, C. 2008. Evaluating L2 readers' vocabulary strategies and dictionary use. *Reading in a Foreign Language* 20(2): 216–231.

Lupescu, S., and Day, R. R. 1993. Reading, dictionaries, and vocabulary learning. *Language Learning* 43: 263–287.

Summers, D. 1988. The role of dictionaries in language learning. In Carter, R., and McCarthy, M. eds., *Vocabulary and language teaching*, 111–125. London, United Kingdom: Longman.

Lomicka, L. 1998. "To gloss or not to gloss": An investigation of reading comprehension online. *Language Learning & Technology* 1(2): 41–50.

Yohsii, M. 2006. L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology* 10(3): 85–101.

Laufer, B., and Hill, M. 2000. What lexical information do L2 learners select in a call dictionary and how does it affect word retention?. *Language Learning & Technology* 3(2): 58–76.

Ide, J., and Jean, V. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics* 24(1): 1–40.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1): 97–123.

Pedersen, T., and Bruce, R. 1997. Distinguishing word senses in untagged text. In the *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*.

Kulkarni, A., Heilman, M., Eskenazi, M., and Callan, J. 2008. Word Sense Disambiguation for Vocabulary Learning. In the *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi M. 2006. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. In the *Proceedings of the 9th International Conference on Spoken Language*.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly* 34(2): 213–238.

Walter, E., ed. 2005. *Cambridge Advanced Learner's Dictionary*, 2nd Edition. Cambridge University Press.

Parent, G., and Eskenazi, M. 2010. Clustering dictionary definitions using Amazon Mechanical Turk. In the *Proceedings of the 11th NAACL-HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Mohammad, S., and Pederson, T. 2004. Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. In the *Proceedings of the Conference on Computational Natural Language Processing*.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Màrquez, L., Màrquez, L., Escudero, G., Martínez, D., Rigau, G. 2006. Supervised Corpus-based Methods for WSD. In Agirre, E., and Edmonds, P. eds. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1): 10–18.

Freund, Y., and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In the *Proceedings of the 13th International Conference on Machine Learning*.

Platt, J. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C., and Smola, A. eds., *Advances in Kernel Methods - Support Vector Learning*.