Sense-Specific Lexical Information for Reading Assistance

Soojeong Eom
Georgetown University
se48@georgetown.edu

Markus Dickinson
Indiana University
md7@indiana.edu

Rebecca Sachs
Georgetown University
rrs8@georgetown.edu

Abstract

To support vocabulary acquisition and reading comprehension in a second language, we have developed a system to display sense-appropriate examples to learners for difficult words. We describe the construction of the system, incorporating word sense disambiguation, and an experiment we conducted testing it on a group of 60 learners of English as a second language (ESL). We show that sense-specific information in an intelligent reading system helps learners in their vocabulary acquisition, even if the sense information contains some noise from automatic processing. We also show that it helps learners, to some extent, with their reading comprehension.

1 Introduction and Motivation

Reading texts in a second language presents the language learner with a number of comprehension problems, including the problem of interpreting words that are unknown or are used in unfamiliar ways. These problems are exacerbated by the prevalence of lexical ambiguity. Landes et al. (1998) report that more than half the content words in English texts are lexically ambiguous, with the most frequent words having a large number of meanings. The word *face*, for example, is listed in WordNet (Fellbaum, 1998) with twelve different nominal senses; although not all are equally prevalent, there is still much potential for confusion.

To address this, we have designed an online reading assistant to provide sense-specific lexical information to readers. By *sense-specific*, we refer to information applicable only for one given sense (meaning) of a word. In this paper, we focus on the system design and whether such a system can be beneficial. Our experiment with learners illustrates the effectiveness of such information for vocabulary acquisition and reading comprehension.

The problem of lexical ambiguity in reading comprehension is a significant one. While dictionaries can help improve comprehension and acquisition (see, e.g., Prichard, 2008), lexical ambiguity may lead to misunderstandings and unsuccessful vocabulary acquisition (Luppescu and Day, 1993), as learners may become confused when trying to locate an appropriate meaning for an unknown word among numerous sense entries. Luppescu and Day showed that readers who use a (printed) dictionary have improved comprehension and acquisition, but to the detriment of their reading speed.

For electronic dictionaries as well, lexical ambiguity remains a problem (Koyama and Takeuchi, 2004; Laufer and Hill, 2000; Leffa, 1992; Prichard, 2008), as readers need specific information about a word as it is used in context in order to effectively comprehend the text and thus learn the word. Kulkarni et al. (2008) demonstrated that providing readers with sense-specific information led learners to significantly better vocabulary acquisition than providing them with general word meaning information.

We have developed an online system to provide vocabulary assistance to learners of English as a Second Language (ESL), allowing them to click on unfamiliar words and see lexical information—target word definitions and examples—relevant to that particular usage. We discuss previous online

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 316–325,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

systems in section 2. Importantly, the examples we present are from the COBUILD dictionary (Sinclair, 2006), which is designed for language learners. To present these for any text, our system must map automatic word sense disambiguation (WSD) system output (using WordNet senses (Fellbaum, 1998)) to COBUILD, as covered in section 3, where we also describe general properties of the web system.

The main contribution of this work is to investigate whether high-quality sense-specific lexical information presented in an intelligent reading system helps learners in their vocabulary acquisition and reading comprehension and to investigate the effect of automatic errors on learning. We accordingly ask the following research questions:

1. Does sense-specific lexical information facilitate vocabulary acquisition to a greater extent than: a) no lexical information, and b) lexical information on all senses of each chosen word?
2. Does sense-specific lexical information facilitate learners' reading comprehension?

The method and analysis for investigating these questions with a group of 60 ESL learners is given in section 4, and the results are discussed in section 5.

2 Background

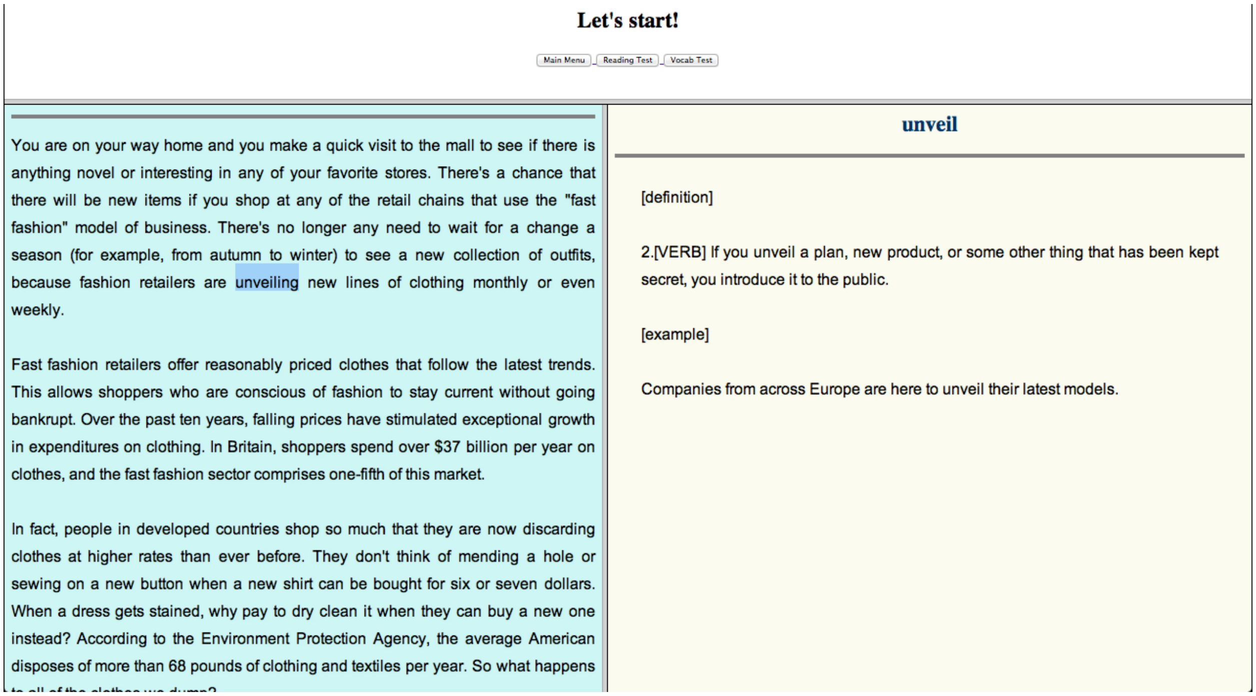
While there are many studies in second language acquisition (SLA) on providing vocabulary and reading assistance (e.g., Prichard, 2008; Luppescu and Day, 1993), we focus on outlining intelligent computer-assisted language learning (ICALL) systems here (see also discussion in Dela Rosa and Eskenazi, 2011). Such systems hold the promise of alleviating some problems of acquiring words while reading by providing information specific to each word as it is used in context (Nerbonne and Smit, 1996; Kulkarni et al., 2008). The GLOSSER-RuG system (Nerbonne and Smit, 1996) disambiguates on the basis of part of speech (POS). This is helpful in distinguishing verbal and nominal uses, for example, but is, of course, ineffective when a word has more than one sense in the same POS (e.g., *face*). More effective is the REAP Tutor (Heilman et al., 2006), which uses word sense disambiguation to provide lexicographic information and has been shown to benefit learners by providing sense-specific lexical information (Dela Rosa and Eskenazi, 2011; Kulkarni et al., 2008).

We build from this work by further demonstrating the utility of sense-specific information. What distinguishes our work is how we build from the notion that the lexical information provided needs to be tuned to the capacities of ESL learners. For example, definitions and illustrative examples should make use of familiar vocabulary if they are to aid language learners; example sentences directly taken from corpora or from the web seem less appropriate because the information in them might be less accessible (Groot, 2000; Kilgarriff et al., 2008; Segler et al., 2002). On the other hand, examples constructed by lexicographers for learner dictionaries typically control for syntactic and lexical complexity (Segler et al., 2002). We thus make use of examples from a dictionary targeting learners.

Specifically, we make use of the examples from the Collins COBUILD Student's Dictionary (Sinclair, 2006), as it is widely used by ESL learners. The content in COBUILD is based on actual English usage and derived from analysis of a large corpus of written and spoken English, thereby providing authentic examples (Sinclair, 2006). COBUILD also focuses on collocations in choosing example sentences, so that the example sentences present natural, reliable expressions, which can play an important role in learners' vocabulary acquisition and reading comprehension. We discuss this resource more in section 3.3.

3 The web system

To support vocabulary acquisition and reading comprehension for language learners, we have designed a system for learners to upload texts and click on words in order to obtain sense-appropriate examples for difficult words while reading, as shown in figure 1. Although the experiment reported upon here focused on 2 preselected texts, the system is able to present lexical information for any content words. Beyond the web interface, the system has three components: 1) a system manager, 2) a natural language processing (NLP) server, and 3) a lexical database.



You are on your way home and you make a quick visit to the mall to see if there is anything novel or interesting in any of your favorite stores. There's a chance that there will be new items if you shop at any of the retail chains that use the "fast fashion" model of business. There's no longer any need to wait for a change a season (for example, from autumn to winter) to see a new collection of outfits, because fashion retailers are unveiling new lines of clothing monthly or even weekly.

Fast fashion retailers offer reasonably priced clothes that follow the latest trends. This allows shoppers who are conscious of fashion to stay current without going bankrupt. Over the past ten years, falling prices have stimulated exceptional growth in expenditures on clothing. In Britain, shoppers spend over $37 billion per year on clothes, and the fast fashion sector comprises one-fifth of this market.

In fact, people in developed countries shop so much that they are now discarding clothes at higher rates than ever before. They don't think of mending a hole or sewing on a new button when a new shirt can be bought for six or seven dollars. When a dress gets stained, why pay to dry clean it when they can buy a new one instead? According to the Environment Protection Agency, the average American disposes of more than 68 pounds of clothing and textiles per year. So what happens to all of the clothes we dump?

unveil

[definition]

2.[VERB] If you unveil a plan, new product, or some other thing that has been kept secret, you introduce it to the public.

[example]

Companies from across Europe are here to unveil their latest models.

Figure 1: A screenshot showing the effect of clicking on *unveiling* and receiving sense-specific information

3.1 System manager

The system manager controls the interaction among each learner, the NLP server, and the lexical database. When the manager receives a raw text as an input from the learner, it first sends the input text to the server and returns an analyzed text (i.e., tokenized, POS-tagged, and sense-tagged) back to the learner, with content words made clickable. Then, when the learner clicks on a word while reading, the manager sends the word with its sense information to the lexical database and brings the word with its sense-specific lexical information back to the learner from the lexical database.

Upon completion of the reading, the manager sends the learner to a page of tests—i.e., a reading test and a vocabulary test, as described in section 4—and records the responses.

3.2 NLP preprocessing

To convert raw input into a linguistically-analyzed text, the system relies on several basic NLP modules for tokenizing, lemmatizing, POS tagging, and collocation identification. Although for some internal testing with different WSD systems we used other third-party software (e.g., the Stanford POS tagger (Toutanova et al., 2003)), our word sense disambiguator (see below) provides tokenization, lemmatization, and POS tagging, as well as collocation identification. Since the words making up a collocation may be basic, learners can easily overlook them, and so we intend to improve this module in the future, to reduce underflagging of collocations.

3.3 Lexical database

The lexical database is used to provide a sense-appropriate definition and example sentences of an input word to a learner. To obtain the sense-appropriate information, we must perform word sense disambiguation (WSD) on the input text. We use SenseRelate::AllWords (SR:AW) (Pedersen and Kolhatkar, 2009) to perform WSD of input texts, as this system has broad coverage of content words. Given that SR:AW does not outperform the most frequent sense (MFS) baseline, we intend to explore using the MFS in the future, as well as other WSD systems, such as SenseLearner (Mihalcea and Csomai, 2005). However, the quality of SR:AW (F-measure of 54–61% on different corpora) is sufficient to explore in our system and gives us a point to work from. Indeed, as we will see in section 5.3, while SR:AW makes errors, vocabulary learning is, in some ways, perhaps not dramatically impeded.

Even with a WSD system, pointing to appropriate examples is complicated by the fact that the database of learner-appropriate examples is from one repository (COBUILD, see section 2), while automatic WSD systems generally use senses from another (WordNet). The lexical database, then, is indexed by WordNet senses, each of which points to an appropriate corresponding COBUILD sense. While we would prefer disambiguating COBUILD senses directly, we are not aware of any systems which do this or any COBUILD sense-tagged data to train a system on. If the benefits for vocabulary acquisition gained by providing learner-friendly examples from COBUILD merit it, future work could explore building a collection of COBUILD-tagged data to train a WSD system—perhaps a semi-automatic process using the automatic system we describe next.

To build a lexical database covering all words, we built a word sense alignment (WSA) system; this is also in line with a related research agenda investigating the correspondences between sense inventories (Eom et al., 2012). Space limitations preclude a more detailed discussion, but the WSA system works by running SR:AW on COBUILD examples in order to induce a basic alignment structure between WordNet and COBUILD. We then postprocess this structure, relying on a heuristic of favoring flatter alignment structures—i.e., links spread out more evenly between senses in each inventory.[1] Iteratively replacing one link with another, to give flatter structures, we weight each type of proposed alignment and accept a new alignment if the weight combined with the probability originally assigned by the WSD system is the best improvement over that of the original alignment structure. After all these steps, the alignments give the lexical database for linking WSD output to COBUILD senses.

We consider alignment structures wherein each WordNet sense maps to exactly one COBUILD sense, to match the task at hand, i.e., mapping each disambiguated WordNet sense to a single set of COBUILD examples. This assumption also makes postprocessing feasible: instead of considering an exponential number of alignment structures, we consider only a polynomial number.

Having collected alignment judgments from linguistics students and faculty, we evaluated the system against a small set of nine words, covering 63 WordNet senses (Eom et al., 2012). The WSA system had a precision of 42.7% (recall=44.5%) when evaluating against the most popular sense, but a precision of 60.7% (recall=36.5%) when evaluating against all senses that seem to be related. We focus on precision since it is important to know whether a learner is being pointed to a correct set of examples or not; whether there are other possibly relevant examples to show is less important. In Eom et al. (2012), we discuss some difficulties of aligning between the two resources in the general case; while some senses go unaligned between the resources, this was not the case for the words used in this study.

For this study, since we use pre-determined input texts, we also created gold-standard information, where each word in the text is manually given a link to the appropriate COBUILD information; note that here there is no intermediate WordNet sense to account for. This lets us gauge: a) whether the gold-standard information is helpful to learners, and b) comparatively speaking, what the effects are of using the potentially noisy information provided by the functioning system.

4 The study

We now turn to evaluating whether this set-up of providing sense-specific lexical information can lead learners to improve their vocabulary acquisition and their reading comprehension.

4.1 Method

4.1.1 Participants

The participants were recruited from three universities and a private institute in Seoul, Korea, giving 60 participants (34 male, 26 female). They ranged in age from 21 to 39 (avg.=23.8) and the length of studying English ranged from 8 to 25 years (avg.=11.32).

The 40 participants from the three universities were taking English courses to prepare for English proficiency testing. The 20 participants from the private institute were mostly university graduates

[1] The general idea is to use information about the alignment structure as a whole; flatter alignments is a convenient heuristic, in lieu of having any other additional information.

taking teacher training courses designed for elementary English teachers. All participants were intermediate-level learners, scoring between 15 and 21 on the reading section of the TOEFL iBT®. We targeted intermediate learners, so as to test the system with learners generally able to understand texts, yet still encounter many unknown words.

The 60 participants were randomly assigned to one of four groups, with 15 participants in each group. The first three received some treatment, while the fourth was a control group:

1. Gold Senses (GS): reading with support of gold standard sense-specific lexical information
2. System Senses (SS): reading with support of system-derived sense-specific lexical information
3. All Senses (AS): reading with support of lexical information of all senses of the chosen word
4. No Senses (NS): reading without any support of lexical information

For example, when presented with the example in (1), if *chains* is clicked, the GS learners see the correct sense, as in (2a), along with associated example sentences (not shown). The automatic system happens to be incorrect, so the SS learners see a related, though incorrect, sense and examples, as in (2b). The AS learners will see those two senses and examples, as well as the three others for *chain*. And the NS learners have no chance to click on a word.

(1) There's a chance that there will be new items if you shop at any of the retail **chains** that use the "fast fashion" model of business.

(2) a. **Gold:** A chain of shops, hotels, or other businesses is a number of them owned by the same person or company.

b. **System:** A chain of things is a group of them existing or arranged in a line.

4.1.2 Materials

Reading texts After piloting various reading texts and drawing on the ESL teaching experience of two of the authors, two texts deemed appropriate for learners at the (high-)intermediate level were adopted: *Fashion Victim* (adapted from *Focus on Vocabulary 1: Bridging Vocabulary* (Schmitt et al., 2011), 589 words) and *Sleep Research* (adapted from *The Official SAT Study Guide* (The College Board, 2009), 583 words).

Fashion Victim	*Sleep Research*
resilient.a, chain.n, conscience.n, cradle.n, expenditure.n, mend.v, outfit.n, sector.n, unveil.v	alternate.a, trivial.a, deliberately.r, aspect.n, fatigue.n, obedience.n, agitate.v, banish.v, indicate.v, resist.v, trigger.v

Table 1: Target words used in the study

The texts were modified to simplify their syntax, to use more ambiguous words in order to allow for a stronger test of the system, and to shorten them to about 600 words. The texts were placed in the online system, and all content words were made clickable.

Target words A total of 20 target words (9 from *Fashion Victim*, 11 from *Sleep Research*) were selected by piloting a number of possible words with 20 learners from a similar population and identifying ones which were the most unfamiliar, which also had multiple senses. They appear in table 1.

Reading comprehension tests For reading comprehension, two tests were developed, each with 4 multiple-choice and 6 true-false questions. The questions focused on general content, and participants could not refer back to the text to answer the questions. For the multiple-choice questions, more than one answer could be selected, and each choice was scored as 1 or 0 (e.g., for 5 choices, the maximum score for the question was 5); for the true-false questions, answers were scored simply 1 or 0. The maximum score for a test was 21.

Vocabulary tests There were one pretest and four immediate posttests, one of which had the same format as the pretest. The pretest and all immediate posttests had the same 30 words (20 target and 10 distractor words). Of 10 distractors, five were words appearing in the text (*obscure.a*, *correlation.n*, *intervention.n*, *discipline.v*, *facilitate.v*), and five were target words but used with a sense that was different from the one used in the reading passage (*deliberately.r*, *chain.n*, *outfit.n*, *mend.v*, *indicate.v*). Each test consisted of a word bank and sentences with blanks (cf. Kim, 2008). For the pretest, the sentences were taken from other sources, whereas the posttest sentences came from the reading texts themselves.

Although we used four posttests in order to test different kinds of vocabulary learning (giving more or fewer hints at meaning), we focus on one posttest in this paper, the one which matches the form of the pretest. Each correct answer was scored as 1; incorrect as 0.

4.1.3 Procedure

The pretest was administered two weeks before the actual experiment and posttests, so as to prevent learners from focusing on those words. Participants who knew more than 16 out of the 20 target words were excluded from the experiment.

After reading one text, learners took a reading comprehension test. Then, they did the same for the second text. After these two rounds, they took the series of vocabulary posttests.

4.1.4 Data analysis

We ran a variety of tests to analyze the data.[2] First, we ran Levene's test of homogeneity of variances, to test whether the variances of the error between groups were equal at the outset of the study. This makes it clearer that the effects from the main tests are due to the variables of interest and not from inherent differences between groups (Larson-Hall, 2010).

Secondly, to test the first research question about whether participants show better vocabulary acquisition with sense-specific lexical information, we used a repeated-measures analysis of variance (RM ANOVA). Time (pre/post) was the within-subject variable and Group (GS, SS, AS, NS) was the between-subject. Post-hoc pairwise comparisons were run in the case of significant results, to determine which groups differed from each other. We also examined the pre-post gain only for the target words which were clicked and for which we might thus expect more improvement.

Thirdly, to test the second research question about whether participants improved in reading comprehension, we used a one-way ANOVA, with reading comprehension scores as a dependent variable and the four groups as an independent variable, to explore if there was any significant main effect of the group on reading comprehension scores. Post-hoc tests were then used, in order to determine specifically which groups differed from each other.

	Pretest		Posttest	
	Mean	SD	Mean	SD
GS	10.73 (54%)	3.43	15.93 (80%)	3.96
SS	10.93 (55%)	2.82	15.47 (77%)	3.80
AS	10.87 (54%)	3.34	13.47 (67%)	3.83
NS	10.87 (54%)	3.25	11.27 (56%)	3.39

Table 3: Descriptive statistics across groups for vocabulary acquisition (*Mean* = average, *SD* = standard deviation, percentage out of all 20 answers in parentheses)

In order to gauge the effect of automatic system errors—distinguishing the SS (System Senses) and GS (Gold Senses) conditions—on vocabulary acquisition, we also examined target words where the system gave incorrect information.

5 Results and Discussion

5.1 Vocabulary acquisition

Since the first research question is to examine the improvement between the pretest and the posttest, the test of homogeneity of variance was carried out to ensure that the pretest/posttest scores of the participants across the four groups showed similar variances. Levene's test of homogeneity of variances suggested that the 4 groups could be considered to have similar variances on both the pretest ($F(3, 55) = 0.49$, $p = 0.69$) and the post-test ($F(3, 56) = 0.13$, $p = 0.94$), meaning that this assumption underlying the use of ANOVA was met.

Looking at the descriptive statistics in table 3, none of the groups differed from each other by more than a quarter of a point (or 1 percentage point) on the pretest. Thus, the groups are also comparable with respect to their levels of performance on the pre-test.

Turning to the results of the treatments in table 3, the four groups show larger differences on their posttest. The GS and SS groups show the clearest gains, suggesting greater vocabulary acquisition than the AS and NS groups, as expected. If we look at percentage gain, GS gained 26% and SS 23%,

[2] We used SPSS,version 20.0, `http://www-01.ibm.com/software/analytics/spss/`

Source	df	$df2$	F	p	Partial Eta^2	Obs. Power
	Test of Within-Subjects Effects					
Time	1	56	62.67	<**0.01**	0.53	1.00
Time*Group	3	56	7.20	<**0.01**	0.28	0.98
	Test of Between-Subjects Effects					
Group	3	56	1.71	**0.18**	0.08	0.42

Table 2: Results of RM ANOVA comparing vocabulary test scores across the four groups over time

while AS gained only 13% and NS 2%.

In order to examine whether the above differences among groups were statistically significant, a repeated-measures ANOVA was run on those pretest and posttest scores, with Group as the between-subject variable and Time as the within-subject variable. The results of the RM ANOVA are presented in table 2.

With respect to the within-subject variable, the effect of Time shows a statistically significant difference ($F(1, 56) = 62.67$, $p < .001$, partial eta^2= 0.53). In other words, not considering Group, there is evidence of improvement from pre to posttest.

Most crucially related to the first research question about whether the groups would have different amounts of vocabulary acquisition over time, we see a significant Time*Group effect ($F(3, 56) = 7.20$, $p < .001$, partial eta^2= 0.28). The partial eta^2 values for Time (0.53) and Time*Group (0.28) in table 2 represent large effect sizes which thus provide strong evidence for the differences.

Two sets of post-hoc comparisons were conducted. The first comparisons, in table 4, show significant mean differences between the pretest and posttest for three groups (GS, SS, AS), whereas no significant difference is observed in the NS group, meaning that the three groups who received lexical information showed improvement whereas the group who received no information did not.

Then, a second set of post-hoc tests were run to compare the three groups which showed significant pre-post gains (GS, SS, AS). In table 5, the Contrast Estimate (Est.) looks at the differences in the mean pre-post gains and shows that the GS group is significantly different from the AS group, whereas the difference between the mean gains of the SS and AS groups is not quite significant. (The GS-SS contrast is non-significant.) In other words, these post-hoc comparisons on the Time*Group interaction effect found a significant difference between the GS and AS groups in their vocabulary learning over time, with the GS group showing greater pretest-posttest improvement, whereas the SS's group apparent advantage over the AS group with their mean gains fell slightly short of statistical signficance.

Group	I	J	Mean Diff.	Std. Error	p
GS	pre	post	-5.20	0.80	<**0.01**
SS	pre	post	-4.23	0.80	<**0.01**
AS	pre	post	-2.60	0.80	<**0.01**
NS	pre	post	-0.40	0.80	0.62

Table 4: Post-hoc comparisons for Time*Group, for vocabulary acquisition

Group Contrast	Est.	Sig.
GS-AS	2.60	**0.02**
SS-AS	1.93	0.09
GS-SS	0.67	0.56

Table 5: Contrast results for Time*Group, where the dependent variable is the difference in mean pre-post gains

Clicked words In addition to analyzing learners' performance on the overall scores of their pretest and posttest, we examine their performance over their pretest and posttest only on words they clicked while reading, as well as how much they clicked. In the three treatments, we find: GS, 28.27 words clicked on average (7.00 target words); SS, 21.80 (5.93); and AS, 20.87 (5.60). Although these differences are not statistically significant, the apparent trend may suggest that the GS group realized

	Pretest		Posttest		
	Mean	SD	Mean	SD	Gain
GS	40%	32%	85%	22%	45%
SS	25%	18%	81%	25%	56%
AS	23%	25%	68%	32%	45%

Table 6: Descriptive statistics for vocabulary acquisition for clicked words (percentage correct)

they could get high-quality lexical information from clicking words and so clicked more often.

Examining only clicked target words, the test of homogeneity confirmed the error variance of all participants were equivalent at the outset of the study ($p = 0.15$). The percentages correct of the words that were clicked in the pretest and posttest are in table 6. The pre to post gain here conveys a general trend: for the words participants clicked on, they showed improvement, with larger gains than for all words (compare the best gain of 26% in table 3). As with all words, in the RM ANOVA the effect of Time shows a statistically significant difference ($F(1, 42) = 96.20$, $p < 0.01$). However, the effect of Time*Group shows no significant difference in this case ($F(2, 42) = 0.60$, $p = 0.55$).

Despite non-significance, two potentially interesting points emerge which can be followed up on in the future: 1) descriptively speaking, the SS group shows the largest gain between pretest and posttest (56%); and 2) the AS group shows as much improvement as the GS group (45%). This may come from the fact that the number of senses listed for many clicked words was small enough (e.g., 2–3) to find an appropriate sense. Future work could investigate a greater number of target words to verify and shed more light on these trends.

Discussion In sum, our results suggest a positive answer to the first research question about whether sense-specific lexical information leads learners to better vocabulary acquisition. The results from several different analyses suggest that: 1) learners provided with lexical information during reading have more vocabulary acquisition, with sense-specific information having a greater increase; 2) learning gains appear to be greater for the subset of *clicked* target words than for all words (though further research is needed to substantiate this); and 3) they seem to check the meaning more when disambiguated correctly (again needing further research).

	Mean	SD
GS	35.80 (85%)	3.98
SS	37.07 (88%)	2.46
AS	34.93 (83%)	3.08
NS	33.27 (79%)	3.69

Table 7: Descriptive statistics for reading comprehension

Source	df	$df2$	F	p
Group	3	56	4.01	**0.01**

Table 8: Results of one-way ANOVA for reading comprehension scores

5.2 Reading comprehension

The second research question explores whether sense-specific lexical information facilitates reading comprehension. The descriptive statistics for reading comprehension mean scores of the four groups are in table 7. The difference among the reading comprehension mean scores of the four groups was within about 4 points, corresponding to a 9% difference (SS, 88%; NS, 79%). The GS and SS groups have the highest values, but only small differences.

In order to examine whether the above differences among groups were statistically significant, a one-way ANOVA was run on reading comprehension scores. The test of homogeneity of variances confirmed the error variances were equivalent ($p = 0.42$). The results of the one-way ANOVA are in table 8.

As shown, the effect of Group shows a statistically significant difference, indicating that the groups are different in their reading comprehension ($F(3, 56) = 4.01$, $p = 0.01$). With this significant difference in reading comprehension performance, it is necessary to locate where the differences existed among the groups. Tukey post0hoc tests compared all four groups in pairs and revealed a significant difference between the SS group and the NS group ($p = 0.007$), with no significant differences between the other pairs.[3]

To some extent, the results support the idea that

[3] GS vs. SS: $p = 0.68$; GS vs. AS: $p = 0.87$; GS vs. NS: $p = 0.12$; SS vs. AS: $p = 0.24$; AS vs. NS: $p = 0.46$.

System	Pretest	Posttest	Accuracy
Appropriate	+ (16)	+ (14)	88% (14/16)
	- (42)	+ (32)	76% (32/42)
Inappropriate	+ (12)	+ (10)	83% (10/12)
	- (18)	+ (9)	**50%** (9/18)

Table 9: Pre/Posttest performance for SS condition, summed over learners, broken down by whether system sense was appropriate (+ = learner got correct; - = learner got incorrect; numbers in parentheses = actual values)

sense-specific lexical information facilitates learners' reading comprehension. Curiously, the GS group, which received more accurate sense information than the SS group, was not found to outperform the control group ($p = 0.12$)—despite descriptively showing slightly higher reading comprehension scores. This issue warrants future investigation.

5.3 Quality of sense information

We have observed some differences between the Gold Senses (GS) and System Senses (SS) conditions, but we still want to explore to what extent the learners in SS group were impacted by words which were incorrectly disambiguated. There were nine words which the automatic system incorrectly assigned senses to (*inappropriate target-sense words*),[4] and eleven words which it correctly assigned. One can see the different performance for these two types in table 9, for words that learners clicked on.

There are two take-home points from this table. First, when learners were correct in the pretest, they generally did not un-learn that information, regardless of whether they were receiving correct sense information or not (88% vs. 83%). This is important, as it seems to indicate that wrong sense information is not leading learners astray. However, the second point is that when learners were wrong in the pretest, they were in general able to learn the sense with correct information (76%), but not as effectively when given incorrect information (50%). This, unsurprisingly, shows the value of correct sense information.

6 Summary and Outlook

We have developed a web system for displaying sense-specific information to language learners and tested it on a group of 60 ESL learners. We showed that sense-specific information in an intelligent reading system can help learners in their vocabulary acquisition and, to some extent, may also help with overall reading comprehension. We also showed preliminary results suggesting that learners might learn more of the words whose definitions they check than words they simply encounter while reading. We can also be optimistic that, while there is still much room for improvement in presenting sense information automatically, errors made by the system do not seem to interfere with language learners' previously-known meanings.

There are a number of avenues to pursue in the future. One thing to note from the results was that the group receiving help in the form of all senses (AS) demonstrated relatively high performance in vocabulary acquisition and reading comprehension, at times similar to the groups receiving sense-specific information (GS, SS). This may be related to the small number of sense entries of the target words (average = 2.95), and a further study should be done on target words with more sense entries, in addition to validating some of the preliminary results presented in this paper regarding clicked words. Secondly, the word sense disambiguation methods and construction of the lexical database can be improved to consistently provide more accurate sense information. Finally, as mentioned earlier, there are preprocessing improvements to be made, such as improving the search for collocations.

Acknowledgments

We would like to thank Stephanie Dickinson and Chelsea Heaven from the Indiana Statstical Consulting Center (ISCC) for their assistance, as well as Graham Katz and the three anonymous reviewers for their useful comments.

References

Kevin Dela Rosa and Maxine Eskenazi. 2011. Impact of word sense disambiguation on ordering dictionary definitions in vocabulary learning tutors. In *Proceedings of FLAIRS 2011*.

[4] *aspect.n*, *chain.n*, *conscience.n*, *expenditure.n*, *sector.n*, *agitate.v*, *banish.v*, *indicate.v*, *resist.v*

Soojeong Eom, Markus Dickinson, and Graham Katz. 2012. Using semi-experts to derive judgments on word sense alignment: a pilot study. In *Proceedings of LREC-12*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Peter J. M. Groot. 2000. Computer assisted second language vocabulary acquisition. *Language Learning and Technology*, 4(1):60–81.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2006. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the 9th International Conference on Spoken Language Processing*.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX-08*. Barcelona.

YouJin Kim. 2008. The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, 58:285–325.

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

Anagha Kulkarni, Michael Heilman, Maxine Eskenazi, and Jamie Callan. 2008. Word sense disambiguation for vocabulary learning. In *Ninth International Conference on Intelligent Tutoring Systems*.

Shari Landes, Claudia Leacock, and Randee I. Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: an electronic lexical database*, chapter 8, pages 199–216. MIT.

Jenifer Larson-Hall. 2010. *A guide to doing statistics in second language research using SPSS*. Routledge, New York, NY.

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

Vilson J. Leffa. 1992. Making foreign language texts comprehensible for beginners: An experiment with an electronic glossary. *System*, 20(1):63–73.

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

Rada Mihalcea and Andras Csomai. 2005. SenseLearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 53–56. Ann Arbor, MI.

John Nerbonne and Petra Smit. 1996. GLOSSER-RuG: in support of reading. In *Proceedings of COLING-96*.

Ted Pedersen and Varada Kolhatkar. 2009. WordNet::SenseRelate::AllWords - a broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of HLT-NAACL-09*. Boulder, CO.

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

Diane Schmitt, Norbert Schmitt, and David Mann. 2011. *Focus on Vocabulary 1: Bridging Vocabulary*. Pearson ESL, second edition.

Thomas Segler, Helen Pain, and Antonella Sorace. 2002. Second language vocabulary acquisition and learning strategies in ICALL environments. *Computer Assisted Language Learning*, 15(4):409–422.

John Sinclair, editor. 2006. *Collins COBUILD Advanced Lerner's English Dictionary*. Harper Collins.

The College Board. 2009. *The Official SAT Study Guide*. College Board, second edition.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259.