

AUTOMATIC PRESENTATION OF SENSE-SPECIFIC LEXICAL INFORMATION
IN AN INTELLIGENT LEARNING SYSTEM

A Dissertation
Submitted to the Faculty of the
Graduate School of Arts and Science
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Soojeong Eom, M.S.

Washington, DC
August 1, 2012

UMI Number: 3524085

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3524085

Copyright 2012 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright 2012 by Soojeong Eom
All Rights Reserved

AUTOMATIC PRESENTATION OF SENSE-SPECIFIC LEXICAL INFORMATION
IN AN INTELLIGENT LEARNING SYSTEM

Soojeong Eom, M.S.

Thesis Advisor: Graham E. Katz, Ph.D., Markus Dickinson, Ph.D.

ABSTRACT

Learning vocabulary and understanding texts present difficulty for language learners due to, among other things, the high degree of lexical ambiguity. By developing an intelligent tutoring system, this dissertation examines whether automatically providing enriched sense-specific information is effective for vocabulary learning and reading comprehension of second language learners. The system developed in this study contributes to an extended understanding of how NLP techniques can be applied more effectively in an educational environment.

The system allows learners to upload texts and click on any content word in order to obtain sense-appropriate lexical information for unfamiliar or unknown words during reading. The system consists of three components: (1) the system manager controls the interaction among each learner, the NLP server, and the lexical database; (2) the NLP server converts a raw input text to a linguistically-analyzed text; (3) the lexical database is used to provide a sense-appropriate definition and example sentences of a word to the learner. To obtain the sense-appropriate information, the system first performs word sense disambiguation (WSD) on the input text. Pointing to appropriate examples tuned for language learners, however, is complicated by the fact that the database of examples is from one repository (COBUILD), while automatic WSD systems generally rely on senses from another (WordNet). The lexical database, then, is indexed by WordNet

senses, each of which points to an appropriate corresponding COBUILD sense. The fact that every sense inventory has its own standards of sense distinction poses a serious problem in integrating these inventories into one. To redirect an input WordNet sense to a corresponding COBUILD sense, thus, a word sense alignment algorithm was developed, following a heuristic of favoring flatter alignment structures.

With this system, an empirical study was conducted with 60 intermediate learners of English as a second language to examine whether this system can lead learners to improve their vocabulary acquisition and reading comprehension. The findings show that learners demonstrated higher performance when receiving sense-specific information. Furthermore, the qualitative examination of the effect of automatic system errors show that, although learners showed learning regardless of the appropriateness of lexical information, they still showed relatively greater learning when given appropriate lexical information.

This dissertation is dedicated to my parents.

ACKNOWLEDGEMENTS

The long and arduous journey that has resulted in this dissertation is the direct result of the efforts and support of many key individuals without whom it would be difficult to imagine such an endeavor could have been possible to even embark upon. To say that I am grateful is a massive understatement and I truly lack the words to properly express my wholehearted feeling of indebtedness to all those around me who gave of themselves thoughtlessly and without reservation.

I would like to thank, first and foremost, my advisors Dr. Graham Katz and Dr. Markus Dickinson, who have guided me and provided their insights without which I would most likely have been lost along the way. Dr. Katz has been incredibly generous with his time and indefatigable as a source of ideas – our long discussions have always been invigorating yet enjoyable and perhaps more importantly, a rigorous training ground for my humble beginnings as a research scientist.

I was also extremely lucky to be able to work with Dr. Dickinson, who has been a constant source of inspiration as I overcame the numerous hurdles throughout my Ph.D. career. I cannot forget the thrilling moment I shared with him when the topic of this research was first determined. His wisdom, support, and friendship have been and will likely continue to be integral to my academic pursuits, for which I am forever grateful. Dr. Dickinson's tremendous enthusiasm and constant encouragement in leading his students have always represented to me the prime model of what it means to be a top researcher and an educator.

My deep gratitude also extends to the members of my dissertation committee, Dr. Paul Portner and Dr. Jeffrey Connor-Linton, for their insightful comments and suggestions. Many thanks in particular to Dr. Portner who opened the path for me to start my graduate life at Georgetown. From the very beginning up to the present, his help as a professor, a committee member, and as the Director of Graduate Studies, was crucial in ensuring the well-rounded progression of my academic and professional life. I also would like to extend a very special thanks to Dr. Connor-Linton for his thoughtful advice and very kind support. I truly respect him for his genuine care for all his students including myself.

I am blessed for having known two other professors: Dr. Elizabeth Zsiga and Dr. Mahendran Velauthapillai. Together, they were my mental and emotional anchors at Georgetown. I truly appreciate Dr. Zsiga for her constant encouragement, especially when I had lost confidence in myself. She inspired me to maintain my life at Georgetown and not give up. Likewise, I am also thankful Dr. Mahendran Velauthapillai for his generosity and willingness to share in not only his profound academic knowledge but also the greater lessons of life.

I am very grateful to the entire Department of Linguistics at Georgetown for all their support, undoubtedly essential in helping me to grow as a researcher until the present. I would like to specially thank our department administrative staff, Erin Esch, for her kind, prompt, and professional help. My time at Georgetown was truly a great and wondrous experience and was also further enriched by the presence of many fellow graduates and friends in the department. My heartfelt thanks in particular go to Dr.

Rebecca Sach for her wonderful friendship, camaraderie, entertainment, emotional support, and most of all, for her enthusiasm in my work.

I would like to gratefully acknowledge revered former professors, colleagues, and friends in Korea for helping me collect the necessary data for this study. I would like to extend my sincere thanks to my formal professors, Dr. Oryang Kwon and Dr. Jin-Wan Kim for their cordial advice and continued support. I also would like to thank my formal colleagues, Dr. Kyungja Ahn, Dr. Da-Yun Nam, and Dr. Young-Soon So for their immediate help and infinite support. Data collection would not have been successful or even possible without their assistance. I also would like to thank Haein Park, Sung-Kong Park, Nick Yon, Joo Yoon Chung, and Jeong-Hoon Yang for their kindness in helping my study go along smoothly.

Lastly and most importantly, I am extremely grateful to my family for their undying love, constant encouragement, and unwavering support throughout this long journey. I would like to thank my parents, Soo Woong Eom and Jin Soon Lee, for their never-ending care and support. Words again cannot suffice for their infinite love; my brother, Soo-Chang Eom, who has always rooted for me and helped to shape me into a better version of myself – he has always been my best counselor. My sister-in-law, Han-Kyoung Kim, who always sends me happiness with a bright smile; my precious little nieces, Chae-Youn and Ye-Dam, who have been patient and understanding enough to realize the importance of their aunt's work despite the limited amount of time spent with them, and last of all my little buddy Pansy, who has kept me company on many nights. Thank you for believing in me and let it be known, I love all of you from the bottom of my heart.

TABLE OF CONTENTS

Chapter I. Introduction.....	1
Chapter II. Previous Research.....	14
2.1 Vocabulary learning and reading comprehension.....	14
2.1.1 Vocabulary learning through reading	14
2.1.2 Dictionary use	16
2.1.3 Assessment of vocabulary learning	19
2.2 Intelligent Computer Assisted Language Learning (ICALL).....	21
2.2.1 ICALL from grammatical to semantic processing.....	22
2.2.2 ICALL for vocabulary learning and reading comprehension	24
2.3 Summary and conclusion.....	30
Chapter III. The System.....	31
3.1 System manager.....	31
3.2 Natural language processing server	37
3.2.1 Linguistic annotation	38
3.2.1.1 Tokenization	38
3.2.1.2 Part-of-Speech tagging.....	41
3.2.1.3 Lemmatization	42
3.2.1.4 Collocation identification.....	44
3.2.2 Word sense disambiguation (WSD).....	46
3.2.2.1 SenseLearner 2.0.....	48
3.2.2.2 WordNet::SenseRelate::AllWords	49

3.2.2.3 Naïve Bayes	50
3.3 Lexical database.....	51
Chapter IV. The Lexical Database.....	53
4.1 Introduction.....	53
4.2 Related works.....	55
4.2.1 Early works of word sense alignment.....	58
4.2.2 Recent works of word sense alignment	62
4.2.3 Word sense alignment evaluation	66
4.2.4 Comparison with previous research.....	68
4.3 Sense inventories	69
4.3.1 WordNet.....	69
4.3.2 English language learners' dictionary.....	70
4.4 Word sense alignment (WSA)	70
4.4.1 Application-specific assumption.....	73
4.4.2 Initial alignment.....	74
4.4.3 The heuristic of the study: adding flatness	79
4.4.4 Calculating probability of alignment structure	84
4.4.5 Adjusting WSD output.....	89
4.4.6 Sense alignment algorithm.....	91
4.4.6.1 Basic WSA algorithm	91
4.4.6.2 Complexity.....	93
4.5 Evaluation	94
4.5.1 Obtaining evaluation data	95

4.5.1.1 Pooling semi-experts.....	95
4.5.1.2 Word selection	96
4.5.1.3 Survey design.....	97
4.5.2 Evaluation data.....	99
4.5.2.1 Overview.....	99
4.5.2.2 Evaluation metrics	103
4.5.3 Evaluating WSA system	105
4.5.3.1 Counting flatness of alignment structure	105
4.5.3.2 Different weight on related meaning	112
4.5.3.3 Evaluation of the system.....	118
4.6 Summary	123
Chapter V. The Empirical Study: Evaluation of the whole system	125
5.1 Research questions.....	125
5.2 Method	126
5.2.1 Participants.....	126
5.2.2 Materials	132
5.2.2.1 Reading texts.....	132
5.2.2.2 Target words	134
5.2.2.3 Reading comprehension tests.....	135
5.2.2.4 Vocabulary tests.....	136
5.2.2.5 User Database	140
5.2.3 Procedure	140
5.2.4 Scoring.....	144

5.2.4.1 Reading comprehension test	144
5.2.4.2 Vocabulary test	145
5.2.5 Data analysis	146
5.2.5.1 Vocabulary acquisition	147
5.2.5.2 Reading comprehension.....	149
5.2.5.3 Effects of the system errors.....	150
5.3 Results and discussion	152
5.3.1 Vocabulary acquisition	152
5.3.2 Reading comprehension.....	172
5.3.3 Effects of the system errors.....	178
5.3.4 Implications.....	180
Chapter VI. Summary and Outlook	182
6.1 The system from a computational perspective.....	182
6.2 The system from a language learning perspective.....	186
APPENDICES	191
Appendix A: Code of computing a probability of alignment structure.....	191
Appendix B: Code of computing word sense alignment	193
Appendix C: Definitions and examples of nine words from WordNet and COBUILD	198
Appendix D: Scores for nine words.....	206
Appendix E: Results of the initial alignment and adjusted alignment over nine words ($\alpha=0.5$, gold standard = all positive, based on SR::AW)	209

Appendix F: Results of the initial alignment and adjusted alignment over nine words ($\alpha=0.5$, gold standard = all positive & top positive, based on SL & NB).....	210
Appendix G: Precision/Recall of WSA based on SL and NB	214
Appendix H: Two reading texts for the empirical study.....	216
Appendix I: Reading comprehension tests	218
Appendix J: Pretest	222
Appendix K: Vocabulary posttest-3 (Post-3).....	226
Appendix L: Vocabulary posttests-1, -2, and -4	230
Appendix M: Informed consent documents.....	237
REFERENCES	246

LIST OF FIGURES

Figure 1.1.	The system architecture	10
Figure 3.1.	Web page to choose a reading text	32
Figure 3.2.	Page source of the example web page (Figure 3.1)	33
Figure 3.3.	Presentation of information by the learner's clicking	34
Figure 3.4.	A part of an analyzed text	35
Figure 3.5.	brdg.py	36
Figure 4.1.	A part of the Index	54
Figure 4.2.	A part of the html repository	54
Figure 4.3.	The lexical database	55
Figure 4.4.	Three kinds of mapping (plant.n)	56
Figure 4.5.	$p(w_n, cb)$ for a given alignment	75
Figure 4.6.	$p(w_n, cb)$ and $p(A_s)$ for a given alignment	83
Figure 4.7.	All possible alignment structure types given five w_n senses and three cb senses and one example for each type	85
Figure 4.8.	The same alignment structure type A_{S4}	86
Figure 4.9.	WSA based on WSD system output (left) and WSA when the WSA algorithm applied (right) for alignments for require.v (dashed line = revised link)	89
Figure 4.10.	Gold standard (left), initial alignment based on WSD system output (middle), and adjusted WSA applied with the flatness (right) for alignments for community.n	90

Figure 4.11.	Question and Choices	98
Figure 4.12.	Confidence Scale	99
Figure 4.13.	Number of times each answer was used for every word	101
Figure 4.14.	Precision & recall average over nine words processed by WSA (AP).....	117
Figure 5.1.	A screenshot showing lexical information, as presented to the GS group	129
Figure 5.2.	A screenshot showing lexical information, as presented to the SS group.....	130
Figure 5.3.	A screenshot showing lexical information, as presented to the AS group	130
Figure 5.4.	A screenshot showing no lexical information, as presented to the NS group	131
Figure 5.5.	Screenshot of the main menu page.....	132
Figure 5.6.	Procedure of the study	143
Figure 5.7.	Scores on the pretest and the post-3 (in percentage)	154
Figure 5.8.	Pre-post-3 gain in percentage	155
Figure 5.9.	Pre-post gain scores for H and L across the four groups.....	160
Figure 5.10.	All words clicked vs. all target words clicked.....	162
Figure 5.11.	Target words clicked for H and L across three groups.....	164
Figure 5.12.	Percentage correct of the clicked target words over the tests.....	166
Figure 5.13.	Pre-Post3 gain among target words clicked (percentage)	167
Figure 5.14.	Reading comprehension scores of the four groups (percentage).....	172

Figure 5.15. Total RC mean scores for the High/Low groups across the
four groups 175

LIST OF TABLES

Table 4.1.	Comparisons between initial alignment and human mapping, showing the COBUILD (c) senses which have more WordNet (w) links than the average	80
Table 4.2.	Comparisons of flatness between initial alignment and human alignment.....	82
Table 4.3.	The various values used in calculating probabilities for alignment types	87
Table 4.4.	Words selected for the present experiment, including number of senses in each inventory.....	97
Table 4.5.	Response Analysis for one WordNet sense of section.n	100
Table 4.6.	Average number of responses for each point on the confidence scale (1=not confident, 5 = very confident).....	101
Table 4.7.	Scores for involve.v ($\alpha = 0.5$)	104
Table 4.8.	Results of the initial alignment and adjusted alignment over nine words ($\alpha=0.5$, gold standard = top positive; improvements in bold).....	106
Table 4.9.	Results of initial alignment and adjusted alignment over three WSD systems (SL, SR::AW, NB) ($\alpha=0.5$, gold standard = all positive; improvements in bold).....	110
Table 4.10.	Results of initial alignment and adjusted alignment over three WSD systems (SL, SR::AW, NB) ($\alpha=0.5$, gold standard = top positive; improvements in bold).....	110

Table 4.11.	Number of link changes from the initial alignments to the adjusted alignments.....	111
Table 4.12.	Precision & recall of words from the initial alignment based on WSD (SR::AW) outputs (AP=all positives, TP=top positive), plus the total number of un-aligned senses	113
Table 4.13.	Precision & recall of words from the adjusted alignment by counting $p(A_s)$ (AP=all positives, TP=top positive), plus the total number of un-aligned senses.....	114
Table 4.14.	Precision & recall average over nine words processed by the WSA algorithm based on three different WSD systems (SR::AW; SL; NB) (AP=all positives, TP=top positive).....	116
Table 4.15.	Results of WSA on 20 words.....	119
Table 4.16.	Average sense numbers of nine words and 20 words	122
Table 5.1.	20 target words.....	135
Table 5.2.	All cases of participants' performance in the SS group.....	152
Table 5.3.	Test of Homogeneity on pretest/Post-3 for the four groups	153
Table 5.4.	Descriptive statistics of pretest and posttest (Post-3) scores across the four groups	153
Table 5.5.	Results of RM ANOVA comparing vocabulary test scores across the four groups over time	155
Table 5.6.	Mean difference between pre-post3 for each group	157
Table 5.7.	Contrast Results for the amount of the mean pre-post gains	158
Table 5.8.	Test of Homogeneity of variances on pretest/Post-3 for all groups	159

Table 5.9.	Descriptive statistics of pretest and posttest (Post-3) scores across the eight groups.....	159
Table 5.10.	Results of two-way RM ANOVA comparing vocabulary test scores across the four groups and the H/L groups over time	161
Table 5.11.	Levene’s Test of Homogeneity of Error Variances	165
Table 5.12.	Descriptive statistics for vocabulary acquisition for clicked words (percentage correct)	166
Table 5.13.	Descriptive statistics of Post-1, Post-2, and Post-4 scores across the four groups	168
Table 5.14.	Levene’s Test of Homogeneity of Variances	170
Table 5.15.	Results of ANOVA analysis	170
Table 5.16.	Tukey post-hoc comparisons for Group differences.....	171
Table 5.17.	Descriptive statistics for reading comprehension mean scores across the four groups	172
Table 5.18.	Levene’s Test of Equality of Error Variances	173
Table 5.19.	Results of one-way ANOVA for reading comprehension scores of the four groups	173
Table 5.20.	Tukey post-hoc comparisons for Group differences.....	174
Table 5.21.	Descriptive statistics for reading comprehension mean scores for the High (H) and Low (L) groups across the four groups.....	175
Table 5.22.	Results of two-way ANOVA for reading comprehension scores for the four groups and HL groups	176
Table 5.23.	Tukey post-hoc comparisons for Group differences.....	177

Table 5.24. Pre to Post-3 gain on clicked words for learners in the SS group 179

I. INTRODUCTION

Natural language processing (NLP) technology is rapidly evolving to be utilized in an increasing number of modern day applications, such as but not limited to machine translation, information retrieval, speech recognition, and dialog systems (Jurafsky and Martin, 2000). One interesting application is to apply various NLP techniques in an educational setting to help language learning, which is known as Intelligent Computer-Assisted Language Learning (ICALL). Research on developing ICALL systems has often focused on supporting learners by encouraging them to practice grammatical forms and functions with individualized feedback (e.g., Heift's E-Tutor, 2001; Nagata's BANZAI, 2002; Amaral and Meurers' TAGARELA, 2006). However, research on ICALL systems specific to the handling of semantic processing (e.g., Bailey and Meurers, 2008) still remains largely unexplored.

As NLP has progressed to the point of enabling the processing of many semantic properties in natural language, ICALL research should be able to aid language learning; this involves understanding the meaning of words and sentences by utilizing more advanced NLP techniques (e.g., word sense disambiguation). As one way of pursuing this objective, this research aims to explore how NLP techniques can be applied in a learning environment which specifically involves the meanings of words, by supporting vocabulary learning and reading comprehension for second language learners.

The ability to retain vocabulary is a fundamental tool for the utilization of any language. Lack of adequate vocabulary and their associated meanings can seriously hinder learners in their efforts at learning a language. Of all other major components in second language proficiency, sufficient vocabulary is viewed as the most crucial for achieving proper reading comprehension (Chanier and Selva, 1998; Coady, 1997; Grabe, 1991; Groot, 2000; Hirsch, 2012; Laufer, 1997; Milton, 2009; Tozcu and Coady, 2004). This explains why a myriad of comprehension problems in second language reading involve the problem of interpreting words that are unknown or used in unfamiliar ways. This is certainly exacerbated by the prevalence of lexical ambiguity. Landes et al. (1998) report that more than half of the content words in English texts are lexically ambiguous, with the most frequent words having a large variety of meanings. Consider example (1) of the meaning contained in the word *deliberately* and example (2) for *cradle*.

- (1) Psychologists have done experiments that involve asking people to yawn *deliberately* in crowded rooms and auditoriums.
- (2) The plan encourages everyone to consider the impact of a fashion item from the *cradle* to the grave.

The word *deliberately* is one that is frequently encountered by learners who may already recognize its most prevalent sense, *carefully*. However, learners may become confused or fail to grasp the intended meaning (i.e., *intentionally*) in context, and this may lead them to completely misunderstand the sentence - i.e., *there are experiments involving careful yawning in crowded places*, rather than *there are experiments involving intentional yawning in crowded places* - where the latter require acquiring the meaning of

the word, *intentionally*. Example (2) also shows a context in which learners may be confused by the word *cradle* used in a way that is related to a more frequent meaning (at least metaphorically) but is clearly distinct. Learners who understand the meaning of the word *cradle* to be defined as *a baby's bed with high sides* would comprehend the sentence as *People are encouraged to consider the impact of a fashion item for all ages from the item for the babies to the one for the dead*, instead of *People are encouraged to consider the impact of a fashion item from the beginning (its design) to the end (its disposal)*.

The confusion witnessed at this fairly elementary level takes a turn for the worse when a word carries an even greater variety of meanings. The word, *face*, for example, is listed in WordNet (Fellbaum, 1998) with twelve different nominal senses (e.g., human face, facial expression, side, surface, boldness, etc.); although not all are equally prevalent, there is still much potential for confusion. The diverse meaning of a word indeed makes it difficult for learners to determine which meaning is the most appropriate when reading a text and learning a word in the context of the reading.

Therefore, research is needed to overcome the problems caused by lexical ambiguity, i.e., multiple senses for a word. Of the many approaches to this question, this dissertation asks the specific question: is it be helpful to provide sense-specific lexical information in learners' reading, in order to improve reading comprehension and vocabulary learning? By *sense-specific*, the study refers to information applicable only for one given sense (meaning) of a word. There are systems which automatically provide sense-specific lexical information (e.g., the REAP tutor (Heilman et al., 2006, Kulkarni et al., 2008), GLOSSER-RuG (Nerbonne and Smit, 1996)) in an attempt to alleviate a learner's

confusion with multiple meanings of an unknown word and avoid further misunderstanding of a reading. They use NLP technology (e.g., word sense disambiguation) to make it feasible, but their approaches have encountered several problems that require treatment. For example, GLOSSER-RuG (Nerbonne and Smit, 1996) employs a part of speech (POS) disambiguation system in order to provide definitions of the words. The limitation of its POS disambiguation lies in the fact that it does not disambiguate all the senses of a word in the same POS; thus, if there are several senses of the same POS for the word, it simply presents all of them, affecting only a small difference from systems without any word sense disambiguation (WSD). This is the one defining element that differentiates the system built for this study in that the WSD module classifies all senses of words thoroughly, i.e., classifies the senses of the word even within the same POS.

As a more effective system for sense understanding, the REAP tutor contains high performance via advanced WSD. Although the WSD system of the REAP tutor shows high accuracy, it is based on an annotated training data, which means their WSD approach only handles words in the annotated data. Accordingly, it still may restrict learners from the opportunity of learning any word in a text (e.g., words not in the annotated data) and may not be very effective for assisting learners' reading comprehension. Also they built their own annotated data to handle as many words as possible, but constructing annotated data itself is quite expensive in time and labor. The WSD approaches of the REAP tutor and the present study are fairly different; the WSD approach taken in this study is not based on annotated data and thus provides sense-

specific information for any content word in a text, which may be more helpful in reading comprehension and learning more words from a reading.

Likely of higher significance is that the quality of lexical information presented to language learners may assist or hamper learning. First, the provided lexical information should contain both definitions and examples of the word. Since the definition displays the meaning of the word *explicitly* and examples show the meaning *implicitly* (Segler, 2007), presenting both should help language learners by providing more illustrations of the meaning. Second, the more examples the learner sees of the words in the context, the more clearly vocabulary acquisition is obtained. Rapaport and Kibby (2002) show that through encountering several examples in contexts, learners can acquire a word's meaning by continually creating and revising a "hypothesis" about its meaning. This is possibly due to the fact that example sentences in various contexts may illustrate a word's meaning more fully (Black, 1991), which leads to successful "semantization" of the word (Beheydt, 1987). Yet, the previously-developed systems (the REAP tutor (Heilman et al., 2006, Kulkarni et al., 2008), GLOSSER-RuG (Nerbonne and Smit, 1996)) do not fully account for the significance of the quality of lexical information presented to language learners; GLOSSER-RuG appears to note the necessity of both definitions and examples to assist learners whereas the REAP tutor seems to overlook the role of examples in acquiring the meaning of words. The REAP tutor offers lexical information from a conventional dictionary, thereby focusing more on providing definitions.

In particular, the quality of examples is of key importance. Examples should make use of simple vocabulary familiar to the learners so as to be easily understood. If the structure or vocabulary of examples is overly and unnecessarily complex, they will be

inappropriate for aiding learners (Kilgarriff et al., 2008; Segler, 2002). In view of this, example sentences taken directly from corpora or web pages appear to be less appropriate as the information presented by them may be less accessible to language learners (Groot, 2000; Kilgarriff et al., 2008; Segler et al., 2002). As seen in (3), the example for use of the word *face* extracted from the web may confuse learners in understanding and learning the word *face*; the example is complicated in structure (e.g., containing a subordinate clause) and also it has a few words that learners may find unfamiliar (e.g., *plenty*, *inappropriately*, *assure*, etc.). Worse still, the use of *deliberately* in (4) is more complicated in structure (e.g., containing a relative clause with missing relative pronouns) and words (e.g., *thwart*, *investigation*, *hacking*, etc.).

(3) You will see plenty of people dressed inappropriately, but be assured that they have lost *face* in the eyes of the Thais around them. (*East Asia Travel*, Jan 15, 2008)

(4) It finds the company *deliberately* tried to thwart the 2005-2006 Metropolitan police investigation into phone hacking carried out by the News of the World.

(*The Guardian*, July 19, 2011)

In terms of the quality, the examples provided in GLOSSER-RuG are not quite helpful. It provided example sentences as a means of helping learners understand the meaning of a word in a reading. But it extracted examples sentences from corpora which are not controlled in terms of structural and lexical complexity. So, learners may further need to use a dictionary to look up unknown words in example sentences. This is not very effective for language learners as it may distract more from the focus on their reading and consequently disrupt their stream of reading comprehension (Koyama and Takeuchi,

2004; Laufer and Hill, 2000; Leffa, 1992; Luppescu and Day, 1993; Prichard, 2008). Indeed, such example sentences from corpora or web sources do not cater to the needs of language learners, although they may, in fact, represent truly authentic use of such words. By comparison, however, examples made up by lexicographers based on their intuitions¹ for learner dictionaries typically control syntactic and lexical complexity (Segler et al., 2002). The REAP tutor uses this form of examples; they extract definitions and examples from a standard dictionary built by lexicographers. However, those examples entirely made up by lexicographers are likely to lose the authenticity and naturalness in the actual usage of words.

Kilgarriff et al. (2008) claim that it may be too difficult to consistently find sentences which satisfy all required criteria of a good example in corpora. So “editorial intervention” is needed to some extent, such as to delete an irrelevant clause, simplify complex names, etc. If so, examples taken from corpora and modified in some way to simplify lexical and grammatical complexities would be ideal (Segler, 2002). Such examples can be authentic, display appropriately the target meaning of the word, and be grammatically simplified without losing any of their illustrative value. In sum, it is seen as beneficial to use a resource in which examples are extracted from corpora, in order to ensure authenticity, and modified by lexicographers, in order to control lexical and structural complexity. In the present study, therefore, the Collins COBUILD Student’s Dictionary (Sinclair, 2006) is selected to extract these types of good examples. The content in COBUILD is based on actual English usage derived from the analysis of a large corpus of written and spoken English, thereby providing authentic examples while

¹ Lexicographers are educated native speakers of the language so their intuition about the typical usage is precise (Laufer, 1992).

retaining control of lexical and structural complexity to some extent for the benefit of language learners (Sinclair, 2006).

One can easily spot the difference between the COBUILD examples in (5) and (6) from those in (3) and (4); the COBUILD example in (6) is structurally and lexically simple and provides additional context (*calmly*) to understand the meaning of *deliberately*.

(5) To cancel the airport project would mean a loss of *face* for the present governor.

(6) The Japanese have acted calmly and *deliberately*.

Yet, using COBUILD examples (and definitions) for vocabulary assistance brings a practical problem in terms of developing the actual system in this study. In order to provide sense-specific lexical information for the words, a state-of-the-art WSD system is employed (Chapter 3): the system is based on the WordNet sense inventory. Unfortunately, the sense inventories used for disambiguating the sense of the words (i.e., WordNet) and for displaying sense-appropriate lexical information to learners (COBUILD) do not match. The fact that every inventory (i.e., dictionary) has its own standards of sense distinction poses a serious problem in integrating these inventories into one. Herein lies the challenge of this research; the standards of sense distinction are different between WordNet and COBUILD, which requires the current study to explore a method to map word senses between the two. That is, in order to consistently present relevant lexical information, the system must link up senses between an automatic WSD

system employing WordNet and a sense inventory containing lexical information more appropriate for language learners, namely the COBUILD dictionary (Sinclair, 2006).

In sum, although vocabulary knowledge is critical for language learning, learning vocabulary and understanding texts present difficulty for language learners due to the high lexical ambiguity frequently found among different words. Some studies try to overcome these problems by employing NLP technology, but they are still limited in their ability to offer an effective setting for supporting learners' vocabulary acquisition and reading comprehension. In addition, different resources used for disambiguating the sense of a word and for presenting lexical information of the word to learners pose a practical challenge in linking these resources by the corresponding sense of the word. The discussion to this point thus suggests a need for research to help resolve and improve on these limitations by automatically providing enriched sense-specific information to language learners as a means of supporting their vocabulary acquisition and reading comprehension.

In this study, an online system is developed to provide vocabulary assistance to learners of English as a second language (ESL) for difficult words during reading and evaluated to test the expectation that this system may offer an improved path for learners to improved vocabulary acquisition and reading comprehension. The system performs this role by allowing learners to click on unfamiliar/unknown words and see lexical information (i.e., target word definitions and examples) relevant to the particular usage given the context of reading. The system aims to assist with any content word.

As a brief overview, the overall system consists of three major components: a system manager, an NLP server, and a lexical database. These are closely interconnected to each other in order to optimize execution of the related tasks. As illustrated in Figure 1.1, the system manager controls the interaction among each learner, the NLP server, and the lexical database.

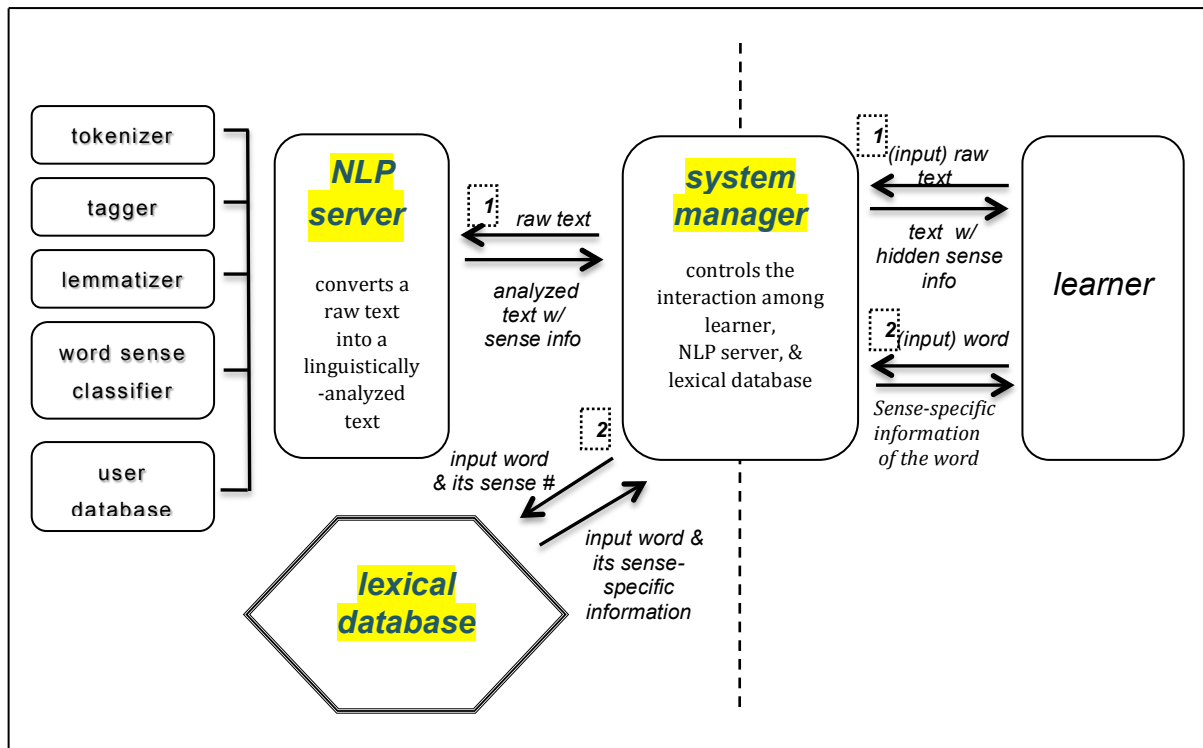


Figure 1.1. The system architecture

When the system manager receives a raw text (passage) as an input from the learner, it first sends the input text to the server, which then returns an analyzed text (i.e. tokenized, POS-tagged, and sense-tagged) back to the learner, along with content words that made to be clickable. Then, when the learner clicks on a word while reading, the system manager sends the word with its sense information (i.e. WordNet sense) to the lexical database and brings the word with its sense-specific lexical information (i.e.

COBUILD definition and examples) back to the learner from the lexical database. One of the critical functions of the system is to ensure that the lexical database redirects WordNet senses for each word to the appropriate COBUILD information. More details on each component are provided in Chapters 3 and 4.

To develop this system, advanced technology for word sense disambiguation (WSD) is utilized to enable the system to identify the meaning of a word chosen by a learner in a text during reading. A lexical database is also developed to present learners with sense-appropriate lexical information extracted from the COBUILD dictionary (Sinclair, 2006). The system is able to provide enriched sense-specific lexical information for any content word in any text, achieved by automatically mapping WSD system output (based on WordNet senses (Fellbaum, 1998)) to corresponding sense-appropriate lexical information in COBUILD. This automatic mapping between WordNet senses and COBUILD senses is performed based on a word sense alignment (WSA) algorithm newly developed in this research. To examine the WSA algorithm, an evaluation data set is constructed by pooling human judgments using a method of *crowdsourcing* (i.e., an online survey). The online system developed in this research is finally tested on groups of students learning English as a second language (ESL).

With the hypothesis that automatic provision of enriched sense-specific information of a word can facilitate learners achieving more successful vocabulary acquisition and reading comprehension, the overarching goals of this research are thus posed as follows:

1. Build a system, utilizing techniques of natural language processing in order to identify the sense of any content word in a text (Chapter 3)

2. Build a lexical database based on a method of word sense alignment developed in this study in order to support enriched sense-specific lexical information for language learners. As an offshoot, a data set is also developed to evaluate word sense alignments generated by the system (Chapter 4)

3. Examine the validity/reliability of the system for language learners' vocabulary acquisition and reading comprehension in a real educational setting; for this empirical purpose, further research questions are as follows (Chapter 5):
 - (a) 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?

 - (b) Does sense-specific lexical information facilitate learners' reading comprehension?

The main contribution of this last point is to investigate whether high quality sense-specific lexical information presented in an intelligent system helps learners in their vocabulary acquisition and reading comprehension. In a broader context, the system developed in this study contributes to an extended understanding of how NLP techniques can be applied more effectively in an educational environment.

This dissertation is organized as follows: Chapter 2 presents a theoretical background of this research and reviews the capacities and limitations of related research. Before explaining how to build the central component of the system (the lexical database),

Chapter 3 describes the technical functions of the system: the system manager and NLP modules such as the tokenizer, POS tagger, lemmatizer and word sense classifiers. Then, the primary component in the system, the lexical database, is explained in detail in Chapter 4: it presents a WSA algorithm and discusses how to build the lexical database based on the algorithm. Also, the evaluation of such an algorithm is treated as a topic in its own right. To demonstrate the validity and reliability of the system in an actual educational setting, Chapter 5 describes how to conduct an empirical study of evaluating the system which provides sense-specific lexical information to learners of English: this chapter addresses the methodology and discusses the results. Lastly, Chapter 6 briefly summarizes the present study and its findings, discussing its implications and suggestions for future research.

II. PREVIOUS RESEARCH

This chapter addresses previous research regarding vocabulary acquisition and reading comprehension related to the overall goal of the present study (2.1). It also reviews past work on Intelligent Computer-Assisted Language Learning (ICALL) that is relevant to the general framework of the present study's overall goal (2.2). Previous research on word sense alignment, a specific topic within a lexical database approach, is discussed separately in its own right in Chapter 4.

2.1 Vocabulary learning and reading comprehension

This section discusses previous research regarding second language vocabulary acquisition through reading (2.1.1) and reviews studies that have investigated the effect of dictionary use in vocabulary acquisition and reading comprehension (2.1.2). This section also looks into previous work on vocabulary assessment that is related to the vocabulary test design used in the empirical study (2.1.3)

2.1.1 Vocabulary learning through reading

Successful reading comprehension is crucially dependent on knowledge of vocabulary (Haynes and Baker, 1993; Coady et al., 1993; Laufer, 1997; Nation and Coady, 1988). Laufer (1997) specifically claimed that a vocabulary of at least 5,000 words is necessary for learners to comprehend any text successfully. Therefore, it is not feasible for learners to attain all necessary vocabulary (e.g., 5,000 words) by means of formal instruction alone. Rather, it would make sense that learners acquire more vocabulary through

reading, because reading allows for a greater chance of being exposed to more words. This is supported by previous studies that claim that ‘incidental vocabulary learning’ or learning vocabulary through reading, is more effective for second language vocabulary acquisition (Huckin and Coady, 1999; Joe, 1998; Krashen, 1989; Matsuoka and Hirsh, 2010; Nagy, 1997; Nagy et al., 1985; Nation and Coady, 1988; Paribakht and Wesch, 1997, 1999; Schouten-Van Parreren, 1989; Wode, 1999). Vocabulary learning by means of reading is regarded as not only incidental but also fundamental (Huckin and Coady, 1999). As such, a reading provides an ideal environment for vocabulary acquisition.

In addition, reading with vocabulary activities (*reading plus*) leads to better acquisition of vocabulary than does reading without vocabulary activities (*reading only*) (Laufer, 2001, 2003, 2005; Paribakht and Wesch, 1997, 1999). Moreover, vocabulary acquisition is not achieved if learners are exposed to an unknown word only once (Rott, 1999; Rapaport and Kibby, 2002). In other words, low-incident words in reading are less likely to be learned. Peter et al. (2007, 2009) have suggested a “vocabulary test announcement” as one way to enhance the learning of words in a reading text, particularly the low-frequency words. In their study, they informed their learners that vocabulary tests would follow the reading task. They found that this announcement, which occurred before the reading task, made learners more focused on words presented, which in turn resulted in a substantial increase of their vocabulary learning. Thus, Peter et al. (2007, 2009) regarded vocabulary acquisition as intentional, as demonstrated by their methodology. Based on these previous studies, it is worthwhile to note that vocabulary learning through reading is more effective than through formal training alone. Furthermore, “vocabulary test announcement” enhances vocabulary acquisition from

reading. This dissertation benefits from these results in terms of the design of the empirical study (as discussed in Chapter 5), as it sets vocabulary learning in the context of reading and administers vocabulary posttests with a test announcement before the reading task.

2.1.2 Dictionary use

The previous section addressed the idea that learners can acquire vocabulary more effectively through reading. If this is the case, when they encounter unknown or unfamiliar words while reading, by what means do they learn the meaning of those words? Some previous studies have suggested that guessing the meaning of words in context could yield actual learning of those words (Fraser, 1999; Li, 1988; Paribakht and Wesch, 1999; Sternberg, 1987). However, for a second language learner, guessing the meaning of words in context is less successful than using a dictionary to achieve acquisition of those words (Bensoussan and Laufer, 1984; Haynes, 1993). Use of a dictionary is a more effective way of learning word meanings, as shown in a number of studies focusing on the value of dictionary use in vocabulary learning and reading comprehension (Bogaards, 1998; Koyama and Takeuchi, 2004; Knight, 1994; Luppescu & Day, 1993; Prichard, 2008; Summers, 1988).

Most of these studies (Bogaards, 1998; Knight, 1994; Luppescu & Day, 1993; Prichard, 2008; Summers, 1988) empirically demonstrated that using dictionaries not only improved learners' reading comprehension, but also facilitated increased vocabulary acquisition. In Luppescu & Day's (1993) study, Japanese speakers learning English performed better on reading and vocabulary posttests when they used dictionaries while

reading. Likewise, Knight (1994) showed that learners of Spanish performed better on reading comprehension and vocabulary posttests and demonstrated a longer retention of the learned words when they used dictionaries. In the same vein, Summers' (1988) study also showed that groups using dictionaries outperformed those who did not in reading comprehension and vocabulary tests. Bogaards (1998) also noted that dictionary use was still more helpful than contextual guessing in finding appropriate word meaning in context. More specifically, Koyama and Takeuchi (2004) showed that using an online dictionary distracted learners less, which resulted in a shorter amount of time needed for reading.

However, some studies have argued that dictionary use during a reading task can induce ineffective vocabulary learning and reading comprehension (Bogaards, 1998; Knight, 1994; Luppescu & Day, 1993). For example, Luppescu & Day (1993) observed that the group using dictionaries while reading took twice as long to complete the reading task than did the group not using them. It takes more time for learners to determine the proper sense of a word with several sense entries, as learners are required to keep the original text in mind for comparing the unknown word with the several possible senses listed in the dictionary. This in turn results in a slower reading pace. Bogaards (1998) also pointed out that dictionary use could lead to ineffective vocabulary learning and reading comprehension because learners often failed to select the contextually correct meaning from the dictionary's multiple sense entries. As a whole, these researchers (Bogaards, 1998; Knight, 1994; Luppescu & Day, 1993) found that using a dictionary while reading often interrupted learners from focusing completely on the text, even if it resulted in a better performance on reading and vocabulary posttests.

It has been suggested that utilizing electronic dictionaries can be one way to alleviate the problems of longer reading time and distraction from reading (Ellis, 1995; Prichard, 2008). Leffa (1992) compared the use of a computer-mediated electronic glossary to that of a traditional paper dictionary for reading comprehension. Leffa (1992) found that learners who used electronic glossaries showed an increased efficiency in reading comprehension when compared to learners who used paper-based dictionaries. Leffa (1992) also found that while electronic dictionary users understood 86% of the text in 50% less time, the traditional paper-based dictionary users understood only 62% of the text. In a similar vein, Lyman-Hager et al. (1993) demonstrated the superiority of electronic dictionaries over paper-based ones by showing that learners using online dictionaries scored significantly better on a vocabulary quiz than did learners using paper dictionaries. Indeed, online dictionary use does not disturb learners' reading comprehension process, but rather enables them to switch attention between a dictionary and a text without requiring much working-memory (Ellis, 1995).

However, the problem of confusion with numerous sense entries (Bogaards, 1998; Luppescu & Day, 1993; Tang, 1997) has not been solved. For example, Luppescu & Day (1993) found that dictionary use caused confusion for learners when a word had multiple sense entries. This confusion often led to improper reading comprehension and ineffective vocabulary learning (Bogaards, 1998).

As discussed above, there are clearly advantages to using dictionaries for reading comprehension and vocabulary acquisition. However, there are still some limitations of dictionary use that have not been overcome. These problems raised in second language

research can be resolved by utilizing advanced natural language processing technologies, as discussed in section 2.2.

2.1.3. Assessment of vocabulary learning

As stated previously, vocabulary knowledge plays an essential role in language learning. It is accordingly important to determine appropriate measures for assessing a learner's vocabulary. In second language learning, there are generally two measurements for assessing knowledge of vocabulary: how many words learners know and how well learners know those words (Nation, 2001; Read, 1998, 2000; Qian, 1999). Of the two, as Read (1993) addressed, simply measuring learners' vocabulary size is not meaningful for testing how well learners know the particular words; especially when words have multiple meanings, a test of approximate vocabulary size is not sufficient. As such, Read (1997) suggested two approaches to assess effectively how much learners know about the words. The first approach is a multi-dimensional measure of word knowledge. As discussed in several previous works (Coombe and Hubley, 2003; Nation, 2001; Schmitt, 2000; Wesch and Paribakht, 1996), this approach tests learners' vocabulary knowledge by analyzing various aspects of words, such as meaning, spelling, pronunciation, morphological and syntactic properties, and so on.

While the first approach is not appropriate for the present study, the second approach seems to be beneficial in terms of vocabulary test design. This second approach tests learners' vocabulary knowledge at each stage of learning. In other words, how well learners know a word is assessed at each of the developmental stages of vocabulary knowledge.

In terms of eliciting students' perceived knowledge, this approach is similar to the Vocabulary Knowledge Scale (VKS) developed by Wesch and Paribakht (1996). The VKS measures learners' level of vocabulary knowledge using a five-point scale, in which 1 is 'not familiar at all', 2 is 'familiar but meaning is not known', 3 is 'synonym or translation is given', 4 is 'semantic appropriateness in a sentence', and 5 is 'semantic and grammatical appropriateness in a sentence'. Although the VKS is a self-report assessment in that learners measure their own vocabulary knowledge, the procedure was intended to assess learners' incremental vocabulary knowledge gains, thereby demonstrating their knowledge of target words (Wesch and Paribakht, 1996). Since it is helpful to track learners' vocabulary learning (Kim, 2008; Read 2007; Wesch and Paribakht, 1996), the idea of the VKS aids in vocabulary test design in the sense that tests are formed to examine degrees of learners' knowledge of a target word. For example, the study in this dissertation has four vocabulary tests: one test is related to the VKS #2, another is related to the VKS #3, and the remaining two are related to the VKS #4. Although it assesses various degrees of learners' vocabulary knowledge, the present study's main interest is to examine if they learn the meaning of words in context, as is related to the VKS #4.

This in turn raises the issue regarding which test format should be used. The test format is important for vocabulary testing (Coombe, 2011) and thus should be carefully determined. Coombe (2011) suggested three points to consider when designing the format of vocabulary tests: the format should 1) be familiar to the learners, 2) be easy to grade, and 3) have "positive backwash effect," which means that a test provides learning effects similar to those from teaching (Hughes, 1989; Nation, 2001). The four vocabulary tests of the present study do have a "positive backwash effect"; by taking four tests about

the same target words, learners receive the effect of repetition of these words, thereby yielding their acquisition. In terms of the remaining two suggestions, familiarity for learners and ease of grading for teachers, the most commonly used format is the multiple choice question (MCQ). MCQs are advantageous in that they 1) are a familiar format to students, 2) are reliable because each question has one clear answer, 3) are practical by allowing teachers to grade easily, and 4) can be applicable for testing proficiency at various levels. (Coombe, 2011).

However, despite these advantages, MCQs do not quite fit the scope of the present study's test design, as the *pretest* should avoid exposure of target words to learners as much as possible. To account for this, the present study adapts the method used in Kim's (2008) study, in which she gave a word bank and sentential question items with a blank; learners have to use a word from the word bank in order to fill in the blank in the sentence. All words (both target words and distractors) in the word bank play the role of distractor to one another. The test design satisfies the two points of Coombe's (2011) suggestion, as the format is familiar to students (i.e., gap filling, matching), and there is only one correct answer, which can be easily graded. The tests designed in the present study are explained in more detail in Chapter 5.

2.2 Intelligent Computer Assisted Language Learning (ICALL)

This section first addresses ICALL research involving semantic processing to date (2.2.1) and reviews ICALL systems for vocabulary acquisition and reading comprehension, which are directly related to the present study (2.2.2).

2.2.1 ICALL from grammatical to semantic processing

With the development of NLP technology, research on ICALL has largely focused on providing practice of grammatical forms for language learners (e.g., Heift's German Tutor, 2001; Heift and Nicolson's E-Tutor, 2001; Nagata's BANZAI, 2002; Amaral and Meurers' TAGARELA, 2006). Some ICALL systems (e.g., Heift's German Tutor, 2001) were designed to provide feedback on learners' grammatical forms by *simple* pattern matching between learners' answers and target answers that were pre-programmed in the system. This indicates that these types of systems do not deal with the meaning of the learners' answers. They thus have a limited scope, as they cannot handle learners' errors that do not match pre-stored "target-like" answers or errors that are grammatically correct but semantically inappropriate. These kinds of ICALL systems, which process linguistic exercises by simple pattern matching, perform well provided that there are no grammatical variations beyond their pre-programmed scope. By embedding more sophisticated NLP techniques, ICALL systems become able to not only determine how a learner's intended expression should be matched with a target-like answer in the system, but also to convert the learner's expression into a more target-like one without ruining the intended meaning (see discussion in Dickinson et al., 2008).

By precisely diagnosing learners' errors and providing detailed individualized feedback, ICALL systems have become "intelligent" in supporting language learning. However, when a task of ICALL systems involves semantic processing, it becomes much more challenging to make it feasible. There are a few ICALL systems that attempt to handle tasks involving semantic processing (e.g., Herr Kommissar (DeSmedt, 1995), AutoTutor (Graesser et al., 2001), FreeText (L'haire and Faltin, 2003)). DeSmedt

(1995)'s Herr Kommissar is worth noting because it provided relatively sophisticated semantic processing; it was designed as a role-playing detective game in which a learner played the role of a detective and asked the system to find a crime. When the system received the learner's question, it performed lexical identification, syntactic analysis, and semantic interpretation in sequence. For the semantic interpretation aspect, the system mapped the results of its lexical and syntactic analyses to an internal model of the input meaning, and then the knowledge representation schema (KRS), the concept ontology embedded in the system, examined the semantic results. Although it was able to handle a learner's input semantically, the semantic processing was still complicated such that the domain of the system was restricted in order to maintain semantic consistency. Other works (e.g., AutoTutor (Graesser et al., 2001), FreeText (L'haire and Faltin, 2003)) tried to semantically match a learner's response and the target answer of the system, but their works also showed that it was still quite difficult to make full semantic processing feasible. For example, Graesser et al. (2001) acknowledged that "AutoTutor cannot interpret student contributions that have no matches to anticipated content in the curriculum script". L'haire and Faltin (2003) also addressed that FreeText had limitations in detecting meaning errors from a learner's response if the response was not contained in the semantic component of their system.

This brief review of representative ICALL systems for language learning to date gives insight into how NLP techniques have been employed in actual language learning tasks and to what extent NLP techniques can support language learning that requires more complicated operations such as semantic processing. Most significantly, it is worth noting that there is a trade-off between the ability of ICALL systems to provide meaningful and

precise processing for language learning activities and the flexibility that ICALL systems allow for doing so (Dickinson et al., 2008). Although the current NLP technology is still too limited to handle every kind of language learning activity (Bailey & Meurers, 2008), especially those that involve a high degree of semantic processing, a learning task related to the level of word meaning processing can be made feasible using current NLP techniques (e.g. word sense disambiguation). For example, ICALL systems involved in vocabulary learning and reading comprehension show relatively effective handling of word meaning by employing advanced NLP technology (e.g., the REAP tutor (Heilman et al., 2006)). The ICALL systems to support vocabulary learning and reading comprehension are reviewed in the following section.

2.2.2 ICALL for vocabulary learning and reading comprehension

Studies in second language learning have shown that vocabulary assistance is fairly influential in vocabulary acquisition and reading comprehension. ICALL has also noted that vocabulary knowledge is one of the most important components for language learning and has thus been an issue (Goodfellow, 1995; Harley, 1996). However, little research has been done on actually building ICALL systems to provide assistance for vocabulary learning and reading comprehension (Ma, 2009). It seems that the reason for this shortage of ICALL systems for vocabulary learning and reading comprehension is that such systems require more sophisticated NLP technology due to semantic processing, such as the processing of word meaning. Despite this shortage, there are several ICALL systems supporting learners' vocabulary acquisition (Gamper & Knapp, 2001; Groot, 2000; Heilman & Eskenazi, 2008; Nerbonne & Smit, 1996; Shei, 2001). While some

systems focused only on vocabulary learning (e.g. Groot's CAVOCA (2000), Gamper & Knapp's ELDIT (2001), Shei's FollowYou! (2001)), other systems pursued vocabulary learning along with other language learning tasks such as reading comprehension (e.g. Heilman & Eskenazi's REAP tutor (2008), Nerbonne & Smit's GLOSSER-RuG (1996)).

The systems focusing solely on vocabulary learning (e.g., Groot (2000), Gamper & Knapp (2001), Shei (2001)) showed the typical use of a simple style of vocabulary learning exercises (Groot, 2000) or the simple use of an online dictionary (Gamper & Knapp, 2001; Shei, 2001). These systems were designed to provide general training for memorizing word meaning. For example, CAVOCA (Groot, 2000) was developed to bring a learner's conscious attention to the learning of new words. CAVOCA consisted of three stages in the vocabulary learning process: *deduction*, *usage*, and *examples*. In the *deduction* and *usage* stages, a learner was exposed to several sentences in which a target word occurred. The learner was then required to do a few exercises, such as finding synonyms and the correct use of a target word, and then received feedback immediately. In the *example* stage, the learner was exposed to passages containing the target words from the previous two stages. As a sense of the target word provided in all three stages was already fixed for exercises and feedback, word sense disambiguation was not an issue for this system. In general, CAVOCA exhibited an example of how to make good use of example sentences for vocabulary learning. CAVOCA is different from other systems (e.g. ELDIT, FollowYou!, GLOSSER-RuG) in utilizing example sentences for vocabulary learning as it used example sentences for exercises, whereas ELDIT, FollowYou!, and GLOSSER-RuG used example sentences as explicit lexical information for a word.

Unlike CAOVCAs, ELDIT (Gamper & Knapp, 2001) provided texts containing words to be learned. It allowed learners to practice new words by linking them to corresponding dictionary information, such as different meanings of the target word, collocations, translations, and illustrative examples extracted from the electronic learner's dictionary, which was developed in conjunction with the larger program. While GLOSSER-RuG and the REAP tutor gave definitions with the correct word sense when learners chose a word (explained in more detail later), ELDIT gave all sense definitions and example sentences regardless of sense for a given word in the text. Likewise, FollowYou! (Shei, 2001) provided definitions, collocations, synonyms, and example sentences when a learner clicked on a word in a text. However, the provided information was not sense-specific but rather was for all senses of a given word. Gamper and Knapp (2001) and Shei (2001) did not specifically discuss the quality of example sentences.

As shown, these systems used a computer as a simple tool of pattern matching for grammar practice or as an electronic dictionary for providing lexical information. These systems did not employ NLP techniques involved in semantic processing to handle various learning situations. Despite having tried to approach vocabulary learning in a reading text, these systems also overlooked the important role of the specific meaning of a word in a context.

On the other hand, there are some systems that have been aware of the importance of meaning disambiguation in vocabulary learning and reading comprehension (Heilman et al.'s REAP tutor (2008), Nerbonne & Smit's GLOSSER-RuG (1996)). These systems have been implemented to disambiguate the meaning of a word by employing NLP technologies such as word sense disambiguation (WSD). 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.

GLOSSER-RuG (Nerbonne and Smit, 1996) was designed to support learners' reading comprehension; it was developed to assist learners when they selected unknown or unfamiliar words while reading. It provided dictionary definitions of the words from a bilingual French-Dutch dictionary and example sentences extracted from corpora. In order to present definitions and examples of the correct sense for a given word, GLOSSER-RuG first disambiguated the sense of a word in the text on the basis of its part of speech (POS). POS disambiguation 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*). Nerbonne and Smit (1996) did not address this problem caused by multiple senses in the same POS; GLOSSER-RuG simply presented definitions for all senses if the given word had several senses associated with the same POS. In terms of a typical function of WSD, their disambiguation method was not truly well-functioning. It was not able to disambiguate the multiple senses of a word in the same POS, whereas the present study's system performs WSD at the level of word sense in full degree, by disambiguating each individual sense of every word.

Meanwhile, GLOSSER-RuG seemed to note the importance of presenting examples in addition to definitions; in order to present these examples, it extracted them from corpora, which provided examples more concretely. However, GLOSSER-RuG did not control the lexical or structural complexity of the examples for language learners, which indicates that they overlooked the quality of example sentences in terms of informativeness, readability, and typicality (Kilgarriff et al., 2008; Segler, 2002). In the case that learners

wanted to look up unknown or unfamiliar words in the examples, GLOSSER-RuG provided assistance in getting this lexical information by means of a dictionary. This solution could have a negative effect on learners' reading by disturbing their pace (Bogaards, 1998; Knight, 1994; Luppescu & Day, 1993). Unlike GLOSSER-RuG, the present study's system intends to provide examples that are authentic as well as grammatically-controlled for language learners without losing their illustrative value.

As a more advanced system with many NLP subcomponents for handling various functions (e.g., selecting reading texts according to learners' preference and proficiency, supporting vocabulary learning, building assessments, etc.), the REAP tutor (Heilmand et al., 2006; Heilman & Eskenazi, 2008; Kulkarni et al. 2008; Dela Rosa and Eskenazi, 2011) presented a more effective approach in providing a definition for the correct sense of a given word in context to support vocabulary learning. The REAP tutor utilized an advanced WSD method and, as a result, showed high performance in providing sense-specific information of a word. However, the REAP tutor had a limited scope; it provided sense-specific information exclusively for pre-determined words to be used for learning, and learners thus were not provided with sense-specific information for other words in the text. When considering the purpose of the REAP tutor to be assisting in vocabulary learning as well as its high performance of WSD (e.g., 88%), this approach is fine for the system to perform its task. However, the WSD methods employed in the REAP tutor required training data to classify the sense of words, and if the words were not contained in the training data, their WSD did not function. It was thus limited in its ability to extend coverage to any content words in a reading text, and was thereby less effective as a system for additional assistance in reading comprehension. Consequently, it restricted the

chances for learners to gain additional words beyond those that were predetermined. In the present study, the system provides sense-specific information for any content words in a text.

In addition, the REAP tutor did not provide specific examples; it simply provided definitions and examples (if any) extracted from a dictionary (i.e., *Cambridge Advanced Learners Dictionary*), which indicates that it initially overlooked the importance of example sentences for vocabulary learning. A later study, in the line of REAP tutor research by Kulkarni et al. (2008) became aware of the importance of example sentences in vocabulary learning; they recommended that example sentences be included in future projects. Moreover, Kulkarni et al. (2008) demonstrated that the REAP tutor was effective for vocabulary learning by providing readers with sense-specific lexical information for words. They experimented to see if there was any difference in vocabulary acquisition achievement when a definition with the correct sense was provided, as compared to when all sense definitions for a given word were provided. They found that when sense-specific definitions were provided, vocabulary acquisition was facilitated to a greater extent.

As reviewed thus far, there are only a few systems that specifically utilize WSD to process word meaning for vocabulary learning and reading comprehension. There remains much work on ICALL research for supporting vocabulary learning and reading comprehension to be done. As an extension of this research, the present study's system strengthens beneficial aspects of GLOSSER-RuG as an aid for reading comprehension and of the REAP tutor as an aid for vocabulary learning, thereby developing a more effective system to facilitate vocabulary learning and reading comprehension for second

language learners. In addition, the system developed in the present study is fairly light, focusing on the word level so that it does not require any complicated processes such as processing meaning of a whole text in order to provide a text fitted to a learner's level or interest. This in turn yields more consistent results.

2.3 Summary and conclusion

For second language learners, it is infeasible to attain all words from formal learning practices such as drill tasks to memorize word meanings. Rather, extensive vocabulary learning can be achieved through reading; when vocabulary learning and reading comprehension come together, they yield a synergistic effect. More enhanced vocabulary learning through reading can be accomplished by provision of more examples for a given word.

When learners perform tasks related to vocabulary learning or reading comprehension, the most prevalent problems are caused by learners' insufficient vocabulary knowledge. These problems can be reduced by the use of dictionaries, both paper and electronic. However, dictionary use still cannot resolve problems caused by lexical ambiguity (i.e., multiple senses of a word). ICALL research provides a solution for this problem by utilizing NLP technology. Some ICALL systems have tried to provide sense-specific lexical information by employing WSD technique, and their positive effect for learners' vocabulary learning and reading comprehension has been demonstrated by previous studies. Drawing upon the merits of these previous studies, the system of the present study is developed to assist vocabulary acquisition and reading comprehension for second language learners.

III. THE SYSTEM

The system built in the present study consists of a system manager, an NLP server, and a lexical database. This section describes in detail how those components were developed and what functions each performs to execute the overall goal of the study: the system manager (3.1), the NLP server (3.2), and the lexical database (3.3).

3.1 System manager

The system manager has the function of enabling a continuous interchange of information among a learner, the server, and the lexical database. Without the system manager, the server, the learner, and the lexical database cannot interact with each other to achieve the final goal of the system.

The system manager is realized through the web interface that is built in HTML (Hypertext Markup Language) as shown in Figure 3.2 for the page in Figure 3.1. The web page contains URLs (Uniform Resource Locators) and CGI (Common Gateway Interface) scripts, which control all transactions between the server and the learner. For example, in Figure 3.2, one of the URLs on the source page is *http://cl.indiana.edu/~se48/frame_11.html*, which shows a path to the source page and this is embedded in “*Fashion Victim GS*” on the page in Figure 3.1. So, when the learner clicks “*Fashion Victim GS*” in Figure 3.1, the embedded URL directs the learner to the page, “*frame_11.html*”. Likewise, a CGI script is executed when the learner clicks. For example, when the learner clicks a word in a text as shown in Figure 3.3, the CGI script, “*brdg.py*”, is called as shown in Figure 3.4. Then the CGI script is executed as

programmed in real-time on the server. It passes the value (i.e., word) chosen by the learner, runs its task on the server, and delivers its output to the learner (more descriptions are followed with the example, *brdg.py*, in Figure 3.5). All CGI scripts used in the present system were written in python.

The following is a step-by-step explanation of the system manager. After entering their personal information (e.g. name and email), the learner is directed to the main page for choosing a reading text, as shown in Figure 3.1. After the learner either uploads his own reading text from a local computer or chooses a text from the list of texts, the system manager sends the chosen raw text to the server.

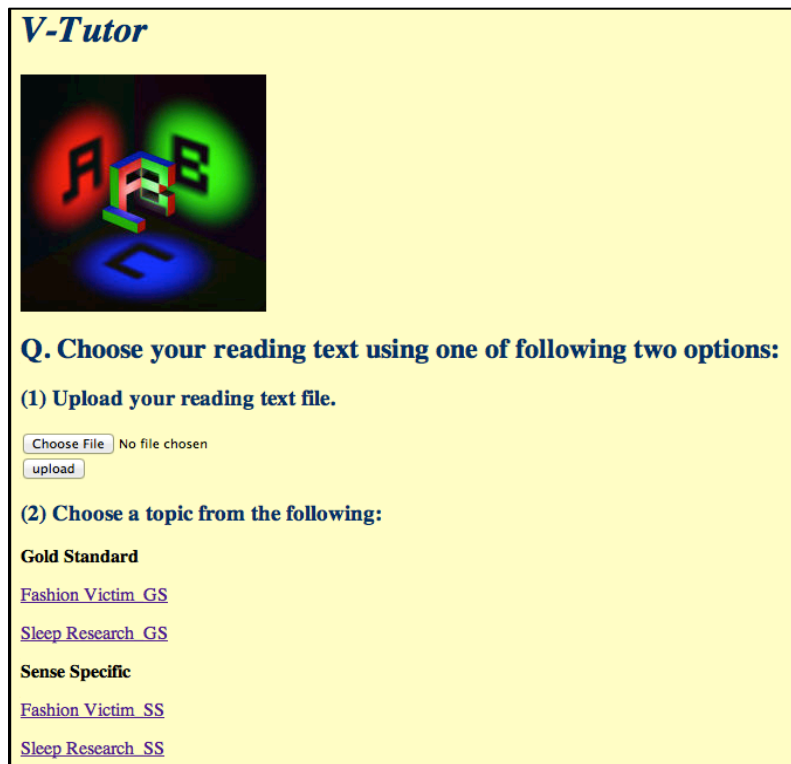


Figure 3.1. Web page to choose a reading text

As shown in Figure 3.2, the file from the local computer is uploaded by the learner and goes through a processing stage. The upload process is controlled by a CGI script (upload.py), which is called immediately after the file has finished uploading to the server (NLP processing, word sense disambiguation, linking all content words to the lexical database, etc.). Every text in the list is linked to its source page by URL.

```

<HTML>
<HEAD>
<TITLE> Have fun!! </TITLE>
</HEAD>

<BODY bgcolor = "#FFFFCC">
<h1> <i> <font color="#003366"> V-Tutor </font> </i> </h1>

<form>
<p> <h2> <b>

</b> </h2> </p>
</form>

<p><b><h2> <font color="#003366"> Q. Choose your reading text using one of following two options: </font> </h2> </b> </p>
<form enctype="multipart/form-data" action="upload.py" method="post">

<p><b><h3> <font color="#003366"> (1) Upload your reading text file. </font> </h3> </b> </p>
<input type="file" name="filename" >
<br>
<input type="submit" value="upload">
</form>

<p><b><h3> <font color="#003366"> (2) Choose a topic from the following: </font> </h3> </b> </p>

<p><b> Gold Standard </b></p>
<a href="http://cl.indiana.edu/~se48/frame_11.html"> Fashion Victim_GS </a></p>
<a href="http://cl.indiana.edu/~se48/frame_12.html"> Sleep Research_GS </a></p>

<p><b> Sense Specific </b></p>
<a href="http://cl.indiana.edu/~se48/frame_11_1.html"> Fashion Victim_SS </a></p>
<a href="http://cl.indiana.edu/~se48/frame_12_1.html"> Sleep Research_SS </a></p>

<p><b> All Senses </b></p>
<a href="http://cl.indiana.edu/~se48/frame_11_2.html"> Fashion Victim_AS </a></p>
<a href="http://cl.indiana.edu/~se48/frame_12_2.html"> Sleep Research_AS </a></p>

<p><b> No Sense </b></p>
<a href="http://cl.indiana.edu/~se48/showtext_11.html"> Fashion Victim_NS </a></p>
<a href="http://cl.indiana.edu/~se48/showtext_12.html"> Sleep Research_NS </a></p>

</BODY>
</HTML>

```

Figure 3.2. Page source of the example web page (Figure 3.1)

When the server processes the raw text and returns an analyzed text (Figure 3.4), the system manager moves the learner to the next page (Figure 3.3), presenting the text with information about the senses for all content words hidden and set to be clickable (Figure 3.4). As shown in Figure 3.3, the main task page is divided into three frames. The top frame provides a guide among web pages (e.g., main menu, reading text, vocab test); each of those buttons embeds a link (URL) to the relevant page and moves the learner to the page upon clicking. The left frame presents a reading text and the right frame displays lexical information for the word that the learner clicks in the reading text on the left frame. All content words in the reading text are set to be clickable so that upon being clicked it will provide relevant lexical information.

Let's start!	
Main Menu Reading Test Vocab Test	
<p>Q. Click a word that you want to look at its gloss</p> <p>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.</p> <p>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.</p> <p>In fact, people in developed countries shop so much that they are now discarding</p>	<p style="text-align: center;">expenditure</p> <p>[definition]</p> <p>1.[NOUN] [FORMAL] Expenditure is the spending of money on something, or the money that is spent on something.</p> <p>[example]</p> <p>Policies of tax reduction must lead to reduced public expenditure. They should cut their expenditure on defence.</p>

Figure 3.3: Presentation of information by the learner's clicking

When the learner clicks on a word, the system manager delivers the word with its associated information (e.g., part of speech, sense number) to the lexical database (i.e., html repository, see section 3.3 and Chapter 4). After receiving the word's sense-specific lexical information (e.g., definition and examples) from the lexical database, the system presents it to the learner (Figure 3.3) on the right frame. Whenever the learner clicks a word during their reading, the system manager passes every clicked word to the server (i.e., a user database) to record the learner's performance for later reference.

As shown in Figure 3.4, all content words in the text have their own POS and sense information appropriate to the context hidden in the code of the page (shaded parts). So, when the learner clicks the word, the CGI script (*brdg.py*) is called, passing the parameter (shaded part in Figure 3.4) to the server (1) to display the relevant html page on the right frame in Figure 3.3 and (2) to store the word the learner clicked.

```
...  
<a href="#" onclick="parent.Content.location='brdg.py?passing_word=price_n_1.html'">prices </a>  
have  
<a href="#" onclick="parent.Content.location='brdg.py?passing_word=stimulate_v_1.html'">stimulated </a>  
<a href="#" onclick="parent.Content.location='brdg.py?passing_word=exceptional_a_2.html'">exceptional  
</a>  
<a href="#" onclick="parent.Content.location='brdg.py?passing_word=growth_n_5.html'">growth </a>  
in  
<a href="#" onclick="parent.Content.location='brdg.py?passing_word=expenditure_n_2.html'">expenditures  
</a>  
on  
...
```

Figure 3.4. A part of an analyzed text

That is, as presented in Figure 3.5, the CGI script, *brdg.py*, performs its task in that it retrieves the relevant html page of the parameter (i.e., clicked word) received from the

web (shaded part in Figure 3.5) and stores the parameter in a user database (*record.txt*) for later tracking of the learner's performance.

```
#!/usr/bin/env python
# brdg.py : when a user clicks the word in the web, this calls relevant html page received as parameter
and stores those clicked words and action time in the server (record.txt)

import cgi
import cgitb; cgitb.enable()
import time
import urllib2

form = cgi.FieldStorage()
word = form.getvalue('passing_word')
r_word= word.split('_')

record =open('record.txt', 'a')
record.write('==word_clicked=====\n')
today = time.ctime()
record.write(today)
record.write("\n")
record.write(word)
record.write("\n")

url = 'http://cl.indiana.edu/~se48/Lexicon_htmls/'+ word
usock = urllib2.urlopen(url)
data = usock.read()
usock.close()

print "Content-type:text/html\r\n\r\n"
print "<html>"
print "<head>"
print "<title> hello </title>"
print "</head>"
print "<body>"
#print "<h2> hello %s </h2>" %(word)
print data
print "</body>"
print "</html>"
```

Figure 3.5. brdg.py

Upon the learner's completion of the reading, the learner may participate in taking a test of what they have learned. The tests are linked by URLs embedded in the button in the top frame (Figure 3.3). Thus, when the learner clicks the button for the test on the top frame, the system manager takes the learner to the web page of the test, delivering a

relevant test from the server to the learner (for the use of tests and materials, see Chapter 5).

Fundamentally, the system manager plays the role of a communication bridge by handling the flow of information among the learner, the server, and the lexical database.

3.2 Natural language processing server

The natural language processing (NLP) server consists of NLP modules and they are built in a UNIX environment. The NLP server functions as the linguistic intelligence of the system, processing the input data to present all requested information to learners. When the server receives a raw text (passage) from the web (learner), the NLP modules in the server – i.e. namely, the tokenizer, POS tagger, lemmatizer, and collocation finder convert the raw input text into a linguistically-analyzed text in that all necessary linguistic properties such as POS, lemmas, and collocations are annotated. Then, based on these linguistics properties, the word sense classifier disambiguates a sense (based on WordNet senses) of all the content words in the text. With a WordNet sense yielded from a WSD classifier, the server looks into the *Index* in the lexical database (see Chapter 4) to find a COBUILD sense corresponding the WordNet sense. Finally it returns the input text with a context-appropriate sense number (COBUILD) for every content word, with those sense numbers hidden from the learner, linked to the lexical database. The server also stores all words clicked by the learner during the learner's reading to track their performance later.

The NLP pre-processing functions associated with linguistic annotation including tokenization, POS tagging, lemmatization, and collocation finder are described in the

following section (3.2.1). One of the major NLP modules, word sense disambiguation (WSD) is discussed in the section 3.2.2. These modules employed in the NLP server are considered to be *state-of-the-art* or can be updated to be so depending on future developments.

3.2.1 Linguistic annotation

In the annotation phase of converting a raw input text to a linguistically-analyzed text, the system relies on several basic NLP modules for tokenizing, lemmatizing, POS tagging and identifying collocations. This section discusses those linguistic annotation tasks and the issues that arise at each step.

Before delving into those NLP modules, one remark should be made regarding the linguistic annotation in this research. As stated earlier, those linguistic annotation tasks are pre-processed prior to WSD. That is, the WSD system is applied to a pre-processed text. However, some WSD systems take a raw text as an input, some take a linguistically-annotated text from the pre-processing step, or some take either a raw input or a linguistically-annotated one. Thus, if the WSD system requires a linguistically-annotated text, then the modules described in the following sections (3.2.1.1 to 3.2.1.4) are applied in the pre-processing stage.

3.2.1.1 Tokenization

When the system receives an input text, the first task is to break it down into smallest parts which are called *tokens*. Accordingly, the system first splits the input text into *tokens*, which often correspond to words. This processing is called *tokenization* and this

is necessary for further processing such as POS tagging. Since *tokens* can be typically identified using whitespaces, it can be considered trivial for a language like English which delimits words by whitespaces.

However, simply splitting a text by whitespace is not a fully reliable solution for *tokenization*. For instance, if the text is split by whitespaces, a word in a possessive form is not separated (e.g. {world's} → [world's]), which should be split into a word and a possessive marker (e.g. [world, 's]). Likewise, contracted forms such as *n't* words (e.g. *can't*, *doesn't*) should be handled with their appropriate forms respectively (e.g. [can, n't], [does, n't]), and so should other contracted forms (e.g. 's, 'll). Apostrophe used in contracted/possessive forms and apostrophe used as a punctuation mark are treated differently; though it is rare, there is a case of using apostrophe as a punctuation mark, indicating certain forms of plurals (e.g. Mind your p's and q's) and in this case the apostrophe is treated as a separate token.

Indeed, punctuation marks need be treated as a separate *token*. Thus, if a word is followed by a punctuation mark such as a period or a comma, punctuation marks should be separated from the word as a *token* (e.g. {waste.} → [waste, .], {ground,} → [ground, ,]). However, at the same time, abbreviations (e.g. *A.*, *D.*) and hyphenated words (e.g. *World-wide*, *high-level*) should be handled differently. That is, if abbreviations and hyphenated words are tokenized as punctuation marks are treated, they are tokenized inappropriately - e.g., [A , .], [D , .], [World, - , wide], [high, - , level]), which should be tokenized as [A., D.], [World-wide], and [high-level]. The examples (1) and (2) illustrate how abbreviations, contracted forms, hyphenated words, and punctuations are appropriately tokenized

(1) input = “The world's nuclear plants have accumulated vast stocks of highly radioactive waste. World-wide, high-level waste is currently stored above ground, and no government has a clear policy on its eventual disposal.”

tokenized= ['The', 'world', "'s", 'nuclear', 'plants', 'have', 'accumulated', 'vast', 'stocks', 'of', 'highly', 'radioactive', 'waste', '.', 'World-wide,', 'high-level', 'waste', 'is', 'currently', 'stored', 'above', 'ground', '.', 'and', 'no', 'government', 'has', 'a', 'clear', 'policy', 'on', 'its', 'eventual', 'disposal', '.']

(2) input = “Mount Vesuvius, a volcano located between the ancient Italian cities of Pompeii and Herculaneum, has received much attention because of its frequent and destructive eruptions. The most famous of these eruptions occurred in A. D. 79.”

tokenized= ['Mount', 'Vesuvius', ',', 'a', 'volcano', 'located', 'between', 'the', 'ancient', 'Italian', 'cities', 'of', 'Pompeii', 'and', 'Herculaneum', ',', 'has', 'received', 'much', 'attention', 'because', 'of', 'its', 'frequent', 'and', 'destructive', 'eruptions', '.', 'The', 'most', 'famous', 'of', 'these', 'eruptions', 'occurred', 'in', 'A.', 'D.', '79', '.']

Since *tokenization* is the first step in NLP processing, high accuracy is crucial for the success of further processing, such as part-of-speech (POS) tagging. According to Grefenstette and Tapanainen (1994), a state-of-the-art tokenizer for English achieves an accuracy of 99.7%. The tokenizer in the system is first imported from the Natural Language Toolkit (NLTK) package (Bird, 2002) which is freely available. If a state-of-the-art POS tagger already has its own pre-built tokenizer (i.e., Stanford POS tagger, see the next section) as in this study, *tokenization* is performed by the POS tagger due to a more simple but robust processing.

3.2.1.2 Part-of-Speech tagging

Once the input text is split into tokens, a Part-of-Speech (POS) tagger is prompted to assign a POS tag to each of the tokens in a process called ‘POS tagging’. Careful and accurate POS tagging is a crucial pre-processing step for word sense disambiguation (WSD) due to words potentially having more than one POS. As shown in the following example (3) from a preliminary experiment, if a POS tagger assigns the wrong POS tag to a word, the WSD module would perform inappropriately based on the incorrect POS information.

(3) the world ’s nuclear *plants* have accumulated vast *stocks* .

NNS²/ VBZ³

NNS/VBZ

In example (3), if the POS tagger assigns VBZ to *plants*, the WSD module disambiguates its meaning as one of its verb senses (e.g. ‘to put a seed, plant, or young tree into the ground so that it will grow there’) which is incorrect in this context. The POS tag should be NNS and the meaning should be ‘a factory or a place where power is produced’. Likewise, if the word, *stocks*, is assigned with VBZ, the incorrect POS information would lead the WSD module to inappropriately disambiguate it as one of its verb senses (e.g. ‘to fill a cupboard, shelf, or room with food or other things’).

In this research, POS tagging is also important for identifying content words. Since the intent of this research is to disambiguate the sense of the content words, the input text

² NNS: plural noun

³ VBZ: 3 singular present verb

would first contain the POS information for every word and then the system would delete the POS tags from the function words and keep the POS tags only on the contents words for further processing, i.e., word sense disambiguation, as illustrated in (4).

(4)

the	world	's	nuclear	plants	have	accumulated	vast	stocks
DT	NN	POS	JJ	NNS	VBP	VBN	JJ	NNS



the	world	's	nuclear	plants	have	accumulated	vast	stocks
	NN		JJ	NNS		VBN	JJ	NNS

Among the freely available taggers, this study employed the Stanford POS tagger (Toutanova et al., 2003), based on the previous study that showed its accuracy of 97.24% (see discussion in Toutanova et al. 2003) on the data from the Penn Treebank Wall Street Journal. In the present study, the performance of the Stanford POS tagger was examined on the 1,005 tokens of the three sample texts (extracted from TOEFL reading texts) and the result gave an accuracy of 99%. The Stanford POS tagger is based on the Penn Treebank Tagset (Toutanova and Manning, 2000).

3.2.1.3 Lemmatization

Lemmatization is the task of converting various inflected forms in each content word to their base form (lemma). As you can see in the following example, the inflected form, *dog* and *dogs*, have the lemma “dog” in common and *walked*, *walking*, *walks* have “walk” as their common lemma.

(5) dogs, dog = “dog”

walked, walking, walks, walk = “walk”

The processing of lemmatization is applied to a POS tagged text, because the text has content word information based on the POS tagging results. If the content word is not lemmatized, the system would not be able to extract a word’s lexical information (e.g. definition and examples) from a lexical database, because the lexical database in this system is designed like a standard dictionary where one looks for a word meaning via its base form (lemma). Thus, the system cannot extract a word’s corresponding lexical information from the lexical database unless it receives the word’s lemma from the previous module, a morphological analyzer. Thus, all the content words in the POS-analyzed text are to have their base forms and POS tags as follows.

(6)

the	world	's	nuclear	plants	have	accumulated	vast	stocks
	(NN, world)		(JJ, nuclear)	(NNS, plant)		(VBN, accumulate)	(JJ, vast)	(NNS, stock)

Two freely available lemmatizers are compared: one from the NLTK package (Bird, 2002) and one from MontyLingua (Liu, 2004). When they are run on the 1,005 tokens of the three sample texts (extracted from TOEFL reading texts), NLTK showed an accuracy of 98.1% whereas MontyLingua showed 99.5%. Although MontyLingua showed very high accuracy, it still gave a consistent error on the comparative and superlative cases; for example, it lemmatized *fastest* as is, *fastest*. On the other hand, NLTK performed inconsistently on the plural nouns; sometimes it lemmatized correctly (e.g. stocks →

stock) but sometimes not (e.g. stocks → stocks). It, however, consistently performed correctly on the comparative/superlative forms. Thus, this system uses a combination of them; MontyLingua is first run and then NLTK is run for only the comparative and superlative forms.

3.2.1.4 Collocation identification

A collocation is a combination of two or more words which co-occur regularly. Usually, the combined words are composed of basic words which leads to learners easily overlooking them thinking that they know their meaning. Collocations are one of the major factors that stymie learners from their proper acquisition, because the meaning of collocations cannot be predicted from the meanings of their constituent words, as a new meaning is added when the words are combined. In this regard, the meaning of collocations is not fully compositional. The most extreme cases of non-compositionality are idioms (Manning & Schutze, 1999).

Because collocations are often non-compositional, they pose a significant challenge to research. If the WSD module, without a collocation identifier, annotates a word's sense, it results in an incorrect understanding of the text as follows.

(7) You shouldn't worry about the exam because you have prepared well and you will probably find it *a piece/8 of cake/3*.

If the words *piece* and *cake* are respectively disambiguated as '8. a serving that had been cut from a larger portion' and '3. made from or based on mixture of flour and sugar and

eggs' without collocation identification, they are interpreted literally. However, they should be interpreted together as a collocation meaning *very easy*.

Thus, before moving forward to the word sense disambiguation phase, the system needs a collocation identifier in order to boost its performance in disambiguating a word's sense. To successfully identify collocations and present their corresponding lexical information, the system needs two sub-components: a collocation identifier and a collocation database. The collocation identifier is designed to find collocations in an input text by pattern matching and disambiguating based on a collocation database. Then it changes the result of the POS tagged data replacing the POS tag(s) with collocation markers, as illustrated in (8).

(8) You should n't *worry* about the *exam* because you have *prepared* well and you

(VB, worry) (NN, exam) (VBN, prepare)

will probably *find* it ***a piece of cake***.

(VB, find) (NN, **piece**) (NN, **cake**)

→ You should n't *worry* about the *exam* because you have *prepared* well and you

(VB, worry) (NN, exam) (VBN, prepare)

will probably *find* it ***a _piece_of_cake***.

(VB, find) (COL)

The current system uses simple collocation modules which come pre-built as part of the WSD system (i.e., SenseRelate::AllWords (Pedersen and Kolhatkar, 2009)) employed in this research.

3.2.2 Word sense disambiguation (WSD)

Word sense disambiguation (WSD) plays two fundamental roles in this research. The first is to disambiguate the sense of all content words (e.g. nouns, verbs, adjectives, and adverbs) in a given text, as in (9). As shown in the example, the WSD module takes an input word, determines its sense by its use in its input context, and outputs the word with its appropriate sense used in its given context.

(9) The world/4 's nuclear/1 plants/1 have accumulated/2 vast/1 stocks/4 of highly/1 radioactive/1 waste/5.

The other role is to provide a basis for word sense alignment (WSA, see Chapter 4); the WSD module provides sense probability distributions, which are used in the processing of WSA. In the present study, the lexical database is built based on WSA, in that the WSA algorithm performs on sense probability distributions yielded by WSD. That is, when the lexical database is built, WSD is first processed on example sentences extracted from COBUILD, generating sense probability distributions for each of the examples; WSD gives the probabilities of every WN sense for each example sentence. For example, if the word *area* in a given COBUILD example sentence has six WN senses in total, WSD generates the sense probability distribution of six WN senses for the given COBUILD example, as in (10).

(10) You will notice that your baby has two soft *areas* on the top of his head.

[(wn₁, 0.20), (wn₂, 0.08), (wn₃, 0.25), (wn₄, 0.11), (wn₅, 0.21), (wn₆, 0.14)]

wn₁: a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography)
wn₂: a subject of study
wn₃: a part of an animal that has a special function or is supplied by a given artery or nerve
wn₄: a particular environment or walk of life
wn₅: a part of a structure having some specific characteristic or function
wn₆: the extent of a 2-dimensional surface enclosed within a boundary

A WSA algorithm operates on the sense probability distributions in the WSA step, which will be explained in more detail in Chapter 4. For the disambiguation role of providing sense probability distributions, a few different WSD systems are employed to test the effectiveness of performing alignment based on these systems. In general, there is a tradeoff between precision and recall with all of these systems, i.e., higher-accuracy systems tend to have less coverage. Taking this balance into consideration, three WSD systems (i.e., SenseRelate::AllWords, SenseLearner 2.0, and Naïve Bayes) were selected to experiment with; since distributions of senses as output are also needed, as mentioned, how to extract them is addressed, if it is not obvious.

Performance of each of the following WSD systems is evaluated and discussed with respect to WSA in Chapter 4. As addressed in Chapter 4, SenseRelate::AllWords (Pedersen and Kolhatkar, 2009) outperformed the other two WSD systems (i.e., SenseLearner 2.0 and Naïve Bayes). Thus, SenseRelate::AllWords (Pedersen and Kolhatkar, 2009) is employed for both (1) disambiguating the sense of the words in the text and (2) generating sense probability distributions for word sense alignment (see Chapter 4). Although the quality of SenseRelate::AllWords (F-measure of 54-61%) is not

much satisfactory, it is regarded as one of state-of-the-art WSD systems in terms of broad coverage of content words and thus sufficient to give a point to work from.

3.2.2.1 SenseLearner 2.0

SenseLearner 2.0 (Mihalcea and Faruque, 2004; Mihalcea and Csomai, 2005) is a state-of-the-art large-coverage WSD system (85.6% coverage), that uses relatively little sense-annotated data. Rather than building a separate classifier for every individual word, it builds general “semantic models” for groups of words sharing some common syntactic or semantic properties based on a small sense-annotated corpus, SemCor (Miller et al., 1994). Consequently, the semantic models become general enough to be able to disambiguate words that are covered by the word categories as well as words presented in the training corpus so that they can have a larger coverage.

The algorithm of the SenseLearner system starts with a text preprocessed by linguistic annotation such as tokenization, POS tagging, and collocation identification. At the same time, semantic models are trained for predefined word categories. The current SenseLearner system has seven semantic models and new models can also be defined and trained using a template included in SenseLearner. During the training, a feature vector, consisting of the target word and its corresponding sense, is created for each semantic model based on the sense-annotated data (training data). Similar feature vectors are constructed for all content words in the input text (test data). Then, in a separate training process, each vector, constructed for words in the input text, is labeled with a predicted word and sense by the TiMBL memory based learning algorithm (Daelemans et al. 1998; 2001; 2010). Finally, if the sense predicted by TiMBL is the same as the sense predicted

by a semantic model in SenseLearner, the predicted sense is used for sense-annotating the target words in the input text. If the predicted sense by TiMBL is different from the sense assigned by a semantic model, sense-annotation is not produced.

To obtain the sense probability distribution, in this study, TiMBL is directly trained using the same data and options as described in SenseLearner’s documentation.

3.2.2.2 WordNet::SenseRelate::AllWords

SenseRelate::AllWords (SR::AW) also has broad coverage (Pedersen and Kolhatkar, 2009). It relies solely on a knowledge source (WordNet), without requiring training with sense-annotated data; accordingly, it covers all the words in WordNet.

The SR::AW system first sets a window of context for a target word and measures similarity scores between all the target word’s possible senses and the possible senses of each word in the window, using WordNet::Similarity⁴ (Pedersen et al., 2004). Then, the system selects the sense of the target word with the maximum score.

The SR::AW system can start with a few kinds of input format. It can either be a raw text or a POS-tagged text (noun, verb, adjective, adverb) converted from a text tagged with Penn Treebank tagsets as shown below:

(11) (raw) The astronomer married a movie star.

(wn-tagged) The astronomer#n married#v a movie-star#n

- (from Pedersen & Kolhatkar, 2009)

⁴ WordNet::Similarity (Pedersen et al., 2004) provides six measures of similarity and four measures of relatedness. Measures of similarity are based on the information of “is-a hierarchy of concepts” and limited to comparisons words within the same part of speech. Measures of relatedness are based on information beyond “is-a hierarchy of concepts” and can compare words across parts of speech.

As shown above, SR::AW identifies each WordNet compound and tags it as a noun in a preprocessing stage (e.g. movie_star) to enhance overall accuracy. After compounding, the stoplist check is done by asking a user if the user wants to use the user's own stoplist or the default stoplist. Once preprocessing is complete, the SR::AW algorithm starts its disambiguation job, providing the user with a few options; the choices of the measure of similarity and relatedness, the window size of the context for determining a word's sense, and several context scoring thresholds. For the option of context scoring thresholds, one can set "a minimum threshold that a sense of the center word should achieve with all the words in the context in order to be selected." (Pedersen & Kolhatkar, 2009, p.19). Based on the options chosen by the user, the SR::AW system disambiguates the input text by measuring similarity scores and finally generates the output text with WordNet sense tags on all content words.

In the present study, the system generates sense probability distributions by having SR::AW output non-zero scores for each sense and converting the scores to relative frequencies. Also, the system uses this WSD module for disambiguating the sense of words in an input text.

3.2.2.3 Naive Bayes

A Naïve Bayes (NB) WSD system has been built in the present study, trained on the DSO corpus (Ng and Lee, 1997). The Naïve Bayes WSD system uses a simple statistical method to disambiguate senses of a word based on frequencies found in a sense-tagged corpus; despite its simplicity, it has obtained high accuracy (Leacock et al., 1993; Mooney, 1996; Ng and Lee, 1996; Pedersen and Bruce, 1997).

The NB WSD system uses *Bayes Decision Rule* - i.e., *Decide s_k if $P(s_k|c) > P(s'|c)$ for $s_k \neq s'$* - which is if s_k is bigger than s' given context c , choose s_k as a sense of the word. That is, the NB classifier decides the word sense, s_k , that has the maximal probability of the sense given the context. In order to compute the maximal probability, the system calculates two probabilities; the probability estimated as the ratio between the number of examples of sense s_k and the total number of examples in the training set; and the probability of observing the context features given the observed sense s_k . While it can have higher accuracy, it tends to have less coverage than the other systems because the NB WSD system only covers words presented in training, being dependent on sense-tagged training data.

3.3 Lexical database

The most important component of the system, the lexical database, is used to provide a sense-appropriate definition and example sentences of an input word to a learner. When the system receives the input word from the learner, the system looks over the word and its specific sense number in the lexical database and presents a definition and examples of the appropriate sense of the word to the learner.

The lexical database has two parts: the *Index* and the *HTML repository* (see Chapter 4). In order to link automatic WSD systems (WordNet sense) with learner-appropriate examples (COBUILD sense), the *Index* is built with the list of WordNet-COBUILD pairs by a word sense. When the WSD classifier disambiguates a word based on WordNet sense, the system finds the COBUILD sense paired with the given WordNet sense via the *Index* in the lexical database. The *Index* then sends the COBUILD sense to the *HTML*

repository, which is composed of *HTML pages* containing lexical information of all content words organized by their COBUILD senses (e.g. mend_v_1.html). The critical function of the lexical database is to redirect a WordNet sense for each word to its corresponding COBUILD lexical information. The lexical database is discussed in much greater detail in the following chapter.

IV. THE LEXICAL DATABASE

4.1 Introduction

The lexical database is used to provide sense appropriate lexical information (definition and example sentences) for an input word to a language learner; lexical information thus should be customized for language learners. As stated in the previous chapters, to obtain sense appropriate information, a WSD system is first applied to the input text. With an output sense (WordNet sense) generated by the WSD system, pointing to appropriate lexical information (COBUILD) is complicated due to the fact that the database of lexical information is from one sense inventory (tuned for language learners, COBUILD) while automatic WSD systems generally use senses from another (WordNet).⁵ The lexical database, then, is indexed by WordNet senses, each of which points to an appropriate corresponding COBUILD sense. To build this lexical database, a method of word sense alignment (WSA) is explored here, thereby aligning two sense inventories: WordNet and COBUILD. Focusing on aligning WSD output with the lexical database of learner-appropriate lexical information, the WSA algorithm in this research is applied to the outputs from the WSD system, thereby building from the state-of-the-art. The WSA approach addressed here is a unique contribution of this research.

In order to evaluate the WSA system built in this research, the evaluation data was built by collecting alignment judgments from linguistics students and faculty (semi-

⁵ It could be feasible to develop a WSD system directly based on COBUILD senses, as there has been previous research trying to build a WSD system that was not based on annotated training data (e.g., Navigli and Lapata (2010)) and this can be a possible avenue pursue to directly build a COBUILD classifier. However, the current research is more interested in looking into different lexical resources and subsequently is focused on that.

expert) by online surveys. The evaluation data was a small set of nine words, covering 63 WordNet senses.

The lexical database constructed by the WSA approach of this study has two piece of information; the *Index* that included a word, its part of speech (POS) and its WN-COBUILD sense pairs, which are listed by word, as shown in Figure 4.1 and the *html repository* that has COBUILD lexical information by word sense as shown in Figure 4.2.

action	n	('w1', 'c5') ('w2', 'c4') ('w3', 'c5') ('w4', 'c4') ('w5', 'c3') ('w6', 'c2') ('w7', 'c1') ('w8', 'c3') ('w9', 'c1') ('w10', 'c2')
activity	n	('w1', 'c1') ('w2', 'c1') ('w3', 'c2') ('w4', 'c3') ('w5', 'c3') ('w6', 'c2')
admit	v	('w1', 'c3') ('w2', 'c3') ('w3', 'c4') ('w4', 'c4') ('w5', 'c1') ('w6', 'c2') ('w7', 'c2') ('w8', 'c1')
agitate	v	('w1', 'c2') ('w2', 'c3') ('w3', 'c2') ('w4', 'c1') ('w5', 'c3') ('w6', 'c1')
air	n	('w1', 'c3') ('w2', 'c4') ('w3', 'c3') ('w4', 'c1') ('w5', 'c4') ('w6', 'c1') ('w7', 'c3') ('w8', 'c3') ('w9', 'c3')
allow	v	('w1', 'c4') ('w2', 'c3') ('w3', 'c1') ('w4', 'c5') ('w5', 'c2') ('w6', 'c6') ('w7', 'c5') ('w8', 'c6') ('w9', 'c1') ('w10', 'c3')
alternate	a	('w1', 'c3') ('w2', 'c1') ('w3', 'c2') ('w4', 'c2')
ambition	n	('w1', 'c1') ('w2', 'c2')
....		

accord_a_1.html
accord_a_2.html
accord_a_3.html
accord_a_4.html
accord_a_5.html
accumulate_v_1.html
action_n_1.html
action_n_2.html
action_n_3.html
action_n_4.html
action_n_5.html
activity_n_1.html
activity_n_2.html
activity_n_3.html
actor_n_1.html
admit_v_1.html
admit_v_2.html
admit_v_3.html
admit_v_4.html
agitate_v_1.html
agitate_v_2.html
agitate_v_3.html

Figure 4.1. A part of the *Index*

Figure 4.2. A part of the *html repository*

When the system, for instance, receives a word, its POS, and WN sense (e.g. *agitate_v_3* = *agitate_verb_WordNet* sense #3) from the WSD module in the previous stage, it finds and sends its corresponding COBUILD sense (e.g. *agitate_v_2* = *agitate_v_COBUILD* sense #2) to the *html repository* as shown in Figure 4.3. All html pages in the repository are composed of COBUILD lexical information. The system then

retrieves a relevant html page (e.g. agitate_v_2.html) from the html repository and presents its lexical information to the learners. Everything is processed automatically in order to provide coverage for any content word in the inventories.

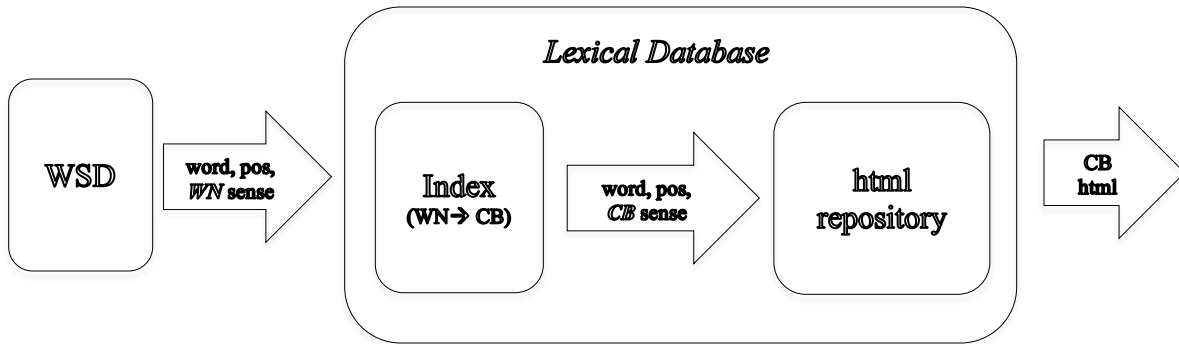


Figure 4.3. The lexical database

Section 4.2 provides the background for the present study in terms of WSA methodology and evaluation. Sense inventories employed in this research are described in section 4.3. Section 4.4 provides the alignment algorithm, discussing the assumptions based on the context of the present study (i.e., an online system providing vocabulary assistance). The WSA method developed in this research is finally evaluated on a dataset, which is discussed in section 4.5.

4.2 Related works

Before addressing previous WSA works related to this research, a general concept of alignment is first introduced. Like other general mappings, there are three kinds of mappings in sense alignment: one-to-one, one-to-zero, and one-to-many alignments (Ide

and Veronis, 1990). As shown in Figure 4.4, each inventory has its own list of senses for every word; for example, WordNet has a list of four senses and COBUILD has a list of three senses for a word, *plant.n*. Between two lists of senses for the word, one-to-one mapping is a direct mapping between one sense in the one inventory and its corresponding sense in the other inventory; in Figure 4.4, mapping between the sense in WordNet (W2) and the sense (C1) in COBUILD shows one-to-one mapping.

WordNet		COBUILD
W1. plant, works, industrial plant -- buildings for carrying on industrial labor		C1. A plant is a living thing that grows in the earth and has a stem, leaves, and roots
W2. plant, flora, plant life -- a living organism lacking the power of locomotion		C2. A plant is a factory or a place where power is produced
W3. plant -- something planted secretly for discovery by another	?	C3. Plant is a large machinery that is used an in industrial processes
W4. plant -- an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience	?	

Figure 4.4. Three kinds of mapping (*plant.n*)

The other two kinds of mapping (i.e., one-to-zero and one-to-many) often occur due to the relatively fine-grained nature of one inventory compared to another. In Figure 4.4, for example, a one-to-zero mapping is the case where the WordNet senses (W3, W4) have no

corresponding sense in COBUILD. Conversely, a one-to-many mapping is the case where one WordNet sense (W1) maps onto two or more senses in COBUILD (C2, C3). These three kinds of alignments have nothing to do with mapping direction between inventories; for example, if WordNet and COBUILD switch sides, the mapping categories would still be the same.

The major difficulty in counting practicalities in integrating different inventories, lies in the fact that every sense inventory has its own purpose (e.g., for lexicography, computational disambiguation systems, or language learners) and thus has its own particular design, with some including hierarchical information, some having illustrative examples for senses, some based on thesaurus information, some providing etymologies and translations, and so forth.

Partly owing to these differences, there has been much work in aligning senses between inventories for a number of applications, for example, in building a large-scale lexical database for machine translation (Knight and Luk, 1994); comparing the performance of different natural language processing (NLP) systems characterizing lexical semantics (Nastase and Szpakowicz, 2001); or reducing the granularity of an inventory (WordNet) for NLP (Navigli, 2006). Moreover, increasing the scale of lexical resources is an ongoing task, especially for newer resources such as Wikipedia and Wiktionary (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011; Ponzetto and Navigli, 2010).

This section reviews automatic word sense alignment works to date. Section 4.2.1 (Early WSA) and 4.2.2 (Recent WSA) review previous studies focusing on their WSA methodology. Section 4.2.3 (WSA evaluation) goes over previous works in terms of their

evaluation methods. Lastly, section 4.2.4 (Conclusion) states the differences and advantages of the present study's WSA work as compared to the previous WSA works.

4.2.1 Early works of word sense alignment

To start with, Ide & Veronis (1990) combined the Collins English Dictionary and Oxford Advanced Learner's Dictionary to create a comprehensive knowledge base. They proposed a spreading activation algorithm to remedy the shortcomings of the Lesk algorithm (Lesk, 1996). The main idea behind the Lesk algorithm is to disambiguate words by finding overlaps among their sense definitions (Mihalcea, 2006). However, the original Lesk algorithm has some limitations in that it does not resolve cases when no shared words are present or when the same number of shared words appears with more than one sense distinction (Ide & Veronis, 1990).

Their spreading activation algorithm created a network for the words from one dictionary's sense definition by building word-to-sense-to-word links; the network for a given word was constructed by words (nodes) of sense definitions in which every word was further linked to its sense definitions, and so forth. Once they constructed a network for every word, they iterated the process by giving weights on the nodes corresponding to the input word. In the end, one sense node with the strongest relation was determined. Although they showed 97% accuracy and were aware of one-to-many mapping, they only dealt with one-to-one mapping. Their algorithm seemed limited in being able to extend to cover a large number of words because the algorithm required a great deal of memory space; the network in the algorithm contained 13,627 total transitions for 59 input senses.

Therefore, the spreading activation algorithm they proposed does not appear efficient enough to be applied to real sense mappings like the one used in the present study.

Knight & Luk (1994) constructed a large-scale knowledge base for machine translation by merging existing resources – WordNet and Longman Dictionary Of Contemporary English (LDOCE). They proposed a definition match algorithm and a hierarchy match algorithm. Their definition match algorithm was also based on the Lesk algorithm; if there were overlapped words between two sense definitions, those two senses were matched. They were also aware of the issues of the Lesk algorithm - e.g., no shared words, shared words on more than two senses, so they used additional information to resolve the issues of the Lesk algorithm; they extracted WordNet's synonyms and superordinates. Then a dimensional matrix was constructed for each inventory (i.e., WordNet and LDOCE) and the two matrices were multiplied, yielding a similarity matrix (*SIM*). Eventually the *SIM* which had the largest value was determined for mapping. They noted their algorithm was not sufficient due to the fact that sense definitions between two inventories often have no words in common, which is a typical problem of the Lesk algorithm. Also they addressed a trade-off between correct mapping and mapping coverage; for example, by using the definition match algorithm, they yielded 90% correct mappings but with 27% coverage. To complement it, they additionally applied the hierarchy match algorithm. The basic concept of the hierarchy match algorithm was that once two senses were matched (from the results of the definition match algorithm), the system could check their respective ancestors and descendants for further matches. The algorithm handled 11,128 noun sense matches at an accuracy of 96% by operating in several iterative phases. Even though the accuracy of the results was

impressive, research along the lines of Knight and Luk (1994) may not fit the scope of this research because they used the sense hierarchy of WordNet and LDOCE which is different from the present study, in which WordNet and COBUILD are employed and COBUILD does not provide those kinds of hierarchy (e.g., ‘semantic code’ such as human (H), plant (P), etc., and ‘genus sense’ (=head noun)) available from LDOCE. Like Ide and Veronis’ (1990) work, they focused on one-to-one mapping, not on the overall structure of alignment that is a concern of the present study.

Kwong (1998) also aligned LDOCE and WordNet in a way similar to Knight and Luk (1994), but further incorporating Roget’s Thesaurus. By linking up three different kinds of resources - WordNet, LDOCE, and Roget’s Thesaurus, Kwong (1998) tried to construct one comprehensive lexical resource organized in a suitable way for a variety of NLP tasks, providing all required information. Kwong (1998) tested three groups of nouns, divided by a degree of polysemy where each group had 12 random nouns. Following Knight and Luk’s similarity measure (*SIM*), he used a similarity method in that he built a similarity matrix multiplying a matrix for LDOCE by extracting definition words and a matrix for WordNet by extracting definition words, hypernyms and coordinates. The average accuracy across three groups was about 60% (Hi-polysemy (13.18 senses); 52.96%, Med-polysemy (7.05); 65.63%, and Lo-polysemy (1.18); 64.77%). Like the other studies, Kwong’s (1998) study focused on one-to-one mapping between senses, not on the overall structure of alignment.

While the above studies could not handle one-to-zero and many mappings, Navigli (2006) attempted to deal with one-to-zero mappings with relatively various content words (e.g. 466 nouns, 231 verbs, 50 adjectives, 16 adverbs). Like the above studies, Navigli’s

mapping method was also based on semantic relatedness between two inventories; in his work, WordNet and the Oxford Dictionary of English; he first built a “sense description, $d_D(S)$ ” for each dictionary based on textual definitions, hypernyms, and domains between dictionaries: “ $d_D(S) = \text{def}_D(S) \cup \text{hyper}_D(S) \cup \text{domain}_D(S)$, where D is either WN or ODE and S is a given sense” (Navigli, 2006, p.842), and computed similarity between sense descriptions of both dictionaries. Then he acquired best matching by his defined function as follows: “ $\hat{m}(S) = \underset{S' \in \text{Sense}_{ODE}(w)}{\text{argmax}} \text{match}(S, S')$ ” (Navigli, 2006, p.842). What has to be noted here is that he also defined mapping μ for one-to zero mapping: “ $\mu: \text{Sense}_{WN} \rightarrow \text{Sense}_{ODE} \cup \{\epsilon\}$, where Sense_D is the set of senses in the dictionary D and ϵ is a special element assigned when no plausible option is available for mapping (e.g., when the ODE encodes no sense entry corresponding to a WN sense)” (Navigli, 2006, p.842). As seen in his function μ , Navigli (2006) handled one-to-zero mapping using special element ‘ ϵ ’. The result showed an accuracy of 66.08%. However, although Navigli (2006) tried to handle a greater variety of content words and one-to-zero mapping, his study focused more on clustering fine-grained senses of WordNet into the more coarse-grained ODE. This work is thus quite different from the WSA work of the present study which deals with two inventories in a more balanced way. Also, the source information used for their WSA is not available in the inventories used in the present study (e.g., COBUILD does not provide information about hypernyms or domains)

In sum, those previous works on sense alignment have primarily used elaborations on the Lesk algorithm as a starting point (Ide & Veronis, 1990; Knight & Luk, 1994; Kwong, 1998; Nastase & Szpakowicz, 2001; Navigli, 2006). Previous studies either developed from the Lesk algorithm (Ide & Veronis, 1990; Knight & Luk, 1994; Kwong,

1998; Nastase & Szpakowicz, 2001; Navigli, 2006) – e.g., they used more information such as synonyms, hypernyms, semantic hierarchy, domains, etc., beyond words of sense definitions, or implemented an additional algorithm to boost its performance (e.g., Knight & Luk (1994)'s hierarchy match algorithm). The key to note here is that these studies used information (for WSA processing) that is not available from COBUILD. By that, one can also note that each WSA work is dependent on the nature of inventories used, not to mention the purpose of its own WSA work. Although their methods are hard to make it feasible in the present study, it is worth keeping in mind that the most favorable method for early WSA studies is a similarity measure between two inventories. Besides, most of them (Ide & Veronis, 1990; Knight & Luk, 1994; Kwong, 1998; Nastase & Szpakowicz, 2001) attempted to account only for one-to-one mappings between single senses, thus not handling one-to-many mappings that appear to be common, nor the overall structure of sense alignment between inventories. More recent works tried to examine those issues, which are reviewed in the following section.

4.2.2 Recent works of word sense alignment

In the last few years, there has been a rapid increase in the number of works involved in aligning WordNet with Wikipedia or Wiktionary (e.g. Meyer and Gurevych, 2011; Neimann and Gurevych, 2011; Navigli and Ponzetto, 2010; Ruiz-Casado et al., 2005; Toral et al., 2009). Since Wikipedia covers a wide variety of information and Wiktionary offers a great deal of linguistic information, they have turned out to be very promising resources for many NLP applications (Meyer and Gurevych, 2011).

The first work in this vein of aligning WordNet and Wikipedia is Ruiz-Casado et al. (2005)'s in that they aligned WordNet synsets and simple Wikipedia entries. Ruiz-Casado et al. utilized a method of overlapping, in particular, a vector-based similarity metric between Wikipedia entries⁶ and WordNet synset's gloss. Taking a slightly larger resource of Wikipedia (i.e., Wikipedia categories⁷), Toral et al. (2009) tried to align WordNet synset (noun only) and Wikipedia categories using semantic similarity based on textual entailment and semantic relatedness between WordNet synset's definitions and Wikipedia articles (i.e., entries) or abstracts that belonged to categories. So, in addition to inventories to align, these two studies (Ruiz-Casado et al., 2005; Toral et al., 2009) are similar in alignment method. They both used a semantic similarity metric to align WordNet and Wikipedia.

With a different purpose from those previous ones (Ruiz-Casado et al., 2005; Toral et al., 2009) which tried to enlarge resources, Navigli and Ponzetto (2010) aligned WordNet and Wikipedia to construct BabelNet to support Machine Translation (MT). They tried to map WordNet senses (concepts) and synsets (relations) to Wikipedia pages/entries (concepts) and hyperlinked texts (relations). They used a conditional probability $p(s|w)$, with WordNet sense s given Wikipedia article w .

While all of them (Ruiz-Casado et al., 2005; Toral et al., 2009; Navigli and Ponzetto, 2010) did not allow multiple mapping, that is, they assigned a single, most likely WordNet sense to each of the Wikipedia categories/articles, Neimann and Gurevych

⁶ "A Wikipedia *article*, or *entry*, is a page that has encyclopedic information on it." (http://en.wikipedia.org/wiki/Wikipedia:What_is_an_article%3F).

⁷ "Wikipedia *categories* are intended to group together pages on similar subjects. This helps readers to navigate, sort, find related articles and see how information is organized." (<http://en.wikipedia.org/wiki/Help:Category>).

(2011) noted that there are three kinds of mapping; one-to-zero mapping, one-to-one mapping, and multiple mappings. Having this in mind, they tried to map WordNet and Wikipedia articles. They extracted all alignment candidates from both resources and matched candidate alignments by computing cosine word overlap similarity, which is similar to the method used in the above studies (Ruiz-Casado et al., 2005; Toral et al., 2009).

The alignment methods proposed in those previous studies (Ruiz-Casado et al., 2005; Toral et al., 2009; Navigli and Ponzetto, 2010) are worthy of note with respect to the fact that they tried to do the alignment work on a large scale (integrating full resources), which is remarkably different from other previous works (Ide & Veronis, 1990; Knight & Luk, 1994; Kwong, 1998; Nastase & Szpakowicz, 2001; Navigli, 2006). However, the resource (Wikipedia) employed to align with WordNet is quite different from the resource that is used in the present study. In other words, Wikipedia is characteristically different from a standard dictionary in many aspects (e.g. organization and kinds of information). Those works would be more useful for providing information to NLP works involved in, for example, ‘named entity recognition’ (Meyer and Gurevych, 2011). Overall, information which is input to the WSA algorithm between those previous studies and the present study are fairly different; some information does not even exist in COBUILD (e.g., semantic hierarchy, domains, etc.), and thus it is hard to employ or adapt their method to the approach of the present study. However, their key idea using similarity scores or semantic relatedness would be worthwhile to consider in designing alignment methods.

Meyer & Gurevych (2011) noted that Wiktionary was a more enriched lexical resource and tried to align word senses between WordNet and Wiktionary. This work of aligning WordNet and Wiktionary is more interesting to the present study with respect to the fact that they tried to integrate two standard kind of lexical resources (i.e. dictionaries) on a full scale. Meyer & Gurevych (2010) first tried to manually align a small number of senses between WordNet and Wiktionary and then proposed the first fully *automatic* alignment method between WordNet and Wiktionary (Meyer & Gurevych, 2011). Following Meyer & Gurevych's (2011) study, Niemann & Gurevych (2011) performed alignment work in two steps: candidate extraction and candidate alignment. In the step of candidate extraction, they extracted all candidate alignments: all synsets in WordNet and all word senses from Wiktionary. Then they aligned candidate alignments in the candidate alignment step, by two similarity measures; the cosine similarity and the personalized PageRank based measures employed from Agirre and Soroa's (2009) work. The performance of Niemann and Gurevych's (2011) study was as follows: precision =.674, recall=.649, and f-measure =.661. (cf., their baseline (most frequent sense); precision=.329, recall=.508, f = .399). As addressed, Niemann and Gurevych's (2011) method allows many-to-many mapping. Besides, they were able to avoid search space problem that is typical in alignment works. In other words, in the two alignment steps, for example, they took two possible candidates and asked yes or no concerning whether there was an alignment between them. Thus if there are five senses and five senses, the possible alignments are up to 25 cases⁸. However, their work still has the limitation that they did not consider the alignment structure as a whole.

⁸ WordNet sense_1 can be mapped to Wiktionary sense_1, sense_2, sense_3, sense_4, sense_5, respectively (5 mappings). The other four WordNet senses can be mapped to each of the five Wiktionary

As shown in the review of previous studies, there have been few commonly used resources across any single kind of NLP task, meaning that most works are still limited and end up being application-specific. In that sense, the approach of the present study is another piece of the puzzle. First, a different set of inventories (WordNet and COBUILD) is employed in the present study, which thus increases the scope of WSA work. Second, the purpose of alignment in the present study is unique: instead of increasing the size of a knowledge base (e.g., a sense inventory) as most previous studies have done, the present study intended to find the areas of commonality by alignment between resources, in order to automatically extract appropriate information from one of the resources.

4.2.3 Word sense alignment evaluation

This section reviews some literature, focusing on what evaluation data has been used in determining sense alignment quality. In the previous work introduced in the section above (Ide and Veronis, 1990; Knight and Luk, 1994; Kwong, 1998; Nastase and Szpakowicz, 2001; Ruiz-Casado et al., 2005), the evaluation seemed to be done manually by a single annotator or was not specified.

Much recent work has focused on having multiple annotators perform this task to better gauge levels of agreement (Meyer and Gurevych, 2011; Navigli, 2006; Navigli and Ponzetto, 2010; Niemann and Gurevych, 2011). To take one recent example, Meyer and Gurevych (2011) aligned WordNet and Wiktionary and asked for judgments on 2,423 sense pairs about whether the senses have the same meaning or a different meaning. The approach of the present study, on the other hand, allows for some graded notion of

senses (i.e., $4*5$). Thus, a total of possible alignments between five WordNet senses and five Wiktionary senses is 25.

meaning, i.e., a related meaning category. This is in line with what Meyer and Gurevych (2011) noted in their error analysis about mis-alignment, where one often wanted to link senses with related meanings: “Future work could distinguish between sense alignments sharing the same meaning and sharing a highly related meaning” (Meyer and Gurevych (2011), p.7).

Similar in spirit to the present work is research comparing judgments on word senses across different contexts. Notably, Erk and McCarthy (2009) and Erk et al. (2009) explored graded word sense judgments by allowing annotators to select the degree of similarity for a word sense on a given task, not just restricting the task to selecting a single sense. Erk et al. (2009), for instance, performed two experiments. In the first, WSSim (Word Sense Similarity), they asked annotators to read sentences and, for every WordNet sense, assign a similarity score, between 1 (completely different) and 5 (identical). This allowed annotators to grade all senses, instead of making a binary choice for each sense or even selecting a single sense. In the second experiment, USim (Usage Similarity), annotators were given pairs of sentences and asked to rank how similar in meaning the two usages of a given word were (using the same 5-point scale). What they found was that “[t]he annotators made use of the full spectrum of ratings.” (Erk et al. (2009), p.17) The experiment of the present study is similar in spirit and design to the USim, in that annotators are asked to compare two potentially distinct usages (in the present study case, dictionary definitions) and rate how similar they are.

Although other databases exist, predominantly for ones linking WordNet and Wikipedia or Wiktionary (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011; Wolf and Gurevych, 2010; Fernando and Stevenson, 2010; Toral et al., 2009), no

database exists for the types of inventories the present study is interested in. Thus, though it is small, this research tries to develop a database (i.e., a gold standard of alignments) between the inventories of interest (i.e., WordNet and COBUILD).

4.2.4 Comparison with previous research

There are several differences between the WSA approach of the present study and much of the other alignment approaches. First, the present system does not simply combine two (or more) inventories to expand a resource and make it bigger as other studies did (e.g. Ide & Veronis, 1990; Knight & Luk, 1994; Kwong, 1998; Meyer & Gurevych, 2011; Navigli & Ponzetto, 2010; Niemann & Gurevych, 2011). Instead, the present system uses two inventories for their own purpose respectively (i.e., WordNet for WSD and COBUILD for providing lexical information) and consistently map from WordNet to COBUILD in order to automatically extract appropriate information from one of the resources (i.e., COBUILD). Second, the sources of information for WSA are different between WSA of this study and other WSA works; this study uses a WSD classifier to draw information for alignment from the inventory (WordNet) whereas other alignment works draw information directly from the inventories (e.g., Ide & Veronis, 1990; Knight & Luk, 1994; Kwong, 1998; Meyer & Gurevych, 2011; Nastase & Szpakowicz, 2001; Navigli, 2006; Navigli & Ponzetto, 2010; Niemann & Gurevych, 2011; Ruiz-Casado et al, 2005; Toral et al., 2009). This brings the third difference that these alignment works have used much information extracted from inventories (e.g. sense definitions, semantic relations, semantic hierarchy, textual entailment, domains, etc.). On the other hand, since the state-of-the-art WSD classifier employed in this study already

uses all necessary information from WordNet, COBUILD example sentences are the only information to process the alignment work. This makes the WSA work of this study fairly simple but robust.

4.3 Sense Inventories

As addressed earlier, two sense inventories are employed in this study; WordNet (4.3.1) and COBUILD (4.3.2). WordNet is used for WSD while COBUILD is used for providing lexical information for language learners.

4.3.1 WordNet

All natural language processing (NLP) modules as well as WSD systems utilized in this research use WordNet. WordNet (Fellbaum, 1998) is a lexical knowledge resource that has been most widely used in the NLP field. The unique feature of WN is that every sense of words in WN is represented in synsets (i.e., a set of synonymous words), which are constructed for NLP systems to clearly distinguish between different senses of a word (Fellbaum, 1998). For example, the synsets *{job, employment, work}* and *{workplace, work}* make a clear distinction between two senses of the noun ‘work’. Each synset comes with a gloss (i.e., a textual definition), often followed by short usage examples.

In addition to synsets, semantic relations such as antonymy (opposite), hyponymy (subsets), hypernymy (superset), meronymy (parts), holonymy (whole) and entailment are defined and those also often help the WSD systems to determine a proper sense of the word in a particular context. WordNet has an extensive coverage of the English language

containing more than 144,600 words (e.g. WordNet 3.0 has 155,000 words) and has been most widely employed in WSD models.

4.3.2 English language learners' dictionary

Since the system specifically targets learners of English as a second language (ESL), the lexical information should be appropriate to that type of learners. Examples constructed by lexicographers for learners' dictionaries typically control for syntactic and lexical complexity (Segler et al., 2002), in addition to containing naturalistic examples. In this respect, these kinds of examples are extracted from the Collins COBUILD Student's Dictionary (Sinclair, 2006), as it is widely used by ESL learners.

The content in the Collins COBUILD Student's Dictionary (COBUILD) is based on actual English usage, derived from analysis of a large corpus of written and spoken English, thereby providing a large number of authentic sentential examples taken from the corpus (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. Those practicalities are a major consideration for employing COBUILD in providing lexical information by the system of this research.

4.4 Word sense alignment (WSA)

As is well known, different word sense inventories contain non-trivial mappings between them. As one example, Palmer et al. (2000) discussed various problems in aligning the senses of *shake* in Hector (Atkins, 1993) with those in WordNet (Fellbaum,

1998), such as the *TREMBLE* and *MOVE* distinctions in Hector being conflated in WordNet. In addition to the fact that defining a mapping between individual senses is non-trivial, automatic sense alignment between different sense inventories raises a more general issue; the number of senses can make it nearly intractable to sort through all possible alignments to find the best one. In other words, with m senses from one sense inventory and n senses from another sense inventory, there are $m \times n$ possible pairs and an alignment structure could consist of any combination of these pairs unless some restrictions are applied. The number of possible alignments is, thus, the size of the power set of the pairs, 2^{mn} . For example, if there are two senses in an inventory A and two senses in an inventory B for a given word, all possible pairs between the senses of A and B are four and the size of their power set (= possible alignments) is 2^4 , as shown in (1):

- (1) $A = \{a1, a2\}, B = \{b1, b2\}$
 Pairs = $\{(a1, b1), (a1, b2), (a2, b1), (a2, b2)\}$
 Power set = $\{\{\}, \{(a1, b1)\}, \{(a1, b2)\}, \{(a2, b1)\}, \{(a2, b2)\}, \{(a1, b1), (a1, b2)\}, \{(a1, b1), (a2, b1)\}, \{(a1, b1), (a2, b2)\}, \{(a1, b2), (a2, b1)\}, \{(a1, b2), (a2, b2)\}, \{(a2, b1), (a2, b2)\}, \{(a1, b1), (a1, b2), (a2, b1)\}, \{(a1, b1), (a1, b2), (a2, b2)\}, \{(a1, b1), (a2, b1), (a2, b2)\}, \{(a1, b2), (a2, b1), (a2, b2)\}, \{(a1, b1), (a1, b2), (a2, b1), (a2, b2)\}\}$

In the power set, an empty set element is barely acceptable in real alignments between two inventories; while theoretically possible, an alignment that has no mapping at all between two inventories for a given word hardly exist. Also, intuitively the alignment, $\{(a1, b1), (a1, b2), (a2, b1), (a2, b2)\}$ is not quite proper either, because the set means that two distinct meanings in one sense inventory are basically the same (e.g., $a1$ and $a2$ in A are the same). Even though those two sets are excluded from all possible alignment

structures, it still can be fairly infeasible to track down all possible alignments as the number of senses becomes larger. If there are five senses mapping to five senses, then there are $2^{25} = 33,554,432$ possible alignment structures.

This had been addressed in the previous studies (Ide and Veronis, 1990; Knight and Luk, 1994; Kwong, 1998; Nastase and Szpakowicz, 2001). As a way to resolve the issue, many works on sense alignment were often limited in scope (e.g., number of words) or failed to handle a great deal of one-to-many mappings in such alignments (Ide and Veronis, 1990; Knight and Luk, 1994; Kwong, 1998; Nastase and Szpakowicz, 2001). Meyer and Gurevyvh (2011) recently tried to handle many-to-many mappings. They handled them by first narrowing down what the possible candidates were and then by only considering decisions at the level of an individual sense. In this way, only 25 sense pairs are handled in the above case with five senses mapping to five senses. As one can note here, it should be more efficient to find a best alignment from the possible alignment candidates excluding improper alignments from the beginning (e.g., 50 possible alignments), rather than finding a best one from all possible alignments (e.g., $2^{25} = 33,554,432$). In that sense, Meyer and Gurevyvh's (2011) approach can be noteworthy. However, as stated in the previous sections (4.1 and 4.2.4), how one aligns the senses is often application-specific. While Meyer and Gurevyvh (2011) performed their WSA at the level of an individual sense, WSA in the current study is performed at the level of alignment structure.

Indeed, the present study is the first to focus specifically on trends in alignment structures between two inventories. As a way to perform the alignment between WordNet senses (for WSD use) and COBUILD senses (for providing information) at the level of an

alignment structure, WSA in this study is performed on the outputs of WSD, thereby building from state-of-the-art systems (sections 4.4.2 to 4.4.5). Also, an assumption specific to the context of the current study is made (section 4.4.1).

4.4.1 Application-specific assumption

To present a learner with sense-specific example sentences, the system maps the output of a WordNet-based WSD system to a database of sense-organized COBUILD sentences. Then what if the same WordNet sense maps to multiple COBUILD senses? This would mean that the system may have to take the WordNet sense disambiguated by the WSD classifier and map it to multiple sets of COBUILD senses. While this may be feasible, a simpler solution would be to restrict the alignment such that each WordNet sense is allowed to map to only one COBUILD sense, thereby providing a coherent set of examples to display.

Considering that COBUILD senses have been distinguished so as to assist learners, to conflate two categories might conflict with that motivation. For example, in (2), *community* has a WordNet sense which can be argued to overlap with two COBUILD senses ((3) and (4)). But the current approach is to keep these COBUILD senses separate and map the WordNet sense only to the better of these two, as in (3).

(2) **community.wn1**: a group of people living in a particular local area (“the team is drawn from all parts of the community”)

(3) **community.cb1**: The community is all the people who live in a particular area or place (“He’s well liked by people in the community”)

(4) **community.cb2**: A particular community is a group of people who are similar in some way (“The police haven’t really done anything for the black community in particular”)

Thus, the current method adopts the assumption that for any given WordNet sense, it can only map to one COBUILD sense, resulting in m -to-1 mappings and not m -to- n . For example, with three WordNet senses and two COBUILD senses, each WordNet sense maps to one COBUILD sense by this assumption, so an alignment always has three pairs. Then, each of the three senses can map to two different possibilities in COBUILD, giving 2^3 possible alignments. As seen in section 4.4.6 (Sense alignment algorithm), this assumption makes it straightforward to develop an alignment algorithm.

4.4.2 Initial alignment

Let us define an alignment A as, in the case of the study, a set of pairings of elements from the sets of WordNet and COBUILD senses, (WN, CB) , for a given target word. An alignment is a bidirectional mapping between the two inventories; $A : WN \times CB$. In general, the mapping may be as dense or sparse as required by the linguistic definitions; for the purposes of the present study, however, the number of possible alignments is reduced by the assumption in which each WordNet sense is allowed to map to only one COBUILD sense, as stated in the previous section (4.4.1 Application-specific assumptions). With an alignment defined as a set of pairs, finding a best alignment is initially established as in (5).

$$\begin{aligned}
(5) \quad & \max_{A \in \text{Approp}} p(A) \\
& = \max_{A \in \text{Approp}} p(\text{wn}, \text{cb}) \\
& = \max_{A \in \text{Approp}} \prod_{(i,j) \in A} p(\text{wn}_i | \text{cb}_j) p(\text{cb}_j)
\end{aligned}$$

As shown in (5), an alignment that has the maximum probability is selected as the best alignment among all candidate alignments in *Approp*. *Approp* represents a set of all possible alignments meeting the assumption (see section 4.4.1). For a given alignment *A*, $p(\text{wn}, \text{cb})$ is the probability of the set of links shown in Figure 4.5.

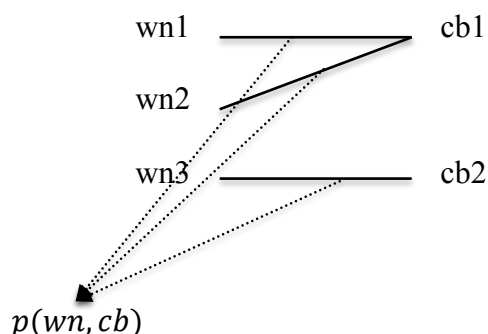


Figure 4.5. $p(\text{wn}, \text{cb})$ for a given alignment

As one can see in formula (5) and in Figure 4.5, $p(\text{wn}, \text{cb})$ is $\prod_{(i,j) \in A} p(\text{wn}_i | \text{cb}_j) p(\text{cb}_j)$, where $p(\text{cb}_j)$ is uniform⁹. Then, in order to compute $\prod_{(i,j) \in A} p(\text{wn}_i | \text{cb}_j)$, the system first has to set up WordNet sense probability distributions for every sense of COBUILD. To initialize the distribution, the system first takes the set of example sentences for each COBUILD sense cb_j and runs a WSD classifier over the COBUILD examples, obtaining probabilities for each WordNet sense

⁹ $p(\text{cb}_j)$ can be ignored in calculation, because it is the same (i.e., uniform) for all wn-cb links per cb sense and consequently it does not affect the probability, $p(\text{wn}, \text{cb})$.

as shown in (6). Given that there are multiple example sentences for each sense cb_j , the algorithm averages over all the WordNet probabilities to obtain a likelihood of the WordNet senses, given each of the COBUILD senses, $p(wn_i|cb_j)$.

(6) 'require.cb₁' (3 example sentences)

'if you **require** further information you should consult the registrar '

$$wn_i = [(wn_1, 0.36), (wn_2, 0.14), (wn_3, 0.23), (wn_4, 0.26)]$$

'this isn't the kind of crisis that **requires** us to drop everything else'

$$wn_i = [(wn_1, 0.40), (wn_2, 0.14), (wn_3, 0.20), (wn_4, 0.26)]$$

'some of the materials **required** for this technique may be difficult to obtain '

$$wn_i = [(wn_1, 0.31), (wn_2, 0.09), (wn_3, 0.14), (wn_4, 0.46)]$$

$$\rightarrow p(wn_i|cb_1) = \{(1, 1): 0.36, (2, 1): 0.12, (3, 1): 0.19, (4, 1): 0.33\}$$

'require.cb₂' (4 example sentences)

'the rules also require employers to provide safety training'

$$wn_i = [(wn_1, 0.41), (wn_2, 0.14), (wn_3, 0.23), (wn_4, 0.23)]$$

'at least 35 manufacturers have flouted a law requiring prompt reporting of such malfunctions'

$$wn_i = [(wn_1, 0.40), (wn_2, 0.15), (wn_3, 0.21), (wn_4, 0.25)]$$

'the law now requires that parents serve on the committees that plan and evaluate school programs'

$$wn_i = [(wn_1, 0.41), (wn_2, 0.15), (wn_3, 0.21), (wn_4, 0.23)]$$

'then he'll know exactly what's required of him'

$$wn_i = [(wn_1, 0.44), (wn_2, 0.17), (wn_3, 0.16), (wn_4, 0.23)]$$

$$\rightarrow p(wn_i|cb_2) = \{(1, 2): 0.42, (2, 2): 0.15, (3, 2): 0.20, (4, 2): 0.23\}$$

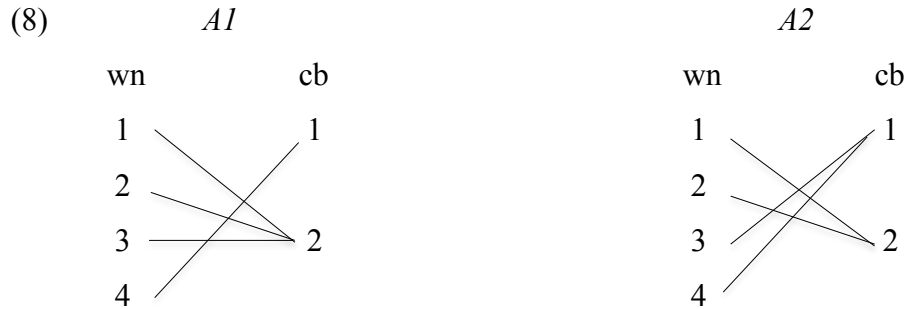
As shown in (6), for example, with a word *require* having two senses of COBUILD and four senses of WordNet, the system takes all given example sentences (e.g., three) for COBUILD sense #1 (*require.cb₁*) and four examples for COBUILD sense #2 (*require.cb₂*). Then the probabilities for every WordNet sense over the COBUILD examples sentences are obtained by running the WSD classifier over them. For a sense of *require* in a given example sentence (*if you **require** further information you should consult the registrar*), the WSD classifier generates a WordNet sense probability distribution as $wn_i = [(wn_1, 0.36), (wn_2, 0.14), (wn_3, 0.23), (wn_4, 0.26)]$. According to this, the probability of WordNet sense #1 (=wn₁) is ‘0.36’, WordNet sense #2 (=wn₂) is ‘0.14’, and so on. Since each COBUILD sense has multiple example sentences, the system averages over all the WordNet probabilities. For the given COBUILD sense #1, all the WordNet probabilities give $p(wn_i|cb_1) = \{(1, 1): 0.36, (2, 1): 0.12, (3, 1): 0.19, (4, 1): 0.33\}$, where the first number in a parenthesis stands for a WordNet sense number and the second is a COBUILD sense number, for example, ‘0.36’ comes from an average of probabilities of WordNet #1 (wn₁) over three example sentences that belong to COBUILD sense #1 (= (0.36+0.40+0.31)/3).

Based on the results in (6) (e.g., $p(wn_i|cb_1)$ and $p(wn_i|cb_2)$), a probability distribution of every WordNet sense for each of the COBUILD senses, $p(wn_i|cb_j)$, is set up as shown in (7).

(7)

$p(wn_i cb_j)$	wn_1	wn_2	wn_3	wn_4
cb_1	0.36	0.12	0.19	0.33
cb_2	0.42	0.15	0.20	0.23

With the information of WordNet sense probability distributions for every sense of COBUILD in (7), let us compute $\prod_{(i,j) \in A} p(wn_i|cb_j)$ for two of the possible alignments, for instance, shown in (8).



$$\begin{aligned} A1: \prod_{(i,j) \in A} p(wn_i|cb_j) &= (wn_1,cb_2) * (wn_2,cb_2) * (wn_3,cb_2) * (wn_4,cb_1) \\ &= 0.42 * 0.15 * 0.20 * 0.33 \\ &= 0.0042 \end{aligned}$$

$$\begin{aligned} A2: \prod_{(i,j) \in A} p(wn_i|cb_j) &= (wn_1,cb_2) * (wn_2,cb_2) * (wn_3,cb_1) * (wn_4,cb_1) \\ &= 0.42 * 0.15 * 0.19 * 0.33 \\ &= 0.0039 \end{aligned}$$

For the alignment *A1* in (8), $\prod_{(i,j) \in A} p(wn_i|cb_j)$ is *0.0042*, and the alignment *A2*, *0.0039*. As a result, based on information from the WSD output, the alignment *A1* is

decided as the best alignment. However, there are empirical evidences that supports a need for this initial alignment to be adjusted, which derives a heuristic of this study discussed in the following section.

4.4.3 The heuristic of the study: adding flatness

Intuitively, if two inventories have the same number of senses and the same granularity of sense distinctions, it should have a one-to-one mapping. If, on the other hand, there are different numbers of senses in the inventories, one might assume that each sense is equally complex. Thus, one could expect that the senses in both inventories would, more often than not, get mapped evenly. Furthermore, given the state-of-the-art, it is clear that WSD output is going to be noisy, so an alignment based on WSD is going to have incorrect links. The question is then *which links will be incorrect?* One hypothesis to consider is that a sense with too many links likely has at least one of them wrong. In the present study, this means that the COBUILD sense with too many WordNet links is likely to have some of them wrong.¹⁰ For instance, for a word with six WordNet senses and three COBUILD senses, if one of COBUILD senses has four links mapped from WordNet senses, some links may be wrong.

Looking ahead to the dataset obtained in this study (see section 4.5.1)¹¹, one can see this to generally be the case. From the initial alignments for nine words¹² (i.e., *area, community, indicate, involve, job, policy, process, require, section*), all COBUILD senses

¹⁰ The other way (i.e., WordNet sense have more than one links to CB sense) cannot be acceptable, given the assumptions (see section 4.4.1.).

¹¹ This analysis was actually carried out after performing WSA experiments.

¹² These nine words were selected to evaluate WSA system (see section 4.5.1.2)

which have more links from the WordNet senses than the average¹³ WordNet links per COBUILD sense were compared with human mapping scores (section 4.5), as shown in Table 4.1.

Word & # of senses	Average	Initial alignments based on WSD (SR::AW)		Human scores for mapping between c-w
		Alignment	COBUILD(c) having more links than average	
area (w6, c6)	1	(0,1,1,1,1, 2)	c4 – w2, w6	(c4,w2) -0.455 (c4,w6) -0.143
community (w6, c3)	2	(1,2, 3)	c2 – w2, w3, w4	(c2,w2) 0.000 (c2,w3) 0.636 (c2,w4) 0.286
involve (w7, c5)	2	(1,1,1,2,2)		
policy (w3,c3)	1	(1,1,1)		
section (w14, c3)	5	(0,5, 9)	c1 - w2, w3, w5, w7, w8, w9, w12, w13, w14	(c1,w2) 0.227 (c1,w3) 0.643 (c1,w5) -0.063 (c1,w7) 0.000 (c1,w8) -0.154 (c1,w9) 0.591 (c1,w1) 0.625 (c1,w13) 0.143 (c1,w14) -0.111
indicate (w5, c6)	1	(0,0,0,1, 2,2)	c2 – w1, w4 c6 – w3, w5	(c2,w1) -0.071 (c2,w4) -0.182 (c6,w3) -0.036 (c6,w5) -0.429
job (w12, c5)	3	(1,2,2,3, 4)	c5 – w1, w8, w9, w10	(c5,w1) -0.667 (c5,w8) -0.833 (c5,w9) -0.875 (c5,w10) -0.857

¹³ The average of links is rounded up from the actual average (e.g., average = 1.4 → 2), so as to more properly count what it means to have “too many” links.

process (w6, c2)	3	(2,4)	c1 – w1, w2, w3, w4	(c1,w1) 1.00 (c1,w2) -0.143 (c1,w3) -0.5 (c1,w4) -0.231
require (w4, c2)	2	(1,3)	c2 – w1, w2, w3	(c2,w1) 0.083 (c2,w2) 0.875 (c2,w3) 0.786

* incorrect alignments by human judgments are noted in bold

Table 4.1. Comparisons between initial alignment and human mapping, showing the COBUILD (c) senses which have more WordNet (w) links than the average

As shown in Table 4.1, eight COBUILD senses have more WordNet links than the average WordNet links per COBUILD sense (*area-c4*, *community-c2*, *section-c1*, *indicate-c2* and *c6*, *job-c5*, *process-c1*, *require-c2*). Comparing them to human mappings, most links between WordNet and COBUILD show negative scores, where higher scores indicate greater confidence in the link (see section 4.5.2). In seven out of eight cases, at least one link is clearly wrong, and in the eighth case (*require*), one of the links is extremely weak (c2, w1), with a score of 0.083. This indicates those links are wrong and need to be revised (i.e., those WordNet senses need to be mapped to other COBUILD senses). This evidence confirms that when there are COBUILD senses having more links from WordNet than the average WordNet links per COBUILD sense, some of those links are highly likely to be wrong and need to be revised. This may suggest that links should be revised towards flatness, because a weak link to a “popular” COBUILD sense needs to be linked elsewhere, because one of the links from the COBUILD sense which has more links may be re-mapped to a COBUILD sense which has less links.

Additionally, the alignment structures of nine words judged by humans (section 4.5) can be analyzed in terms of their flatness, as shown in Table 4.2. For the analysis, how

much each of alignment is deviated from the flattest is first measured. Then the flatness between the system-initial alignment and human alignment is compared.

word	Initial alignment		flatness	Human alignment	
	type	deviation		type	deviation
area	(0,1,1,1,1,2)	$\sigma=0.577$	=	(0,1,1,1,1,2)	$\sigma=0.577$
community	(1,2,3)	$\sigma=0.817$	=	(1,2,3)	$\sigma=0.817$
involve	(1,1,1,2,2)	$\sigma=0.490$	>	(0,1,1,2,3)	$\sigma=1.020$
policy	(1,1,1)	$\sigma=0.0$	>	(0,1,2)	$\sigma=0.817$
section	(0,5,9)	$\sigma=3.682$	>	(1,3,10)	$\sigma=3.859$
indicate	(0,0,0,1,2,2)	$\sigma=0.898$	<	(0,0,1,1,1,2)	$\sigma=0.687$
job	(1,2,2,3,4)	$\sigma=1.020$	<	(2,2,2,3,3)	$\sigma=0.490$
process	(2,4)	$\sigma=1.0$	<	(3,3)	$\sigma=0.0$
require	(1,3)	$\sigma=1.0$	<	(2,2)	$\sigma=0.0$

*deviation shows how far off an alignment is from the flattest alignment

Table 4.2. Comparisons of flatness between initial alignment and human alignment

As shown in Table 4.2, the system-initial alignment and the human alignment show similar flatness for two words (e.g., *area*, *community*) in terms of alignment structure, while system-initial alignment and the human alignment show differences for the rest of the seven words. For the seven words, although it is a small difference (i.e., three words vs. four words out of the nine words), the human alignment shows more flatter alignment structures (e.g., *indicate*, *job*, *process*, *require*). Furthermore, looking just at the human alignments, one can see that the deviation is generally less than 1, with exception of

section and more minor exception of *involve*. In other words, although the distributions are not entirely flat, they are not very skewed either. This analysis suggests that the flat alignment structure would be more likely to be the correct alignment.

From these pieces of empirical evidence, this research found that the system-initial alignments from WSD output may need to be revised toward flatter structures. Accordingly, formula (5) needs to be updated with a new quantity, $p(A_s)$, which reflects the flatness of the alignment structure, as in (9).

$$\begin{aligned}
 (9) \quad & \max_{A \in \text{Approp}} p(A) \\
 & = \max_{A \in \text{Approp}} p(A_s) p(\text{wn}, \text{cb}) \\
 & = \max_{A \in \text{Approp}} p(A_s) \prod_{(i,j) \in A} p(\text{wn}_i | \text{cb}_j)
 \end{aligned}$$

As shown in (9), the maximal probability is computed by $p(\text{wn}, \text{cb})$ and $p(A_s)$, where, $p(\text{wn}, \text{cb})$ is the probability of the set of links and $p(A_s)$ is the probability of the alignment structure. These can be seen explicitly in Figure 4.6.

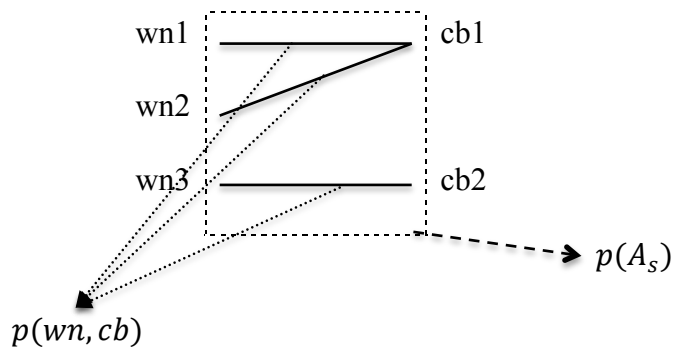


Figure 4.6. $p(\text{wn}, \text{cb})$ and $p(A_s)$ for a given alignment

The value of $p(A_s)$ cannot be calculated directly, as this would require knowing the prior probability of alignment structures. Based on the empirical evidences discussed earlier, an idea of estimating $p(A_s)$ starts from the following heuristic-based mechanism:

Maximize 'flat' alignment

Minimize 'skewed' alignment

The mechanism is designed to give more weight to flat alignments and less weight to skewed alignments; alignments which are flatter have higher values. How to estimate $p(A_s)$ is discussed in the following section.

4.4.4 Calculating probability of alignment structures

According to the heuristic of this study (4.4.3), the results in (8) (i.e., the more skewed alignment, $A1$ have higher probability than the flatter alignment, $A2$) is the reverse. In that case, the probability of alignment structure, $p(A_s)$ would be useful in getting the results such that the flatter alignment gets a higher value, assigning this is correct.

Accordingly, in the present approach, $p(A_s)$ is estimated by measuring how far off a particular alignment is from a flat alignment. What the method estimates is not strictly an empirical probability distribution, in that it does not derive from the measurement of a random variable. It is a metric; as it meets certain formal properties (e.g., a value for each alignment structure is between 0 and 1 and the sum of all possible alignment structures is 1), this method is used to approximate the empirical probability distribution.

Then, how is the *flatness* of the alignment structure defined? With the number of WordNet and COBUILD senses, the total number of links is equal to the number of WN senses by the assumption that each WordNet sense maps to only one COBUILD sense

(see section 4.4.1). So the approach calculates the mean number of links (μ) each COBUILD sense has. If there are three COBUILD senses mapping to five WordNet senses, for example, the mean number of links is $5/3$; on average each COBUILD sense has 1.67 links.

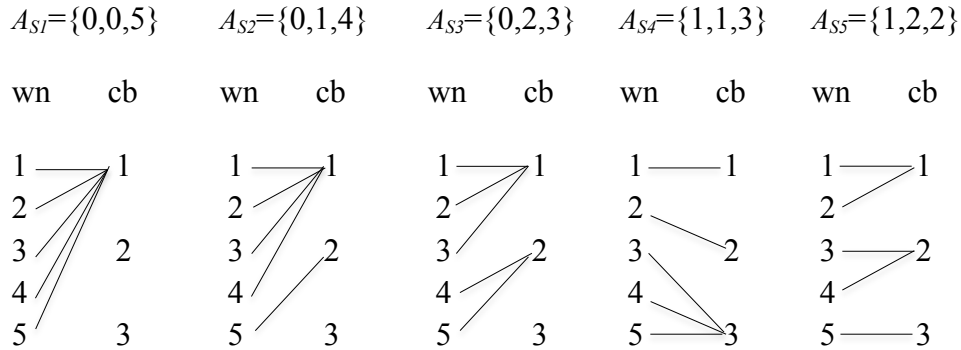


Figure 4.7. All possible *alignment structure types* given five *wn* senses and three *cb* senses and one example for each type

To be flatter, the alignment should be closer to this mean. In the next step, then, the algorithm finds all possible *alignment structure types* (i.e., the distribution of the number of links for each COBUILD sense). To continue the example, the system has the following alignment types: $\{0,0,5\}$, $\{0,1,4\}$, $\{0,2,3\}$, $\{1,1,3\}$, and $\{1,2,2\}$ as shown in Figure 4.7. Taking $\{1,1,3\}$, for instance, one *cb* sense has 1 link, another has one link, and a third has three links.

For each type, it does not matter which COBUILD sense has a particular number of links. Thus, $\{1,1,3\}$, $\{1,3,1\}$, or $\{3,1,1\}$ are, for example, all the same in terms of a structure type as shown in Figure 4.8. Likewise, those shown in Figure 4.7 are one representative example for each type.

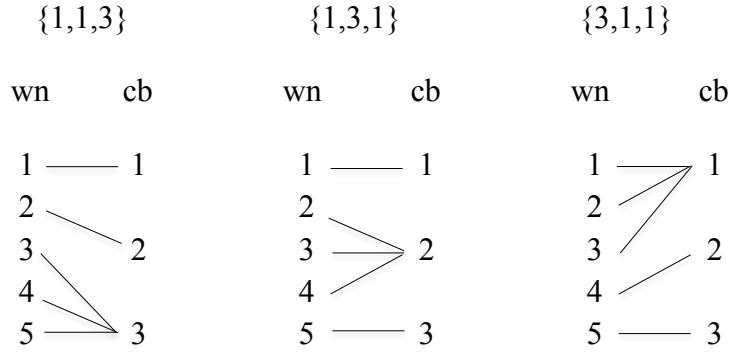


Figure 4.8. The same alignment structure type A_{S4}

To measure how far off each alignment type is, the deviation from the flattest alignment is estimated for each alignment type¹⁴. For $\{1,1,3\}$, for example, the deviation from the flattest is 1.26 , and the most deviated from the flattest (i.e., skewed distribution), $\{0,0,5\}$, has the deviation of 3.16 . To obtain probabilities, the system first converts these deviations to *counts*, as in (9), where each deviation is subtracted from the highest deviation (i.e., most skewed alignment (e.g., $A_{S4}=\{0,0,5\}$)). The reason for subtracting each deviation from the most skewed one is that the most skewed alignment has the highest deviation, which has to be least weighted in terms of the flatness. Formula (9) applies when there are two or more COBUILD senses. When there is only one, $c(A_{Si}) = c(A_{S1}) = 1$. The λ term allows the system to assign some non-zero count to A_{S1} . For simplicity, in the present study, $\lambda = 0$ is first used for the experiments, which rules out the most skewed alignment. Then $\lambda = 1$ is tried to assign non-zero count to the most skewed alignment.

¹⁴ standard deviations (σ) are simply used for estimating the deviation.

$$(9) c(A_{S_i}) = \sigma(A_{S_1}) + \lambda - \sigma(A_{S_i})$$

$$(10) p(A_{S_i}) \approx \frac{c(A_{S_i})}{\sum_j c(A_{S_j})}$$

The probability is then estimated in the usual way, as in (10). Values for each different alignment type in the example are given in Table 4.3.

		$\lambda = 1$		$\lambda = 0$	
$A_{S1} = \{0,0,5\}$	$\sigma(A_{S1})=3.16$	$c(A_{S1})=1.00$	$p(A_{S1})=0.08$	$c(A_{S1})=0.00$	$p(A_{S1})=0.00$
$A_{S2} = \{0,1,4\}$	$\sigma(A_{S2})=2.28$	$c(A_{S2})=1.88$	$p(A_{S2})=0.16$	$c(A_{S2})=0.88$	$p(A_{S2})=0.13$
$A_{S3} = \{0,2,3\}$	$\sigma(A_{S3})=1.67$	$c(A_{S3})=2.49$	$p(A_{S3})=0.21$	$c(A_{S3})=1.49$	$p(A_{S3})=0.22$
$A_{S4} = \{1,1,3\}$	$\sigma(A_{S4})=1.26$	$c(A_{S4})=2.90$	$p(A_{S4})=0.25$	$c(A_{S4})=1.90$	$p(A_{S4})=0.28$
$A_{S5} = \{1,2,2\}$	$\sigma(A_{S5})=0.63$	$c(A_{S5})=3.53$	$p(A_{S5})=0.30$	$c(A_{S5})=2.53$	$p(A_{S5})=0.37$

Table 4.3. The various values used in calculating probabilities for alignment types

Continuing from the example in (8) for computing $\max_{A \in \text{Approp}} p(A_S) p(\text{wn}, \text{cb})$, $p(A_S)$ is now estimated as shown in (11) and added to the quantity of $\prod_{(i,j) \in A} p(\text{wn}_i | \text{cb}_j)$ calculated in (9), shown in (12).

(11) With four senses of WordNet and two senses of COBUILD,

a. WordNet links per COBUILD sense $\rightarrow 2.0$

b. alignment type $\rightarrow [(0,4), (1,3), (2,2)]$, where the two datapoints in each

parenthesis stands for COBUILD senses, and the number for each datapoint is

the number of WordNet senses linked to the COBUILD sense (e.g., (1,3)

means one COBUILD sense has one WordNet sense mapped and one

COBUILD sense has three WordNet senses mapped)

c. Score for each alignment type

c.1. measure deviations for each alignment type

$$= \{(1,3): 1.0, (2,2): 0.0, (0,4): 2.0\} \rightarrow \text{most deviated} = 2.0$$

c.2. put more weights on the less deviated type from the flattest

(i.e., highest deviation – each deviation)

$$= \{(1,3): 1.0, (2,2): 2.0, (0,4): 0.0\}$$

\rightarrow sum of scores = 3.0

e. Estimate proportion of each alignment type (i.e., normalization (=score/sum))

$$= \{(1,3): 0.33, (2,2): 0.67, (0,4): 0.0\}$$

$$\rightarrow p(A_{si}) = \{(1,3): 0.33, (2,2): 0.67, (0,4): 0.0\}$$

According to $p(A_{si})$ in (11), the alignment type of $A1$ and $A2$ in (8) is (1,3) and (2,2) respectively and thus $p(A_{s1})$ is 0.33 and $p(A_{s2})$ is 0.67. With $p(A_{si})$, the computation of $\max_{A \in A_{prop}} p(A_s) p(w_n, c_b)$ is shown in (12), thereby showing the best alignment between the two possible alignments is $A2$ with a higher probability.

$$(12) A1: p(A_s) \prod_{(i,j) \in A} p(w_n_i | c_b_j)$$

$$= 0.33 * 0.0042$$

$$= 0.0014$$

$$A2: p(A_s) \prod_{(i,j) \in A} p(w_n_i | c_b_j)$$

$$= 0.67 * 0.0039$$

$$= 0.0026$$

Clearly, this is a heuristic-based way of calculating alignment probabilities and can be adjusted in the future. However, as one can note from the results (8) to (12), it does reward flatter alignments and penalize more skewed ones. The actual python code for computing $p(A_{si})$ is in Appendix B.

4.4.5 Adjusting WSD output

As shown in the previous sections, WSA is performed on the WSD outputs by applying the probability of alignment structures $p(A_s)$. This allows the system to revise links based on general patterns on alignment structures. That is, the links are revised by the quantity, $p(A_s)$. For example, the best alignment for *require.v* (when $p(A_s)$ is applied) is given in the right side of Figure 4.9, which is a flatter alignment and an alignment structure based on a WSD system output (left) provides a more skewed alignment. As one can note, the links are revised from the structure based on WSD output (treating $p(A_s)$ as a uniform) to the structure after the WSA algorithm is applied (estimating $p(A_s)$), thereby becoming flatter.

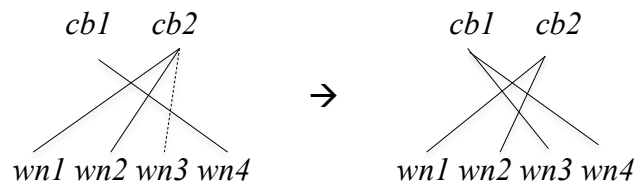


Figure 4.9. WSA based on WSD system output (left) and WSA when the WSA algorithm applied (right) for alignments for *require.v* (dashed line = revised link)

As shown in the example, the present approach biases alignments to be less skewed than they would be by only using $p(wn_i|cb_j)$. As addressed in section 4.4.3, favoring

flatter structures seems like a useful heuristic. Even when the alignment is not flat, such a re-consideration of the WSD output can be helpful. Consider Figure 4.10, where the correct alignment (left) is somewhat skewed, with COBUILD sense 1 aligning to three WordNet senses, while sense 2 aligns to only one. Interestingly, the WSD system output (middle) also has a skewed alignment, but for the *wrong* senses: sense 2 now has three links. links.

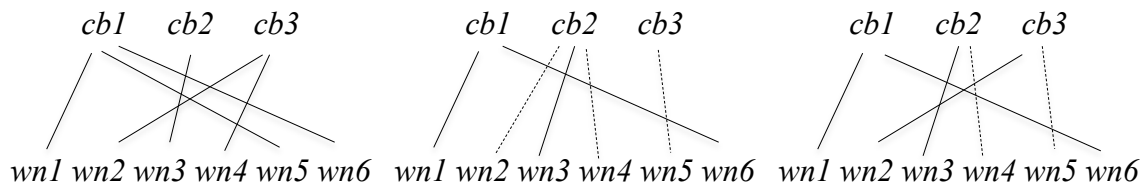


Figure 4.10. Gold standard (left), initial alignment based on WSD system output(middle), and adjusted WSA applied with the flatness (right) for alignments for *community.n*

If the skewed links are incorrect, the hope is that system confidence for them is lower, and they are thus more amenable to adjustment. In this case, for instance, the link between COBUILD sense 2 and WordNet sense 2 is not very strong and by encouraging a flatter alignment (see section 4.3.2), COBUILD sense 3 is correctly re-aligned to this WordNet sense. This corrective measure works best, then, when: a) senses with many links are likely to have some be wrong; b) only minor adjustments are made (i.e., there needs to be good evidence to overturn an link); and c) correct links have competitive probabilities. Indeed, one benefit of examining $p(A_s)$ as a source of information is that it could overcome limitations in the WSD systems by forcing them to reconsider other possibilities.

4.4.6 Sense alignment algorithm

Summing from previous sections (4.4.1 to 4.4.5), with the assumption (section 4.4.1) that each WordNet sense maps to one COBUILD sense, WSA is first performed on the basis of information from WSD outputs (i.e., initial alignments, section 4.4.2). Yet, the empirical evidence shows that the actual alignment structures are more flatted (section 4.4.3). Thus, the initial alignments from WSD outputs are adjusted by the heuristic of this study, which favors flatter alignment structures (section 4.4.4 and section 4.4.5). The algorithm to automatically process WSA has been developed as introduced in the following.

4.4.6.1 Basic WSA algorithm

The basic alignment algorithm works as follows (see the actual code in Appendix C):

1. Obtain probability estimates for each pairing, $p(wn_i, cb_j)$. As mentioned in section 4.4.1, this is done by sense tagging the COBUILD examples to obtain $p(wn_i|cb_j)$.
2. For each WordNet sense, assign its best COBUILD sense for an initial alignment A. Initially, $p(A_s)$ is not applied.

- $p(wn_i, cb_j)$ is used to estimate $p(cb_j|wn_i)$ and rank candidates, because of the equivalence given in (13).

$$\begin{aligned} (13) \max_A \prod_{(i,j) \in A} p(wn_i, cb_j) \\ &= \max_A \prod_{(i,j) \in A} p(wn_i|cb_j) \\ &\propto \max_A \prod_{(i,j) \in A} p(cb_j|wn_i) \end{aligned}$$

- If two COBUILD senses are tied, the system runs steps #3–#5 separate times, thereby considering all possible (tied) candidates.
3. For each pair (wn_i, cb_j) in A , substitute every other COBUILD sense cb_k to obtain a new alignment A' ; calculate its probability, $p(A')$.
 - $p(A')$ is calculated as in (5).
 4. Select the single-best change from step #3, i.e., replace (wn_i, cb_j) with (wn_i, cb_k) , provided that $p(A') \geq p(A)$ and no other replacement provides a greater increase.
 - If the probability is unchanged, the system takes the flatter alignment, i.e., the newer one. (Given that $p(wn_i|cb_j)$ and $p(wn_i|cb_k)$ were originally different, this is unlikely.)
 5. Repeat steps #3 and #4 until no change is made, producing a final revised alignment

If multiple alignments have an equal probability upon completion of the algorithm or its various runs (see note on ties in step #2), the system randomly selects one.

The algorithm above is designed to give us the best set of COBUILD senses, given a set of WN senses, so as to be able to map from WN to COBUILD.

Steps 1 and 2 find the best COBUILD sense for each WN one, ignoring the overall alignment structure (i.e., simply using WSD output); the rest of the steps take the alignment structure ($p(A_s)$) into account.

In steps 3 and 4, the system iteratively finds a better alignment, by checking whether each WN sense could change its link to result in an overall better alignment, thereby also preserving the assumption that a WN sense maps to one COBUILD sense. The substitutions which succeed are the ones which result in a flatter alignment (see section 4.4.3). With the heuristic of giving more weight on the flatter alignment, in step 4, the system verifies that the change is the best change at that iteration and that the overall alignment probability increases.

The system repeats steps 3 and 4 until no change is made, indicating that the probability $p(A')$ does not go up (=maximum) and thus is the final revised alignment (step 5). The system considers every possible change of each wn 's link from one cb to one of other possible cb s at a given iteration, so $p(A_s)$ fluctuates based only on the link patterns. Accordingly, a better $p(wn_i|cb_j)$ probability is preferred over a lower probability with the same type of alignment (e.g., the same flat alignment) to be a better overall alignment. In that way, the system keeps updating the best alignment with the alignment of the maximum probability in steps 3 and 4. Therefore, if a change makes the alignment worse, no further change could improve over what it had been.

4.4.6.2 Complexity

Briefly examining algorithm complexity, let m and n be the number of WordNet and COBUILD senses, respectively, for an alignment. In the first two steps of the algorithm

(section 4.4.2.1), the system looks at an $m \times n$ matrix of senses and picks the m best links, giving $O(mn)$. In steps #3 and #4, the system loops over m sense pairs, substituting no more than $n-1$ COBUILD senses for each one, again $O(mn)$. This is done until no more changes are possible.

Considering that once the system makes a change for a particular WordNet sense's link, the system does not change it again: once (wn_i, cb_j) is replaced by (wn_i, cb_k) , the system does not change the link for wn_i again. If the system were to change the link again, it would be to a different cb_l , which would have either the same alignment with a lower $p(wn_i, cb_l)$ or a less likely alignment structure. Thus, the system should not make more changes than the number of links in the alignment, i.e., m .

Thus, the $O(mn)$ algorithm is run no more than m times, giving an overall complexity of $O(m^2n)$. With fewer than 100 senses in either inventory (and generally much fewer) for a given word, this is minimal. In the worst case, in step #2, every WordNet sense is tied among every COBUILD sense, requiring the algorithm to try all mn combinations in the subsequent steps, giving a worst-case of $O(m^3n^2)$. If in practice there are few ties, the expected case would be closer to $O(m^2n)$.

4.5 Evaluation

In the present study, the upper bound on automatic WSA accuracy was investigated by testing human accuracy. This study has obtained a small set of gold alignments (nine words of varying polysemy¹⁵), based on pooling judgments from semi-experts (i.e. linguistics students and faculty). Section 4.5.1 describes details of how to obtain the evaluation data for this research and the characteristics of the evaluation data are outlined

¹⁵ *area.n, community.n, indicate.v, involve.v, job.n, policy.n, process.v, require.v, section.n*

in section 4.5.2. Section 4.5.2 finally presents how to evaluate the system outputs and discuss its results.

4.5.1 Obtaining the evaluation data

4.5.1.1 Pooling semi-experts

Developing a gold standard with expert annotators can be costly. An alternative for annotation for various NLP tasks is to collect non-expert annotations, i.e., crowdsourcing (Madnani et al., 2011; Wang et al., 2009; Snow et al., 2008). However, the task of assigning sense mappings may be more challenging to most non-experts, especially when they would require deeper linguistic knowledge to do so. Therefore, in this study, the strategy of collecting judgments from *semi-experts*, namely (computational) linguistics faculty and students was pursued (cf. Muhonen and Purtonen, 2011). By surveying linguistics faculty and graduate students to align word senses, people who have at least a basic knowledge of semantics are targeted. No costs are incurred in the present study, as participants are volunteers, and thereby the crowdsourcing problem of obtaining noisy answers (Laws et al., 2011) is mitigated, while at the same time being able to gather a number of annotators for a given sense¹⁶.

One limitation is the smaller potential pool of respondents than with crowdsourcing, and thereby more strictly limiting the amount of data that can be gathered. As an experiment into how resources align, however, the method is straightforward.

¹⁶ More annotators can of course reduce any idiosyncrasies arising from one person's data; see, for example, the discussion in (Erk and McCarthy, 2009).

4.5.1.2 Word selection

The basic words from the Academic Word List (AWL) are considered as a starting point. The AWL consists of 570 word families occurring most frequently over a range of academic texts, namely over 100 times in a 3.5 million word academic corpus¹⁷. These word families are indexed by a particular head word, e.g., *interpret* heads a list containing *interprets*, *interpreter*, *interpretation*, etc. Students who master the AWL thus greatly expand their vocabulary usage.

For the study, the words with at least 3 WordNet (WN) senses are selected, in order to obtain enough complexity to get a grasp on the general properties of alignment. Then three types of words are picked, representing a range of different COBUILD (CB) senses: 1) less senses than in WordNet; 2) (roughly) the same number of senses; and 3) more senses¹⁸. This gives different degrees of alignment *skewedness*, increasing the chances of seeing both zero/null mappings (i.e., where a sense in one inventory maps to nothing in the other) and multiple mappings. Despite being a small set, this break-down to some extent allows the present study to get a handle on word alignment across a diverse set of cases (cf. (Meyer and Gurevych, 2011)), just as Erk and McCarthy (2009) use eight lemmas to evaluate graded word sense assignment¹⁹. The nine selected words are in Table 4.4. In total, there are 63 WordNet and 35 COBUILD senses, incorporating both nouns and verbs (see Appendix D, for definitions and examples of the nine words in WN and CB).

¹⁷ <http://www.victoria.ac.nz>

¹⁸ We only have one instance of this (*indicate.v*); in general, COBUILD is less fine-grained than WordNet.

¹⁹ In the future, one can ensure selection across further criteria, including the so-called Unique Beginner of a word and location within the WordNet taxonomy (Niemann and Gurevych, 2011).

	Word	WN	CB
Balanced	area.n	6	6
	indicate.v	5	6
	policy.n	3	3
Skewed	community.n	6	3
	involve.v	7	5
	job.n	12	4
	process.v	6	2
	require.v	4	2
	section.n	14	3

Table 4.4. Words selected for the present experiment, including number of senses in each inventory

4.5.1.3 Survey design

Taking the 63 WordNet senses, the study consists of seven individual surveys with nine multiple-choice questions each. Each question is a WordNet sense, and the nine different words are distributed across the surveys. The question choices consist of all the COBUILD senses of a word (with examples), as in Figure 4.11. Each question includes examples of the sense; adding examples to the definition helps participants to more readily understand the sense. As shown, there are four options for each choice: *same meaning*, *related meaning*, *no relation*, and *unable to determine*. The last category is important, as it allows one to see how often participants had extreme difficulty in making a decision.

Q3. indicate (v) : to state or express briefly ("He indicated his wishes in a letter")

	Your Answer
1. If one thing indicates another, the first thing shows that the second is true or exists ("A survey of retired people has indicated that most are independent and enjoying life")	<input checked="" type="checkbox"/> Same meaning
2. If you indicate an opinion, an intention, or a fact, you mention it in an indirect way ("U.S. authorities have not yet indicated their monetary policy plans")	<input type="checkbox"/> Related meaning
3. If you indicate something to someone, you show them where it is, especially by pointing to it ("He indicated a chair. 'Sit down.'")	<input type="checkbox"/> Unable to determine
4. If one thing indicates something else, it is a sign of that thing ("Dreams can help indicate your true feelings")	<input type="checkbox"/> No relation
5. If a technical instrument indicates something, it shows a measurement or reading ("The temperature gauge indicated that it was boiling")	<input type="text"/>
6. When drivers indicate, they make lights flash on one side of their vehicle to show that they are going to turn in that direction ("He told us when to indicate and when to change gear")	<input type="text"/>

Figure 4.11. Question and Choices

In the present study, subdividing *related meaning* into specific cases, such as hyponymy was considered at first, but it was kept simple to reduce cognitive load. Furthermore, excluding related meaning was also considered, so as to be a better model of a yes/no judgment task; however, this seemed not to match the author's own intuitions about the nature of the alignments, namely that they may be gradable (Erk and McCarthy, 2009; Erk et al., 2009) or may contain non-exact similarities (Meyer and Gurevych, 2011).

The WordNet sense is used as the question and the COBUILD senses as choices since this study is ultimately interested in working in this direction, i.e., from a WordNet based WSD system to COBUILD examples. In addition, presenting senses in a dictionary format (i.e. as definitions) is based on the purpose of the current study, in which the system tries to map sense definitions between dictionaries.

The final question of every survey is a question about participant confidence for all questions, using a Likert scale from 1 to 5, as shown in Figure 4.12. In addition to the *unable to determine* category, this helps to determine annotator ability and reliability for semi-experts.

10. (optional) On a scale of 1-5 (1=not at all confident, 5=very confident), how confident are you in your answers to this survey?

	1 not at all confident	2 less confident	3 confident	4 more confident	5 very confident
Q1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q5	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q6	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q7	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q8	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q9	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 4.12. Confidence Scale

The surveys were administered to Linguistics faculty and students in the Departments of Linguistics and related fields at Georgetown University and Indiana University. Volunteers completed and anonymously submitted the surveys online. The surveys were administered via a free web service (<http://www.surveymonkey.com/>). While this makes implementing such experiments feasible for research in almost any context, there are distinct limitations, such as not being able to track the same user across different surveys.

4.5.2 Evaluation data

4.5.2.1 Overview

Before delving into detailed evaluation, an example set of responses can be shown as in Table 4.5. This is for the first WordNet sense (W1) of the noun *section*, which has three possible corresponding COBUILD senses. Responses indicate that sense 2 (C2) is a favorite, but C1 is also likely; C3 is divided, leaning towards not related.

section (n) : a self-contained part of a larger composition ...	Same meaning	Related meaning	Unable to determine	No relation
C1. A section of something is one of the parts into which it is divided ...	38.5% (5)	53.8% (7)	7.7% (1)	0.0% (0)
C2. A section of an official document ... is one of the parts into which it is divided ...	76.9% (10)	15.4% (2)	0.0% (0)	7.7% (1)
C3. A section is a diagram of something such as a building ...	0.0% (0)	38.5% (5)	7.7% (1)	53.8% (7)

Table 4.5. Response Analysis for one WordNet sense of *section.n*

This variability is typical of the responses, as one can see in Figure 4.13, where the counts for each type of response for each word are summed up in the present approach. One can also see the differing numbers of annotators in this graph, with *job.n*, for example, receiving more responses than *policy.n* in the present experiment. For a word like *job.n*, the number of responses for *no relation* predominates, but for *community.n*, there are more *related meaning* instances. Most notably, as with the study in Erk et al. (2009), respondents are clearly using not just the extreme categories (same/different), but are making great use of the *related meaning* category. Indeed, in total, *no relation* was the most popular answer (866 responses), followed closely by *related meaning* (828) and then *same meaning* (472); *Unable to determine* (146) was the least popular choice, but still accounted for 6.3% of the responses. This not only supports the author's intuition that sense mapping is not so simple as to be divided into yes or no but it also provides convincing evidence for the findings of previous studies in that sense may be gradable (Erk and McCarthy, 2009; Erk et al., 2009) or may even contain non-exact similarities

(Meyer and Gurevych, 2011).

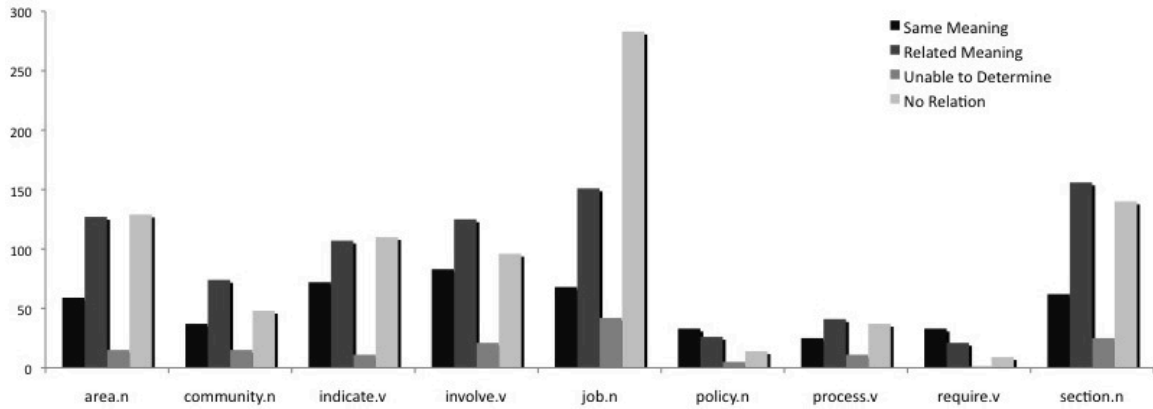


Figure 4.13. Number of times each answer was used for every word

Turning to how well respondents agreed on their answers, when Fleiss' kappa is calculated to test inter-annotator agreement, a value of 0.18 is obtained; according to Landis and Koch (1977), this is only a “*slight*” agreement. This lack of agreement is not surprising if one looks at the participants’ confidence, as in Table 4.6. Around 50% of WordNet senses result in confidence scores of 3 or below. One has to note that the alignment task is tough for even human. Meyer and Gurevych (2011) also noted the difficulty of the alignment task to human annotators; even two most skilled annotators showed an agreement of .80 in their alignment task between WordNet and Wiktionary, comparing their agreement in the alignment task between WordNet and Wikipedia.

1	2	3	4	5
0.21	0.97	1.98	1.63	1.70

Table 4.6. Average number of responses for each point on the confidence scale (1=not confident, 5 = very confident)

Part of the difficulty seems to lie in the fact that within each inventory, senses are related in complicated ways, sometimes causing confusion for annotators in mapping between them. For *community.n*, for example, the three COBUILD senses are:

- C1. The community is all the people who live in a particular area or place
- C2. A particular community is a group of people who are similar in some way
- C3. Community is friendship between different people or groups, and a sense of having something in common

When asked to align the WordNet sense of *common ownership*, then this property can cut across all three definitions, but seems to be describing a different way at looking at *community* completely.

Given the lack of agreement, an immediate question is: can these results be used to evaluate WSA systems? One answer is that the results should be used as weighted scores. That is, when evaluating measures such as precision and recall, instead of counting C2, for example, as a totally correct alignment for W1 of *section.n* (cf. Table 4.5), it counts as .769 of a correct alignment. One can see Madnani et al. (2011) for such a proposal using binary crowdsourced data, and Erk and McCarthy (2009) for different measurements related to graded word senses.

An alternative is to seek whether this approach can obtain higher confidence in the way that the classes are used. To address this, the calculations are adjusted by removing the *unable to determine* cases and combining the *same* and *related* meanings. This reflects the fact that one may want to group them together for particular alignment uses;

this gives a kappa of 0.24 (“fair” agreement). Again, the low agreement is not terribly surprising, given the low confidence reported earlier, and it can indicate at least one of two things: 1) the task was not clear, or 2) these particular sense inventories are difficult to align.

4.5.2.2 Evaluation Metrics

The responses are converted into scores for evaluation, in order to quantify to what extent—according to the various annotators—the senses from the two inventories express the same meaning. Specifically, a weight of 1 is assigned for *same meaning*, -1 for *no relation*, and 0 for *unable to determine*; thus higher scores indicate a greater degree of “sameness”²⁰. For *related meaning*, the system is tested with different weights (α)—1, 0.5, and 0—reflecting differing degrees of their contribution towards a correct alignment. For example, for the W1-C1 mapping in Table 4.5, the system generates: $5+7*1+0 = 12$, $5+7*0.5+0 = 8.5$, and $5+7*0+0 = 5$, respectively.

Participants were not required to complete all surveys, so the number of responses per survey is different. Thus, the scores are normalized by the number of respondents: in this case, with $\alpha = 0.5$, the score for W1-C1 is $8.5/13 = 0.65$. For example, normalized scores for *involve.v* are in Table 4.7 (see Appendix E, for scores of all nine words). The closer to 1 the score is, the greater the strength in aligning those senses.

²⁰ One could also use normalized judgment scores as in (Erk and McCarthy, 2009). In the current context, this means: *same*=2, *related*=1, *none*=0, and normalized score = score/2. Instead of ranging from -1 to 1, it ranges from 0 to 1, but shares the same basic intuition, especially for when $\alpha = 0$, putting *related meaning* exactly halfway between the others.

	C1	C2	C3	C4	C5
W1	0.08	0.50	0.33	0.08	0.25
W2	-0.06	0.75	0.69	0.50	0.13
W3	1.00	-0.14	-0.29	-0.14	-0.43
W4	0.61	-0.11	-0.44	-0.06	-0.39
W5	0.89	0.68	0.25	0.36	-0.07
W6	-0.46	0.23	0.23	0.41	-0.32
W7	-0.57	-0.14	0.29	-0.21	0.00

Table 4.7. Scores for *involve.v* ($\alpha = 0.5$)

Using the scores, the present approach performs two ways of counting different cases as correct alignments, namely counting: 1) *all positive* scores (unshaded cells of Table 4.7); or 2) only the *top positive* score for each WordNet sense (i.e., the highest score reading across a row in bold). The second method, which is referred to as *top positive*, matches the assumption of having only one link for each WordNet sense. A graded notion of what counts as correct to calculate precision and recall could also be explored (Erk and McCarthy, 2009). In the present study, however, the WSA system is used on its own as a categorical one, returning yes or no for each alignment link, so a categorical evaluation is used.

With the present purpose, after defining a set of correct alignments by categorical decisions (i.e., yes/no), precision and recall of alignments are calculated in the usual way. Precision is the number of correct links divided by the number of guessed, and recall divides by the number of links in the gold alignment. Given that false mappings can lead learners astray, precision is generally more of a concern in the present study.

4.5.3 Evaluating WSA system

The system is evaluated in three different ways: (1) by comparing the performance of the WSA system with and without accounting for the flatness of alignment structure, (2) by weighting *related meaning* (i.e. α) differently, and (3) by expanding the evaluation data to the 20 words²¹.

4.5.3.1 Counting flatness of alignment structure

In order to explore how much improvement is achieved from the initial alignment based on WSD output to the adjusted alignment by counting the alignment structure, each of the WSD systems are evaluated by comparing precision and recall over the nine words by (1) running the WSD systems and taking the top COBUILD sense for each WordNet sense (WSD) (= the initial alignment) and (2) after running word sense alignment (WSA) with counting the probability of the alignment structures (= the adjusted alignment). Precision and recall are computed by different gold alignments (i.e., *all positive*, *top positive*). For example, with *related meaning* counted as half of the same meaning (i.e. $\alpha=0.5$), the table of mapping scores between WordNet and COBUILD is generated as shown in Table 4.7 in the previous section. When the gold alignments are set by *all positive* scores between the WordNet and COBUILD mapping (unshaded cells of Table 4.7), the precision for *involve.v* is 0.57; with the system outputs, {(W1, C1), (W2, C2), (W3, C3), (W4, C5), (W5, C1), (W6, C2), (W7, C4)}, the correct matches are {(W1, C1), (W2, C2), (W5, C1), (W6, C2)} and thus the precision is $4/7 = 0.57$. For recall, the gold alignments are 19 mappings (19 positive scores in Table 4.7) and thus the recall is $4/19 =$

²¹ *resilient.a, expenditure.n, mend.v, unveil.v, sector.n, chain.n, conscience.n, cradle.n, outfit.n, agitate.v, fatigue.n, obedience.n, trivial.a, deliberately.r, aspect.n, banish.v, resist.v, indicate.v, alternate.a, trigger.v*

0.21. When the gold alignments are set by *top* positive scores for each WordNet sense over COBUILD senses (i.e., the highest score reading across a row in bold), the precision for *involve.v* is 0.29; with the same outputs of the system, the correct matches are {(W2, C2), (W5, C1)}, and thus the precision is $2/7 = 0.29$. For the recall, the gold alignments are seven mappings (seven top scores for each WordNet sense in Table 4.7) and thus the recall is $4/7 = 0.29$. The following Table 4.8 shows the evaluation results based on human gold standard (e.g., *top* positive) with counting *related meaning* as half of the *same meaning* (i.e. $\alpha=0.5$). Table 4.8 also shows how the alignment structure becomes flatter from the initial alignment to adjusted alignment.

word	Human alignment (top positive, $\alpha=0.5$)	Initial alignment (based on SR::AW)		Adjusted alignment			
			P	R		P	R
area	 (w1,c1),(w2,c6),(w3,c4), (w4,c6), (w5,c3), (w6,c5)	 (w1,c1),(w2,c4),(w3,c5), (w4,c6), (w5,c2), (w6,c4)	0.33	0.33	 (w1,c1),(w2,c4),(w3,c5), (w4,c6), (w5,c2), (w6,c3)	0.33	0.33
type	(0,1,1,1,1,2)	(0,1,1,1,1,2)			(1,1,1,1,1,1) -- flatted		
commu- nity	 (w1,c1),(w2,c3),(w3,c2), (w4,c3), (w5,c1), (w6,c1)	 (w1,c1),(w2,c2),(w3,c2), (w4,c2), (w5,c3), (w6,c1)	0.50	0.50	 (w1,c1),(w2,c3),(w3,c2), (w4,c2), (w5,c3), (w6,c1)	0.67 ↑	0.67 ↑
type	(1,2,3)	(1,2,3)			(2,2,2) -- flatted		

word	Human alignment	Initial alignment	P	R	Adjusted alignment	P	R
indicate	<p>(w1,c4),(w2,c3),(w3,c2), (w4,c1), (w5,c1)</p>	<p>(w1,c2),(w2,c3),(w3,c6), (w4,c2), (w5,c6)</p>	0.20	0.25	<p>(w1,c2),(w2,c3),(w3,c1), (w4,c4), (w5,c6)</p>	0.20	0.25
type	(0,0,1,1,1,2)	(0,0,0,1,2,2)			(0,1,1,1,1,1) -- flatted		
involve	<p>(w1,c2),(w2,c2),(w3,c1), (w4,c1),(w5, c1), (w6,c4), (w7,c3)</p>	<p>(w1,c1),(w2,c2),(w3,c3), (w4,c5),(w5, c1), (w6,c2), (w7,c4)</p>	0.29	0.29	<p>(w1,c1),(w2,c2),(w3,c3), (w4,c5),(w5, c1), (w6,c2), (w7,c4)</p>	0.29	0.29
type	(0,1,1,2,3)	(1,1,1,2,2)			(1,1,1,2,2)		
job	<p>(w1,c1),(w2,c3),(w3,c2), (w4,c3),(w5, c4),(w6, c4) (w7,c5), (w8,c5),(w9, c1) (w10,c2), (w11,c1), (w12, c2)</p>	<p>(w1,c5),(w2,c3),(w3,c2), (w4,c2),(w5, c1),(w6, c4) (w7,c3), (w8,c5),(w9, c5) (w10,c5), (w11,c3), (w12, c4)</p>	0.17	0.22	<p>(w1,c5),(w2,c3),(w3,c2), (w4,c2),(w5, c1),(w6, c4) (w7,c3), (w8,c5),(w9, c1) (w10,c5), (w11,c3), (w12, c4)</p>	0.17	0.22
type	(2,2,2,3,3)	(1,2,2,3,4)			(2,2,2,3,3) -- flatted		
policy	<p>(w1,c1),(w2,c1),(w3,c3)</p>	<p>(w1,c1),(w2,c2),(w3,c3)</p>	0.67	0.67	<p>(w1,c1),(w2,c2),(w3,c3)</p>	0.67	0.67
type	(0,1,2)	(1,1,1)			(1,1,1)		

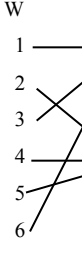
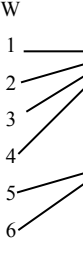
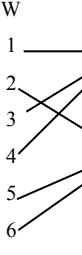
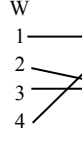
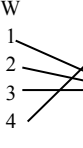
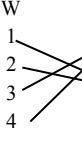



word	Human alignment	Initial alignment	P	R	Adjusted alignment	P	R
process	 <p>(w1,c1),(w2,c2),(w3,c1), (w4,c2), (w5,c2), (w6,c1)</p>	 <p>(w1,c1),(w2,c1),(w3,c1), (w4,c1), (w5,c2), (w6,c2)</p>	0.50	0.50	 <p>(w1,c1),(w2,c2),(w3,c1), (w4,c1), (w5,c2), (w6,c2)</p>	0.67 ↑	0.67 ↑
	type	(3,3)			(2,4)		
require	 <p>(w1,c1),(w2,c2),(w3,c2), (w4,c1)</p>	 <p>(w1,c2),(w2,c2),(w3,c2), (w4,c1)</p>	0.75	0.75	 <p>(w1,c2),(w2,c2),(w3,c1), (w4,c1)</p>	0.50 ↓	0.50 ↓
	type	(2,2)			(1,3)		
section	 <p>(w1,c2),(w2,c3),(w3,c1), (w4,c1),(w5,c1),(w6,c1), (w7,c1),(w8,c3),(w9,c1), (w10,c1)(w11c1)(w12c1) (w13,c1),(w14,c3)</p>	 <p>(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c2), (w7,c1),(w8,c1),(w9,c1), (w10,c2)(w11c2)(w12c1) (w13,c1),(w14,c1)</p>	0.36	0.42	 <p>(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c2), (w7,c1),(w8,c1),(w9,c1), (w10,c2)(w11c2)(w12c1) (w13,c1),(w14,c1)</p>	0.36	0.42
	type	(1,3,10)			(0,5,9)		

Table 4.8. Results of the initial alignment and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *top* positive; improvements in bold)

As shown in 4.8, the adjusted alignment clearly shows more flatness, which manifests that the WSA algorithm does effectively function in reflecting the heuristic of favoring flatness. While most initial alignments are revised to be flatter, the precision is not improved as much as the alignments are flattened; two cases show improvement of its precision among six cases of being flattened. However, although the adjusted alignment shows small changes, it still gives a positive future that the heuristic of favoring flatter alignments has a potential to be enhanced.

The same method of comparing precision and recall over nine words is processed in the *all* positive evaluation (see Appendix E) and the other two WSD systems (i.e., SL, NB) are evaluated by the same method in the *all* positive and *top* positive evaluations (see Appendix F).

Accordingly now the three systems (i.e., SR::AW, SL, NB) are evaluated by comparing precision and recall *average* over the nine words. The results are in Table 4.9 and Table 4.10.

As shown in Table 4.9 and Table 4.10., *SenseRelate::AllWord* (SR::AW) and Naïve Bayes (NB) systems show a mild improvement in both the AP and TP conditions from the initial alignments to the adjusted alignments, as shown in bold (Table 4.9 and Table 4.10). While *SenseLearner* (SL) does not show improvement in the *all* positive evaluation (Table 4.9), the WSA algorithm does help it to find the best sense in the *top* positive evaluation (Table 4.10). Based on the alignment results (i.e., precision and recall) with counting $p(A_s)$ (=adjusted alignment), *F-score*, the “*harmonic mean*” of precision and recall, is computed and among the three, SR::AW is presented as the best system to be a basis for WSA processing, with the highest *F-scores* in both the *all* positive and the top

positive conditions (i.e., 0.456 for the all positive, 0.436 for the top positive).

Method	<i>initial alignment</i>		<i>adjusted alignment</i>		
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>F₁</i>
<i>SL</i>	0.646	0.281	0.624	0.275	0.382
<i>SR::AW</i>	0.583	0.349	0.607	0.365	0.456
<i>NB</i>	0.545	0.23	0.601	0.259	0.362

Table 4.9 Results of initial alignment and adjusted alignment over three WSD systems (SL, SR::AW, NB) ($\alpha=0.5$, gold standard = **all** positive; improvements in bold)

Method	<i>initial alignment</i>		<i>adjusted alignment</i>		
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>F₁</i>
<i>SL</i>	0.339	0.282	0.344	0.294	0.317
<i>SR::AW</i>	0.418	0.436	0.427	0.445	0.436
<i>NB</i>	0.399	0.322	0.414	0.331	0.368

Table 4.10. Results of initial alignment and adjusted alignment over three WSD systems (SL, SR::AW, NB) ($\alpha=0.5$, gold standard = **top** positive; improvements in bold)

Shown in Table 4.10, from the *top* positive evaluation, around 40% of the mappings, in the present experiment, link each WordNet sense to its best corresponding COBUILD

sense, while with the *all* positive evaluation as shown in Table 4.9, around 60% point at least to a relevant sense. In passing, the higher precision of the *all positive* evaluation was noted, due to more possible senses to match. Although the changes are small, there is consistent improvement by counting $p(A_s)$ to the WSD outputs. This is particularly promising, given that the heuristic of favoring flatter alignments has much the potential to be refined.

To better gauge the effect of applying the WSA algorithm, the number of WordNet senses mapped to COBUILD based on WSD outputs for the nine words is counted. As shown in Table 4.11, out of 63 WordNet senses in total, SL generates 44 WordNet senses mapped to COBUILD senses based on its WSD outputs, SR::AW generates all 63 WordNet senses mapped to COBUILD senses, and NB, 44 WordNet senses mapped to COBUILD. Then, the number of link (mapping) changes proposed by counting $p(A_s)$ is shown in Table 4.11. For example, from the initial alignment based on SR::AW WSD outputs (i.e., 63 links), eight links are changed by applying $p(A_s)$ (i.e., 12.7%). As one can see, the system changes 6.8%–18.2% of the links by counting $p(A_s)$, with the most changes coming for the Naive Bayes method—which was the least accurate to begin with and saw the greatest increase in accuracy.

Method	Links	Changes
SL	44	3 (6.8%)
SR::AW	63	8 (12.7%)
NB	44	8 (18.2%)

Table 4.11. Number of link changes from the initial alignments to the adjusted alignments

4.5.3.2 Different weight on related meaning

This section shows how the system performs differently by different handling of *related meaning*. Table 4.12 and Table 4.13 show precision and recall of all nine words on the initial alignment based on SR::AW WSD outputs (Table 4.12) and the adjusted alignment by counting the probability of the alignment structures, $p(A_s)$ (Table 4.13) respectively. Gold alignments are categorized into three by how *related meaning* (i.e. $\alpha=1, 0.5, \text{ or } 0$) is scored. For each different count of *related meaning*, precision and recall are computed by different gold alignments (i.e. all positive, top positive). The same method of computing precision and recall is processed for the other counting of *related meaning* (i.e. $\alpha=1, \text{ or } 0$).

With $\alpha = 1$, related senses are counted fully correct, meaning the system will match more, giving higher precision. As shown in Table 4.12 and Table 4.13, for example, for *involve.v* based on AP gold alignments, precision gets lower as related senses are counted less (i.e., with lower α); 0.71 with $\alpha = 1$, 0.57 with $\alpha = 0.5$, and 0.29 with $\alpha = 0$. Likewise, exclusion of counting *related meaning* (i.e., $\alpha = 0$) subsequently results in fewer alignments, producing generally higher recall. For example, for *policy.n* in Table 4.12 and Table 4.13, recall gets higher as related senses are counted less; 0.38 with $\alpha = 1$, 0.43 with $\alpha = 0.5$, and 0.60 with $\alpha = 0$.

These tables (Table 4.12 and Table 4.13) also report the number of senses that do not align to the other inventory. As shown in the tables, the *un-alignment* between two inventories is generated more as *related meaning* is counted less. For example, when *related meaning* is not counted at all, 19 out of 63 WordNet senses are not aligned to COBUILD senses and six out of 35 COBUILD senses are not aligned to WordNet senses.

This makes sense considering the fact that less counting of *related meaning* gives fewer alignments.

SR::AW		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 0$	
		AP	TP	AP	TP	AP	TP
area.n	P	0.67	0.33	0.33	0.33	0.17	0.17
	R	0.21	0.33	0.18	0.33	0.14	0.20
community.n	P	0.83	0.50	0.67	0.50	0.67	0.50
	R	0.38	0.50	0.44	0.50	0.57	0.60
indicate.v	P	0.60	0.20	0.20	0.20	0.20	0.20
	R	0.16	0.25	0.09	0.25	0.14	0.25
involve.v	P	0.71	0.14	0.57	0.29	0.29	0.14
	R	0.20	0.14	0.21	0.29	0.18	0.17
job.n	P	0.42	0.25	0.33	0.17	0.17	0.08
	R	0.21	0.30	0.22	0.22	0.17	0.14
policy.n	P	1.00	0.67	1.00	0.67	1.00	0.67
	R	0.38	0.67	0.43	0.67	0.60	0.67
process.n	P	0.50	0.50	0.50	0.50	0.50	0.50
	R	0.43	0.50	0.43	0.50	0.60	0.60
require.v	P	1.00	0.75	1.00	0.75	0.75	0.75
	R	0.50	0.75	0.57	0.75	0.60	0.75
section.n	P	0.86	0.50	0.64	0.36	0.43	0.21
	R	0.57	0.50	0.56	0.42	0.55	0.38
un-aligned senses	WN	4	5	8	8	19	19
	CB	0	6	0	5	2	6

Table 4.12. Precision & recall of words from the initial alignment based on WSD (SR::AW) outputs (AP=*all* positives, TP=*top* positive), plus the total number of un-aligned senses

WSA (based on SR::AW)		$\alpha=1$		$\alpha=0.5$		$\alpha=0$	
		AP	TP	AP	TP	AP	TP
area.n	P	0.67	0.33	0.33	0.33	0.17	0.17
	R	0.21	0.33	0.18	0.33	0.14	0.20
community.n	P	0.83	0.67 ↑	0.83 ↑	0.67 ↑	0.67	0.50
	R	0.38	0.67 ↑	0.56 ↑	0.67 ↑	0.57	0.60
indicate.v	P	0.60	0.20	0.40 ↑	0.20	0.40 ↑	0.20
	R	0.16	0.25	0.18 ↑	0.25	0.29 ↑	0.25
involve.v	P	0.71	0.14	0.57	0.29	0.29	0.14
	R	0.20	0.14	0.21	0.29	0.18	0.17
job.n	P	0.42	0.25	0.33	0.17	0.17	0.08
	R	0.21	0.30	0.22	0.22	0.17	0.14
policy.n	P	1.00	0.67	1.00	0.67	1.00	0.67
	R	0.38	0.67	0.43	0.67	0.60	0.67
process.n	P	0.67 ↑	0.67 ↑	0.67 ↑	0.67 ↑	0.67 ↑	0.67 ↑
	R	0.57 ↑	0.67 ↑	0.57 ↑	0.67 ↑	0.80 ↑	0.80 ↑
require.v	P	1.00	0.50 ↓	0.75 ↓	0.50 ↓	0.50 ↓	0.50 ↓
	R	0.50	0.50 ↓	0.43 ↓	0.50 ↓	0.40 ↓	0.50 ↓
section.n	P	0.86	0.50	0.57 ↓	0.36	0.43	0.21
	R	0.57	0.50	0.50 ↓	0.42	0.55	0.38
un-aligned senses	WN	4	5	8	8	19	19
	CB	0	6	0	5	2	6

Table 4.13. Precision & recall of words from the adjusted alignment by counting $p(A_s)$ (AP=*all* positives, TP=*top* positive), plus the total number of un-aligned senses

These results do not indicate the best evaluation. They simply illustrate how the treatment of *related meaning* affects the results. The *all* positive (AP), $\alpha = 1$ evaluation, for instance, indicates how far off a system is from any correct answer, while, on the

other side of the spectrum, the *top* positive (TP), $\alpha = 0$ evaluation indicates how well the best senses are being found. Those two cases would answer two different questions respectively; (1) is the current study leading learners astray or not? and (2) is the current study finding the best possible sense to show learners? To display sense-specific examples for learners, the present research will thus want evaluations across the spectrum to know how often learners will be presented with related examples, as opposed to exact matches.

Examining the results from Table 4.12 and Table 4.13, the results without (Table 4.12) and with (Table 4.13) counting the probability of the alignment structure, $p(A_s)$, show that the system performance is improved from the initial alignments based on WSD outputs to the adjusted alignments by counting $p(A_s)$, as presented by arrows in Table 4.13. This indicates that there is promise in adjusting the output of the WSD systems, which is discussed earlier with Table 4.9 and Table 4.10. There is, however, one exception, *require.v*, which worsens its performance from WSD mapping to WSA mapping. It has two COBUILD senses and four WordNet senses; one COBUILD has one WordNet link and the other COBUILD sense has three WordNet links in the initial alignments based on WSD outputs, which is skewed. By revising the links in the WSA step to be flatter, one link from three links between COBUILD and WordNet is re-mapped to the other COBUILD-WordNet link, resulting in two WordNet senses evenly mapped to two COBUILD senses respectively. However, the changed link is correct before changing and becomes incorrect after changing. Consequently, the performance gets worse. This seems to be caused by inappropriate WSD outputs. The re-mapped link between WordNet and COBUILD has a lower probability than others so that it is changed

while others are stuck to each other. This indicates that the WSA algorithm (i.e., $p(A_s)$) cannot improve the performance with incorrect WSD outputs, meaning that the accuracy of the WSD system is crucial.

The same evaluation with different α is conducted on WSA results based on the other two WSD systems (e.g. SL, NB) (see Appendix F, for detailed precision/recall of the adjusted alignment over nine words based on the other two WSD systems). Finally, the precision and recall average over nine words for their WSA results based on the three WSD systems are presented in Table 4.14. Those results are also graphically displayed in Figure 4.14. For example, the precision of SR::AW, when *related meaning* is fully counted, is *0.751* based on *all* positive gold alignments and *0.437* based on *top* positive gold alignments. Precision is computed by averaging all correspondent precision figures (graded figures) over the nine words in Table 4.13 (e.g., $0.751 = (0.67+0.83+0.60+0.71+0.42+1.00+0.67+1.00+0.86)/9$). Other precision and recall scores are computed in the same way.

WSA		$\alpha=1$		$\alpha=0.5$		$\alpha=0$	
		AP	TP	AP	TP	AP	TP
SR::AW	P	0.751	0.437	0.607	0.427	0.478	0.349
	R	0.353	0.448	0.365	0.445	0.411	0.412
SL	P	0.730	0.391	0.624	0.344	0.441	0.344
	R	0.253	0.323	0.275	0.294	0.295	0.342
NB	P	0.741	0.417	0.601	0.414	0.413	0.353
	R	0.244	0.333	0.259	0.331	0.268	0.318

Table 4.14. Precision & recall average over nine words processed by the WSA algorithm

based on three different WSD systems (SR::AW; SL; NB)

(AP=*all* positives, TP=*top* positive)

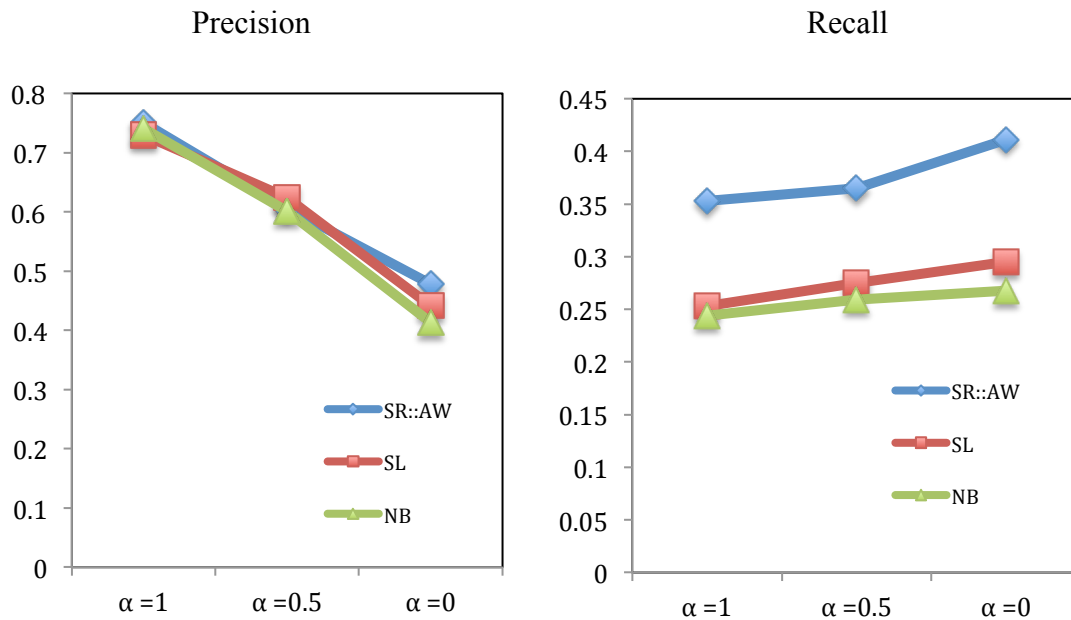


Figure 4.14. Precision & recall average over nine words processed by WSA (AP)

As shown in Table 4.14 and Figure 4.14, one can find the same trend across WSA outputs based on different WSD systems; precision gets higher by higher α , whereas recall gets higher by lower α . Precision changes are relatively big by scores of *related meaning* while recall does not show a relatively steep change over scores of *related meaning*. These findings suggest that how much *related meaning* needs to be counted in gold alignments depends on the purpose of the study; if the study focuses more on the system accuracy, *related meaning* needs to be counted more. The value $\alpha = 0.5$ is used in this study as a compromise between counting *related meaning* as fully as *same meaning* and counting *related meaning* as *no relation*.

4.5.3.3 Evaluation of the system

As demonstrated in previous sections, the WSA system based on SR::AW WSD output shows a higher performance. This WSD module is thus employed in the current system, which is examined by users (i.e., language learners) its role of assisting language learners in vocabulary acquisition and reading comprehension (see Chapter 5).

The system performance is evaluated with 20 target words. The overall results of the initial alignment based on the WSD outputs and the adjusted alignments counting the probability of alignment structure, $p(A_s)$, on those 20 words are presented in Table 4.15. The second column shows the WordNet sense generated by the WSD system (i.e. SR::AW). The third column presents the COBUILD sense mapped from the WordNet sense in the second column, generated by the initial alignment based on WSD outputs (= by taking the top COBUILD sense per each WordNet sense based on sense probability distribution generated by the WSD module, that is, $\max_A p(w_n, c_b)$). The fourth column displays the COBUILD sense corresponding to the WordNet sense (in the second column) by the adjusted alignments processed by applying $p(A_s)$ (i.e., $\max_A p(A_s) p(w_n, c_b)$). The last column shows the gold COBUILD senses manually annotated by humans. In presenting WordNet and COBUILD senses in Table 4.15, w refers to WordNet and c refers to COBUILD, the number next to it represents its sense number, and a lower case alphabet letter in a parenthesis represents POS; for example, $w_2(a)$ means that the word is WordNet sense number 2 and an *adjective*.

word	WSD	Initial alignment	Adjusted alignment	Human alignment
resilient	w_2 (a)	c_1 (a)	c_1 (a)	c_1 (a)
expenditure	w_1 (n)	c_2 (n)	c_2 (n)	c_1 (n)
mend	w_2 ((n))	No c for noun		c_1 ((v))
unveil	w_2 ((n))	No c for noun		c_2 ((v))
sector	w_4 (n)	c_1 (n)	c_1 (n)	c_2 (n)
chain	w_3 (n)	c_1 (n)	c_1 (n)	c_4 (n)
conscience	w_1 (n)	c_2 (n)	c_1 (n)	n_3
cradle	w_1 (n)	c_4 (n)	c_4 (n)	c_4 (n)
outfit	w_1 ((v))	c_1 ((v))	c_1 ((v))	c_1 ((n))
agitate	w_6 (v)	c_1 (v)	c_1 (v)	c_3 (v)
fatigue	w_1 (n)	c_1 (n)	c_1 (n)	c_1 (n)
obedience	w_2 (n)	c_1 (n)	c_1 (n)	c_1 (n)
trivial	w_1 (a)	c_1 (a)	c_1 (a)	c_1 (a)
deliberately	w_2 (r)	c_2 (r)	c_2 (r)	c_2 (r)
aspect	w_3 (n)	c_2 (n)	c_2 (n)	c_1 (n)
banish	w_2 (v)	c_1 (v)	c_1 (v)	c_2 (v)
resist	w_2 (v)	c_4 (v)	c_4 (v)	c_1 (v)
indicate	w_2 (v)	c_3 (v)	c_3 (v)	c_1 (v)
alternate	w_1 ((v))	c_1 ((v))	c_1 ((v))	c_3 ((a))
trigger	w_1 ((n))	c_2 ((n))	c_2 ((n))	c_2 ((v))

Table 4.15. Results of WSA on 20 words

As shown in Table 4.15, six words out of 20 words are presented with sense appropriate lexical information (*resilient, cradle, fatigue, obedience, trivial, deliberately*) and 14 words are in error. However, one should note here that five words (*mend, unveil, outfit, alternate, trigger* – crossed out in the Table) out of 14 words in error are caused by the part of speech (POS) error (the POS error is shown in double parenthesis – e.g., w_2

((*n*)), neither by WSD nor by WSA. Since the WSD system (SR::AW) employed in the current system performs POS tagging using its pre-built tagger in addition to its sense disambiguation, the system is unlikely to remedy the POS error separately. This means, if POS tagging works correctly, the overall performance is more likely to be improved. Let us look at *trigger* in Table 4.15, for example. Although the WSD module improperly tags it as $w_1(n)$ (WN noun, sense #1, in the second column), the WSA algorithm correctly²² mapped it as $c_2(n)$ (COBUILD noun, sense #2, in the fourth column). That is, WN-CB mapping for *trigger* as a *noun* is correct (i.e., $w_1(n) \sim c_2(n)$), although the correct alignment as a *noun* is ignored because the correct POS for *trigger* in the context is a *verb*. This supports the claim that the overall system performance could be boosted when a better performing POS tagger is employed. The system is being refined to perform POS tagging and WSD separately, employing a state-of-the-art POS tagger.

Then, let us look at the nine words in error caused by sense alignment. Among the nine words, one word (i.e., *conscience*) shows changes in mapping from the initial alignment based on WSD output to the adjusted alignments with $p(A_s)$ counted. What the present study focuses on here is where the error is originating from; e.g., is it from the WSD module (=wrongly disambiguated)? Or is it from the initial mapping based on WSD output (before the WSA algorithm applied)? Or is it from the adjusted mapping with counting probability of alignment structure? As you can find in Table 4.15, the error is caused at the WSD stage, disambiguating $w_1(n)$ (second column), which should be $w_2(n)$, based on human annotation²³, or possibly be n_3 in terms of “following rightness” in the following context in (15):

²² WSA output (fourth column) for *trigger* is manually checked by human

²³ WordNet sense for *conscience* is manually checked by human.

(15) Many of us ease our *conscience* about all this waste by donating our surplus clothing to charities.

w_1: motivation deriving logically from ethical or moral principles that govern a person's thoughts and actions

w_2: conformity to one's own sense of right conduct

w_3: a feeling of shame when you do something immoral

[c_3] conscience is a feeling of guilt because you know you have done something that is wrong

According to WSA outputs for *conscience* (i.e., w3-c3) that is also manually confirmed by humans, the WordNet sense (by WSD, 2nd column) should be w_3 (n) in order to be mapped with w_3 (n) (human alignment, 5th column). However, the WSD module disambiguated it as w_1 (n), resulting in c_1 (n) (adjusted alignment, 4th column) by $p(A_s)$ (i.e., w1-c1). This kind of error can also be corrected when WSD performance is improved. Therefore, when WSD and POS tagging are improved, WSA performance would be enhanced.

On the other hand, the rest of the eight words do not show any changes from the initial alignment (3rd column) to the adjusted alignment (4th column). According to the heuristic of favoring flatter alignments in WSA processing, some changes between them are anticipated but these seven out of the eight words do not show changes. This may come

from (1) that the sense probability distributions (i.e. $p(w_n_i|c_b_j)$, see section 4.4.2) are more biased than the probability of the candidate alignments (i.e. $p(A_s)$, see section 4.4.3) not to make any changes while WSA is being processed. Or the un-change of mapping may come from (2) that the small number of senses of those 20 words. Since the 20 words are set for language learners (see Chapter 5), not for computationally evaluating the system performance, the number of senses of those 20 words was not as diverse as those nine words for computational evaluation in terms of various polysemy (see 4.5.1.2) as shown in Table 4.16. The number of senses of those 20 words thus may be too small to draw reliable mapping results. Besides, when the system has a small number of senses, it is really hard to make any change, due to the fact that the system does not overturn the best sense links.

	9 words		20 words	
	total	average	total	average
WN	63	7	80	4
COBUILD	35	3.89	59	2.95

Table 4.16. Average sense numbers of nine words and 20 words

As shown in Table 4.16, the nine words chosen for computational evaluation are more polysemous than the 20 words chosen for learner evaluation. Thus, it would be interesting to conduct a further study to examine the system with more various polysemy words to validate the findings. In the next chapter, one will see how these errors affect learners' vocabulary learning and reading comprehension in the current working system.

4.6 Summary

The task of word sense alignment has been outlined within the context of providing relevant example sentences for language learners. Given the assumptions for this context, the present research developed an efficient algorithm for aligning two resources, in particular, WordNet senses (which a number of available WSD systems use) and COBUILD senses (which were designed with language learners in mind), counting the probability of alignment structures to the output of WSD, in order to build from state-of-the-art systems. Favoring flat alignments over skewed ones, the system generates the best alignment which has a maximum probability from the computation of sense probability distribution (i.e. $p(wn_i, cb_j)$) and possible alignment structure probability (i.e. $p(A_S)$).

The utility of the algorithm was then illustrated on an evaluation set. The study has examined constructing the database of alignments of word senses between two sense inventories, specifically WordNet and COBUILD, by pooling the judgments of semi-experts. Using online surveys, the study presented a sense of a target word from one dictionary with senses from the other dictionary, asking for judgments of relatedness. The system is first gauged by comparing the initial alignment and the adjusted alignment. Although it is small amount, the system performance is improved from the initial alignment to the adjusted alignment by counting $p(A_S)$ which validates the present study's heuristic of favoring flat alignments. Next, the accuracy of the system is gauged by differently weighting *related meaning* (e.g. $\alpha = 1, 0.5, 0$); precision gets higher by counting *related meaning* more no matter what systems are used. This suggests that *related meaning* should be considered in establishing gold sense mappings. Finally the system is evaluated with 20 words. The system showed an overall performance of 55%,

assuming that POS tagging is correctly performed. In the next chapter, the system is examined the effect of presenting sense-specific examples on learners' vocabulary acquisition and reading comprehension and the effects of alignment errors on their learning.

V. THE EMPIRICAL STUDY: EVALUATION OF THE WHOLE SYSTEM

This chapter describes how the present study fulfills its goal and provides details regarding the research questions (5.1), methods (5.2), and results and discussion (5.3).

5.1 Research questions

This empirical study was conducted to evaluate the online tutoring system developed in this research. As presented in Chapters 3 and 4, the current system was built to display sense-appropriate lexical information (i.e. definition and examples) to learners of English for difficult words in order to support their vocabulary learning and comprehension during reading.

The goal of this empirical study is therefore to investigate whether high-quality sense-specific lexical information presented by the intelligent reading system helps learners in their vocabulary acquisition and reading comprehension. Accordingly, the following research questions are posed for the present investigation.

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?

In order to fulfill this goal, the empirical study was conducted with a group of 60 native Korean speakers who were learning English as a second language (ESL) (5.2.1 Participants). Two weeks before the main experiment and posttests, the present study administered a pretest to measure the learners' prior vocabulary knowledge so as to prevent learners from focusing on the target words. During the main experiment, after reading one text, learners took a reading comprehension test. They then did the same for the second text. Following these two rounds, they took vocabulary posttests (5.2.2 Materials, 5.2.3 Procedure). After data from all 60 participants were collected, several statistical tests were run to analyze the data (5.2.4 Scoring, 5.2.5 Data analysis). In order to gauge the effect of automatic system errors, the present study also analyzed the cases in which the system gave incorrect information (5.2.5 Data analysis). More details of the whole process are provided in the following sections.

5.2 Method

This section describes the methodology of the present study in terms of participants, materials, procedure, scoring and data analysis.

5.2.1 Participants

The participants were recruited from three universities and a private institute in Seoul, Korea. As an a priori power analysis to get a reliable sample size for the study, a power analysis calculation (i.e., a G*Power²⁴ calculation) was performed. Based on an alpha

²⁴ The sample size plays an important role in all statistical analyses because if the sample size is not appropriate the results (e.g. differences) drawn from the analysis may not be truly reliable. Thus, in order to determine an appropriate sample size for the study, a statistical power analysis needs to be performed. For this study, G*Power (a statistical power analysis tool) is used to get an appropriate data set.

value of 0.05, a beta value of 0.05 (i.e., power = 0.95), and an effect size²⁵ of 0.4, it recommended a total sample size of 32 with an actual power of 0.96. This suggested that the present experiment required at least eight participants per group in order to get a reliable sample size of participants. The original number of participants was 88; after 28 participants who were not in the target range were excluded, the final number of participants was 60. This total of 60 participants was thus chosen for the four groups of this study (i.e., 15 participants per group).

The 40 participants from the three universities were non-English majors (mostly computer science, engineering, business, science, arts, or education majors). They were taking English courses to prepare for a test of English proficiency (e.g., Test of English Proficiency developed by Seoul National University (TEPS)) at the time of study. The 20 participants from the private institute were mostly university graduates from various backgrounds who, at the time of the present study, were taking teacher-training courses designed for elementary English teachers. Participants from each institution were randomly but proportionally²⁶ assigned into one of four groups. In addition, the Test of Homogeneity of Variances (i.e., Levene's Test) was run to ensure that the variances of the error among groups are equal at the outset of the study.

With the help of the instructors of the English courses in which the learners were enrolled, the author explained the present study and provided an opportunity for learners to volunteer to participate. Learners were required to report their TOEFL iBT® score, and only those whose level of English proficiency was in the target range of the study

²⁵ According to Cohen (1992)'s effect size conventions, for more than two groups (i.e., F-Test (ANOVA)), the effect size (d) of 0.10, 0.25, and 0.40 are considered small, medium and large respectively. The author expected a "large" effect, so the effect size was thus set as $d=0.40$.

²⁶ The author tried to assign participants from each institution into four groups as evenly as possible with the intent of avoiding having all of the participants from one institution in the same group.

were able to participate. The target range of English proficiency for this study was *intermediate*,²⁷ which, according to the ETS TOEFL center, is considered to be a score of 15 to 21 on the reading section of the TOEFL iBT®. Although the prospective participants were selected initially due to their TOEFL iBT® score, participants who received a score of more than 16 out of 20 target words correct in the pretest were eliminated from the final pool of participants (i.e., 28 out of 88 were excluded from the final pool).

Of the 60 participants in the final pool, 34 were female and 26 were male. The age of participants ranged from 21 to 39 years old, with an average age of 23.8 years. The length of their English study ranged from eight to 25 years of study, with an average length of 11.32 years. Finally, the 60 participants were randomly assigned to one of four groups, with 15 participants in each group. Groups 1, 2, and 3 were in the treatment condition and were provided with lexical information during the task, while Group 4 was in the control condition and was provided with no lexical information.

Group 1 (Gold Senses, GS): reading with support of gold-standard sense-specific lexical information

Group 2 (System Senses, SS): reading with support of system-derived sense-specific lexical information (i.e., automatically obtained senses)

²⁷ Intermediate learners were targeted in because this study requires participants to understand texts without much difficulty while still having relatively less vocabulary knowledge. Most beginning learners still have problems with reading comprehension, and most advanced learners already have relatively enough vocabulary knowledge that it would be hard to draw some learning effect from using the system of this study. The author also confirmed this with a pilot study; she had learners from a similar population read the passage (5.2.2.1 Reading texts), answer comprehension questions (5.2.2.3 Reading tests), and circle the words they did not know in order to make sure the reading passages would be comprehensible with an adequate number of unfamiliar words (5.2.2.2 Target words). Based on those facts, intermediate learners of English were decided to be an appropriate target of the study.

Group 3 (All Senses, AS): reading with support of lexical information of all senses of the chosen word

Group 4 (No Senses, NS): reading without any support of lexical information

The following figures are examples of lexical information with which each of the four groups were provided when they clicked, such as *chains*. As shown in Figure 5.1, learners in the GS group were provided with the correct sense-appropriate lexical information (COBUILD sense #4) when they clicked on the word. The system for the GS group is manually modified from the present system, which has some errors. That is, the author manually examined the original system output (Chapter 4) for errors; if she found an error, she then corrected it. For learners in the SS group, the automatic system happened to present incorrect lexical information (COBUILD sense #3), as shown in Figure 5.2. This case presents an example of when the system has errors; when the system has no errors, all lexical information provided to the SS group is the same as that presented to the GS group.

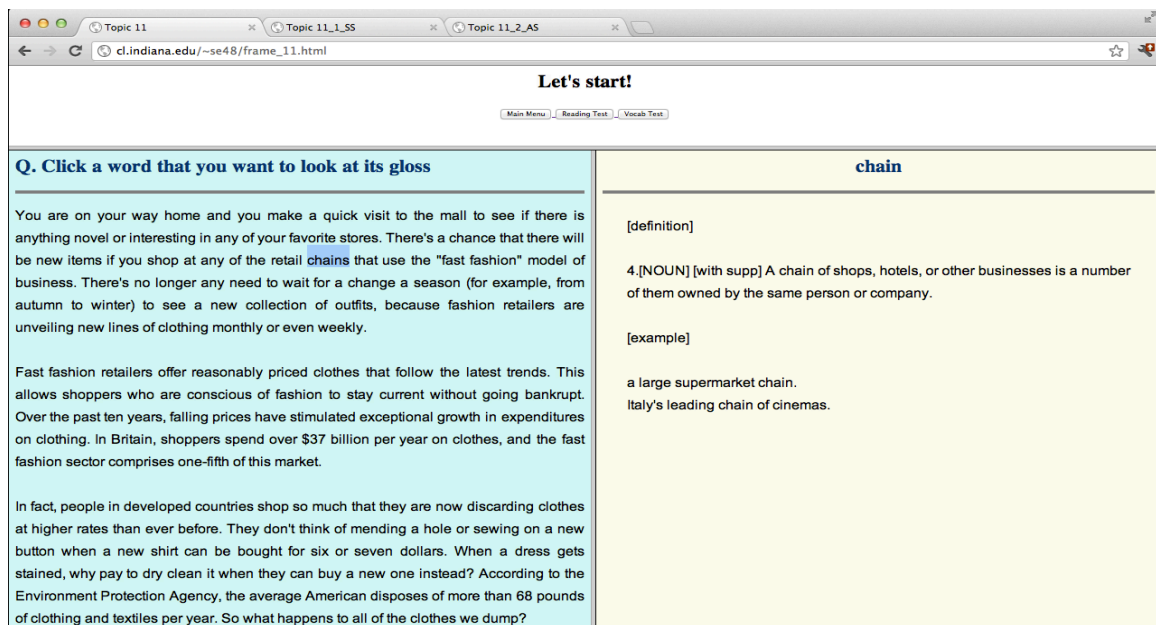


Figure 5.1. A screenshot showing lexical information, as presented to the GS group

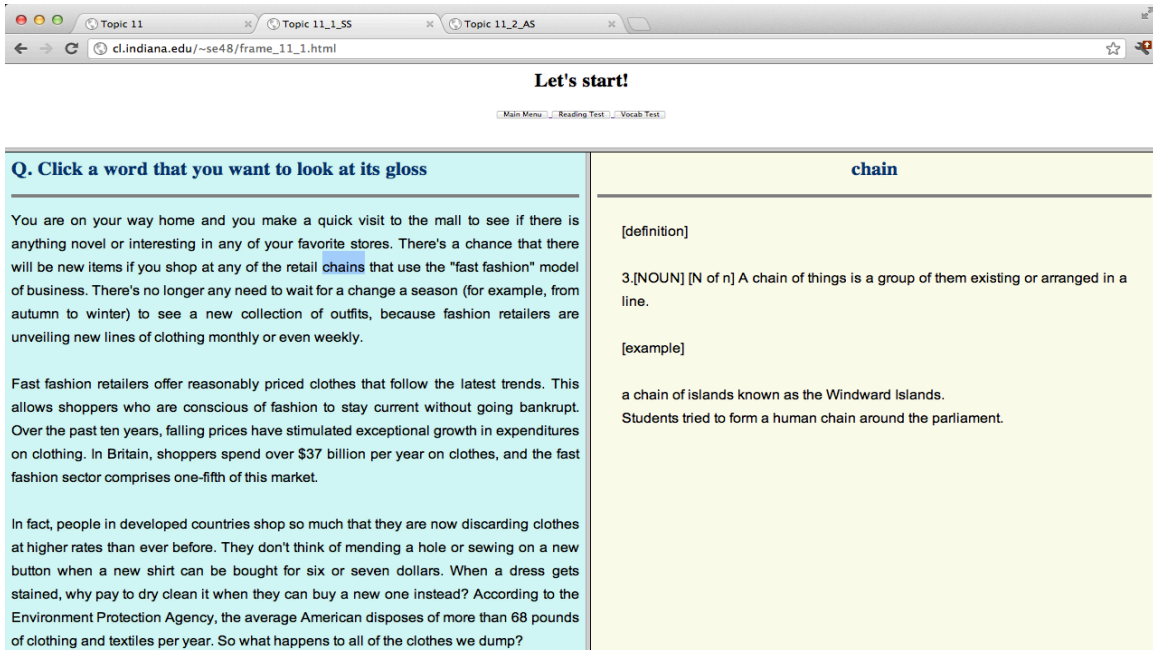


Figure 5.2. A screenshot showing lexical information, as presented to the SS group

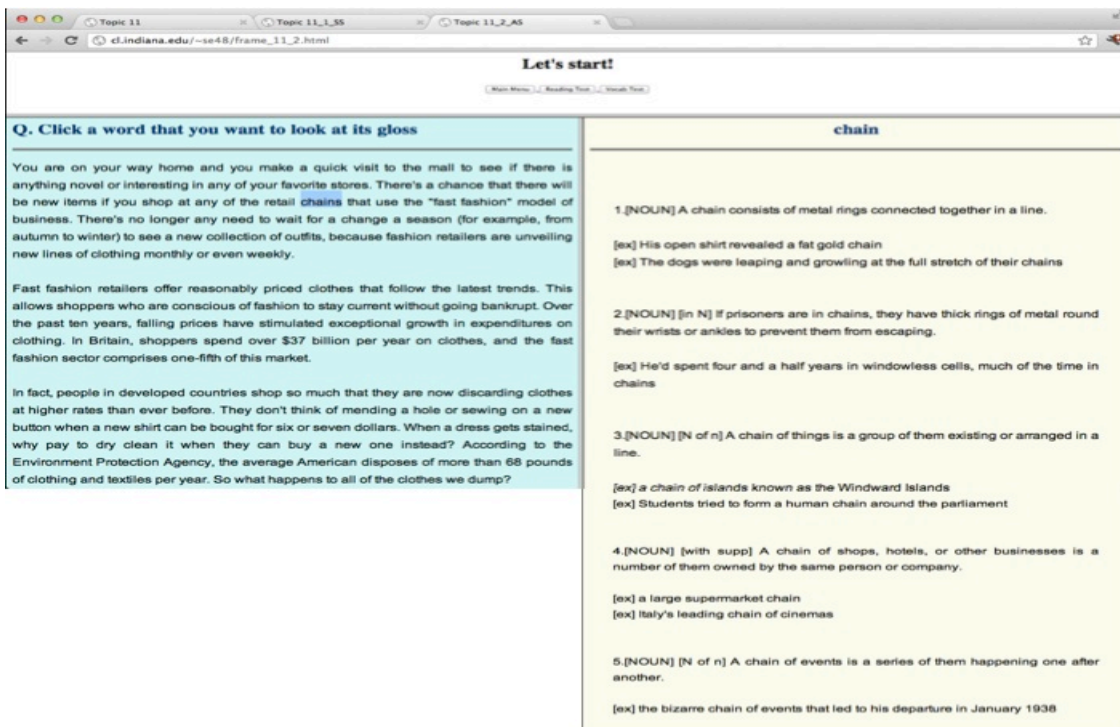


Figure 5.3. A screenshot showing lexical information, as presented to the AS group

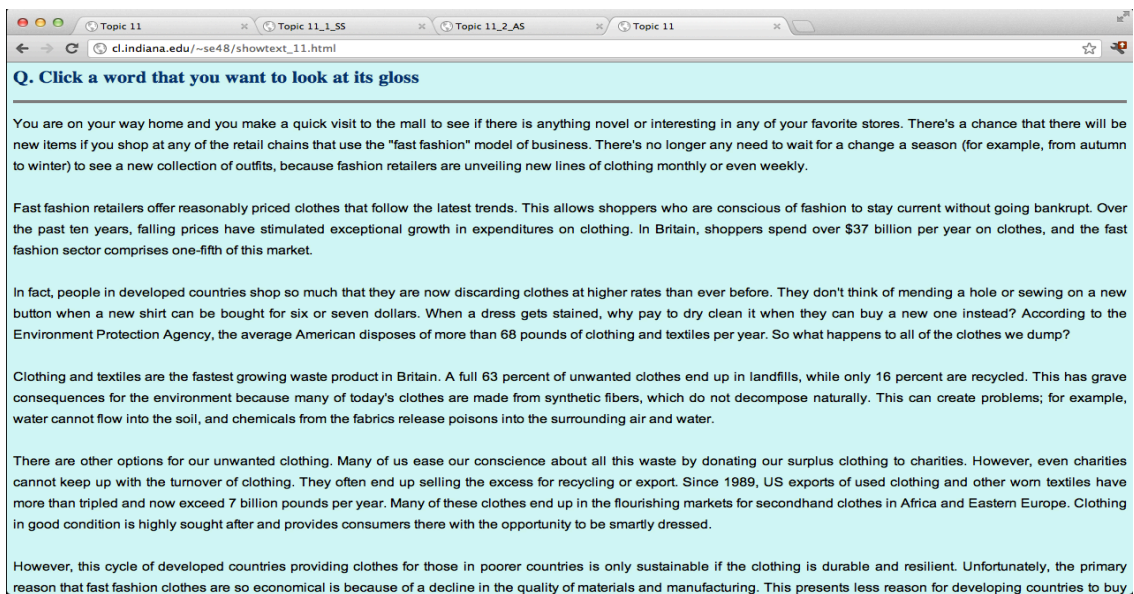


Figure 5.4. A screenshot showing no lexical information, as presented to the NS group

Figure 5.3 displays the lexical information of a word presented to the learners in the AS group when they clicked on the word; they are provided with all sense entries of the word. (COBUILD senses #1-#5). As stated previously, the NS group received no information, as shown in Figure 5.4.

Before going into the reading task page appropriately designed for each of the four groups (Figure 5.1 to 5.4), all participants started on the main menu, as shown in Figure 5.5. On the main page, each learner selected one of two reading texts listed under each title representing his/her group. For example, learners of the GS group chose one of the two texts in the list. Since the learners had to complete tasks for both of the two reading texts, which text they chose first did not really matter (however, for the purpose of balance, half of the group was guided to start with the first text and the other half with the second text – see section 5.2.3 for more details).

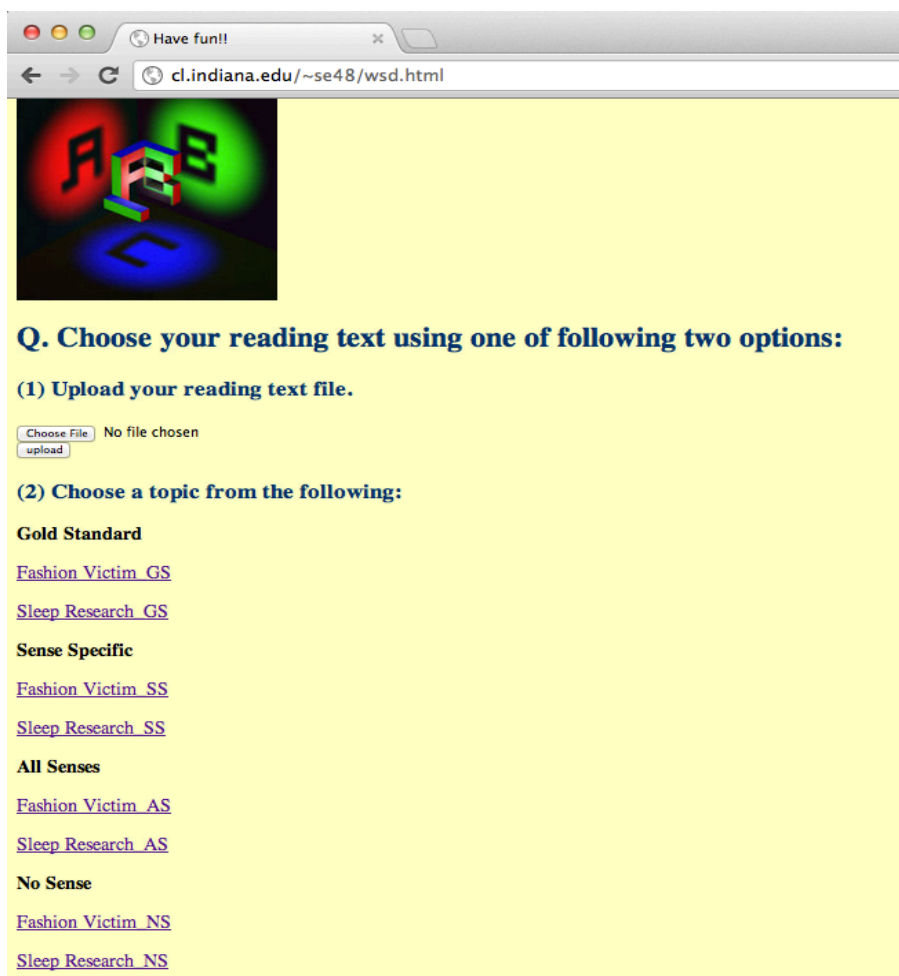


Figure 5.5. Screenshot of the main menu page

5.2.2 Materials

The materials used for the study were reading texts, target words, reading comprehension tests, vocabulary tests, and a user database.

5.2.2.1 Reading texts

Two reading texts for (high-) intermediate ESL/EFL learners and one reading text for SAT preparation were adapted in the first round by the author and another linguist, both of whom have extensive ESL/EFL teaching experience. According to their intuition and

experience, they deemed the level of those texts to be appropriate for (high-) intermediate ESL/EFL learners. The texts were also modified; they simplified syntax if it was too complicated, replaced words with synonyms²⁸ that had multiple senses, and shortened the length of texts down to about 600 words, as based on a previous study (Yanguas, 2009). All three texts were modified to be parallel. Two other TOEFL instructors further confirmed the appropriateness of the revised texts' difficulty level in terms of syntax and vocabulary.

The three texts were also piloted with 20 intermediate ESL/EFL learners who were not included in the main study. They were asked to circle the unfamiliar/unknown words, specify how interesting each passage was on a scale from 1 to 5, and indicate the difficulty of each passage on a scale from 1 to 5. Their reading time was also measured. Based on the results of the pilot study, two texts, *Fashion Victim* and *Sleep Research*, were finally selected. Students indicated that those two passages were more interesting and less difficult in terms of content. They took approximately 10-15 minutes for the learners read completely, which was expected as an appropriate reading time. *Fashion Victim* was adapted from *Focus on Vocabulary 1: Bridging Vocabulary* (Schmitt et al, 2011), and *Sleep Research* was adapted from *The Official SAT Study Guide* (The College Board, 2009). The lengths of the modified texts were 579 words (*Fashion Victim*) and 583 words (*Sleep Research*), respectively (see Appendix G).

As presented earlier in Chapters 3 and 4, the two selected texts were put into the NLP (natural language processing) server and processed to set all of their content words to be linked to the lexicon in the server such that any necessary lexical information (e.g. sense-

²⁸ The content words having only one sense were replaced with their synonyms that have multiple senses due to the purpose of the present study, which investigates the effectiveness of different kinds of lexical information (i.e. all senses vs. one sense appropriate for the context).

specific lexical information for the sense-specific group; lexical information of all senses of a given word for the all-sense group) could be presented to a learner by means of clicking on any content word. The two NLP processed texts were then uploaded online so that learners could access them from any computer with an Internet connection (See Chapter 3).

5.2.2.2 Target words

A total of 20 target words (nine words from *Fashion Victim* and 11 words from *Sleep Research*) were selected through the following steps:

(1) The author and another linguist replaced low-frequency content words that have only one sense with their synonyms that have multiples senses. Content words to needed have multiple senses in order for the study to be able to investigate the difference in vocabulary acquisition and reading comprehension among the groups (i.e., GS, SS) provided with sense-specific lexical information and the group provided with lexical information of all senses.

(2) All content words in the selected two texts were piloted with 20 intermediate ESL/EFL learners who were excluded from the main study. They were asked to circle all the words that they were not sure of or did not know. Based on the results, a list of words ranking from the most unknown/unfamiliar to the most known/familiar to those 20 intermediate learners was created.

(3) Finally, the target words were chosen from this list by fulfilling the required criteria, meaning that the words had to rank most unknown to the learners of the pilot

study and have multiple senses. Nine and 11 words were selected from the two texts, respectively, in order to balance the proportion of target words between texts.

All content words, including target words, in the texts were set to be clickable and, when clicked, available to present their lexical information. The target words consisted of nine nouns, seven verbs, three adjectives, and one adverb, as shown in Table 5.1.

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

Table 5.1. 20 target words

5.2.2.3 Reading comprehension tests

Reading comprehension tests were created by the author and another linguist who had extensive experience as ESL/EFL teachers, based on a set of reading comprehension questions that accompanied the selected texts. The original test sets had two types of questions: multiple-choice questions and true/false questions. The original test sets were modified by the author and the linguist while the format was maintained. The two new test sets thus consisted of four multiple-choice questions and six true/false questions (see Appendix H). Multiple-choice questions had either four or five choices per question and more than one answer could be selected; one question had one correct choice out of four choices, another question had two correct choices out of five choices, and the remaining two questions had three correct choices out of five choices.

Since participants were not allowed to refer back to the text to get help in answering the reading comprehension questions, the test questions focused more on general content rather than the details. The questions thus asked participants about their overall comprehension of the text, which aimed to measure how much they clearly understood its overall content.

The same pool of 20 ESL/EFL learners who were involved in determining reading texts and target words took these adapted reading comprehension tests. They confirmed that the level of difficulty was appropriate for intermediate level learners through their scores (the average score was 32.55 out of a total of 42 points, or an average of 77.5% which can be considered to be intermediate level) and answers to a questionnaire that asked about their perception of the difficulty of the test.

5.2.2.4 Vocabulary tests

There were one pretest and four immediate posttests (i.e., Post-1, -2, -3, and Post-4), one of which had the same format as the pretest (i.e., Post-3). The pretest and Post-3 had the same format in order to be comparable for examining the learning effect. The other three posttests (i.e. Post-1, -2, and Post-4) were intended to measure different attributes, which are explained later in this section. The pretest and all immediate posttests had the same 30 words, of which 20 were target words and 10 were distractor words. All of the 30 words were from the reading texts. Of the 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 were used with a sense that differed from the sense used in the reading

text (*deliberately.r, chain.n, outfit.n, mend.v, indicate.v*)²⁹. The composition of the part of speech (POS) of the distractor words was four nouns, four verbs, one adjective, and one adverb, which was proportional to the composition of the POS of the target words (i.e., nine nouns, seven verbs, three adjectives, and one adverb).

Initially, a test of 40 questions with 20 target words and 20 distractors was created by the author and another linguist, but when those 40 questions were piloted with another pool of 10 intermediate learners who were excluded from the main study, the author found that learners of the pilot test seemed to find the test extremely difficult, which resulted in some of the pilot learners giving up while taking the test. As such, the author reduced the number of distractors to 10. When the learners took the tests with 30 items, they still seemed to have difficulty, but they appeared to work on the tests without feeling extreme difficulty or giving up. Based on the author's judgment of the learners' performance, she decided to make vocabulary tests with 30 items.

The purpose of administering the pretest in the present study was to get a sense of the participants' prior knowledge of target words without cluing them in before receiving any treatments. In addition, the level of acquisition achieved through the treatments could be investigated by comparing the results of the pretest with those of the immediate posttest. In terms of the type of pretest, the production test type seemed to be too difficult for intermediate learners. Multiple-choice questions also did not seem appropriate, as either a question or a choice in a multiple-choice question would expose the definition of a given

²⁹ Since participants in the AS group were given lexical information of all senses, there was a chance that some of them might have chosen (and learned) the alternate sense definition for one of the target words. Including items that tested for alternate senses might allow for credit for having learned *something* even if it was not the definition that was used in the text, to make it more fair for them and to give credit where credit is due (; the idea was that they might have learned something different from that a test focusing on sense-specific definitions would not have picked up). The current study has not done this analysis; it is a possibility for future research.

word, and participants should not receive any clues about the definitions of target words, which could lead them to learn the words from the pretest. The author thus had to find a form of the tests (i.e., pretest and Post-3) overcame those limitations and found a suitable test format from a previous study. The present study adapted Kim's (2008) design, in that the posttest (Post-3) requires learners to fill in a blank for each sentence finding a word from a word bank (included both target words and distractor words).

In designing the pre/posttest in the present study, the idea first came from the fact that there are different levels of vocabulary knowledge; level 1 is simply recognizing that words are familiar or that a learner has seen them before (e.g., simple recognition) and a much higher level is being able to provide the definition or write a sentence with the word (i.e., production task). The posttest (Post-3) is in between these levels – e.g., the ability to recognize whether the word fits in a sentence. Since learners might not acquire the highest level of knowledge, the author needed to ensure that she tested for intermediate levels of knowledge in order to give them credit for whatever minimal knowledge they might have. If they are only tested on the hardest task (e.g., write a definition of a word, write a sentence with a word, etc.), the investigator might get the wrong impression that learners have no knowledge at all, but it is highly likely that the test was too difficult. In this respect, the posttest (Post-3) needs to be designed to require learners to show at least some of their knowledge. As stated earlier, since the pretest should not clue learners in to the target words, this study accordingly can only have comparative pre-post data for a fairly low level. In this respect, Kim (2008)'s idea of using actual sentences from the reading texts with blanks in them (for the posttest) seemed feasible for this study.

Since the format of the pretest and the posttest (Post-3) should be identical for the sake of comparability such that participants' gain score can be calculated (in order to examine the learning effect), the author decided to adapt the format as follows: the test consisted of three equivalent subsections. Each section had a word bank composed of 10 target words and five distractor words, as well as 10 sentential contexts, each of which had a blank to be filled with one of the words from the word bank (see Appendix I, for pretest; Appendix J, for Post-3). All 15 words in the word bank of each subsection thus played a role of distractors for each other.

Although the format of the pretest and the Post-3 should be identical for comparability, sentential contexts for the pretest and the Post-3 were made to be different. This was because although the pretest occurred two weeks before the posttest, it seemed likely that the same context would be remembered. Thus, the sentential contexts for the Post-3 came from the reading texts, following Kim's (2008) method, whereas the sentences for the pretest were taken from other sources such as dictionaries or were composed by the author and the linguist. By taking the pretest and the Post-3 as a learner, a third linguist reviewed and confirmed the reliability of those tests; that is, each of the questions had no other possible answers, enough contexts, and so on.

In addition to the Post-3 corresponding to the pretest, each learner took three additional posttests (i.e. Post-1, Post-2, and Post-4; see Appendix K), which could examine the different levels of vocabulary knowledge that learners could develop. The Post-1 had a list of 30 words consisting of 20 target words and 10 distractor words. It asked participants to mark which words they recognized as having occurred in the reading texts. This Post-1 was to see if they had recognized the target words at the most

superficial level. The Post-2 had the same list of words as the Post-1 and asked participants to provide definitions of these words. This test might have been considered the most difficult because it represented decontextualized production knowledge. Upon completion of the Post-3, participants took the Post-4, which had the same format as the Post-3 but provided additional information by including the definition of an answer word for a blank. Participants thus could find the answer for the blank by using the definition, the sentential context, or both. Participants could only see one test at a time and could not go back to the previous tests.

5.2.2.5 User database

It was necessary to record participants' performance while they were completing tasks in an unobtrusive way. The participants' basic information and all the words they clicked were recorded with the time in log files in the database built in this research (see Chapter 3). This database was later used to keep track of participants' look-up behavior and of which words had been clicked.

5.2.3 Procedure

The data was collected on two distinct days over a two-week period in a pretest-posttest design, as shown in Figure 5.6. The pretest was administered on a first meeting and the task and posttests were carried out on a second meeting that occurred two weeks after the first meeting. The experiment was conducted at three universities and a private institute using the same administrative procedure throughout.

At the first meeting, the author first explained the study clearly, and prospective participants signed an informed consent form (see Appendix L) if they agreed to take part in the study. They also answered a short questionnaire about their educational background. Upon completing the questionnaire, participants took the pretest, which assessed their knowledge of target words and thus allowed the author to find out the number of words with which they were familiar. Participants who knew more than 16 out of the 20 target words were excluded from the experiment. As previously stated, the results of the pretest excluded 28 potential subjects from the pool of participants, which resulted in a final pool of 60 participants. The total amount of time taken for the first meeting was approximately 20 to 30 minutes.

There was a gap of two weeks between the first meeting and the second meeting. Since it was almost impossible to identify words that none of the participants knew, as even the least-known words were still known by some of the learners of the pilot study, it was necessary to have a period of time between the pretest and the main task/posttest. By doing so, the experiment could avoid cluing the participants in to which words they were to be tested on. If the participants realized that they would be tested on the words they had just seen in the pretest, they would know on which words to focus. If this were the case, the experiment would not be truly testing how well they learn vocabulary while reading; rather, it would be testing how well they can memorize words on which they expect to be tested. On the other hand, the time of the pretest needed to be close enough to the time of the experiment (i.e. task/posttest) that the participants would be unlikely to learn the target words in the meantime through other means. If this were the case, the pretest would no longer be a valid measure of their knowledge. The author thus tried to

hold the pretest far enough in advance that the participants (1) perhaps would not realize that the pretest was part of the same experiment, (2) would not remember the target words, and (3) would not be specifically on the lookout for them. A period of two weeks was thus appropriate as the time between the pretest and the task/posttest.

The second meeting of the experiment took place in a computer laboratory setting. Each of the participants worked with the author individually at his/her own workstation. Participants were first familiarized with how to use the computer for this task³⁰. The three participant groups with support of lexical information were also informed that they were free to look up any words they wanted while reading, and were taught how to make the system present lexical information of the words that they chose. The participants in the group without support of lexical information were not allowed to look up lexical information while reading, so they were not informed about this possibility.

When participants were ready to take the task after receiving the instructions, they were given a first reading text (see Appendix G) on the computer. After reading the text, they were given an immediate posttest (see Appendix H) to assess their comprehension of the reading text. They were not allowed to refer back the reading text during the test. They did the same with the second reading text and reading comprehension test. Since two reading texts and tests were given, participants were counterbalanced in each group; half of the participants in a group worked with the reading text-1/test-1 first, while the

³⁰ The participants were introduced the process of the whole task as follows: (1) Open a web page – put their personal information, (2) Go to the main page – click the link that the author guided, (3) (For three groups) In the reading page, the author showed them how to use the mouse to click and get lexical information, as well as how to change the size of windows, (5) The author explained that when they were done reading, they would take a reading comprehension test, but could not refer back to the reading passage, (6) They were also instructed to take the second reading text and reading comprehension test in the same manner, (7) When finishing a reading task, they were going to take four vocabulary texts one by one, (8) For each vocabulary test, the author (pointing to instruction in the test sheet) verbally explained how to take it, though the instructions are also on the test sheet.

other half worked with the reading text-2/test-2 first³¹. This was due to the possibility of differences in a participant's concentration while doing the task (i.e., they are likely to show higher concentration in the beginning, which could lead to better performance initially, rather than later) Each reading task took about 15-20 minutes, and it took approximately 30 minutes for participants to complete the entire reading task.

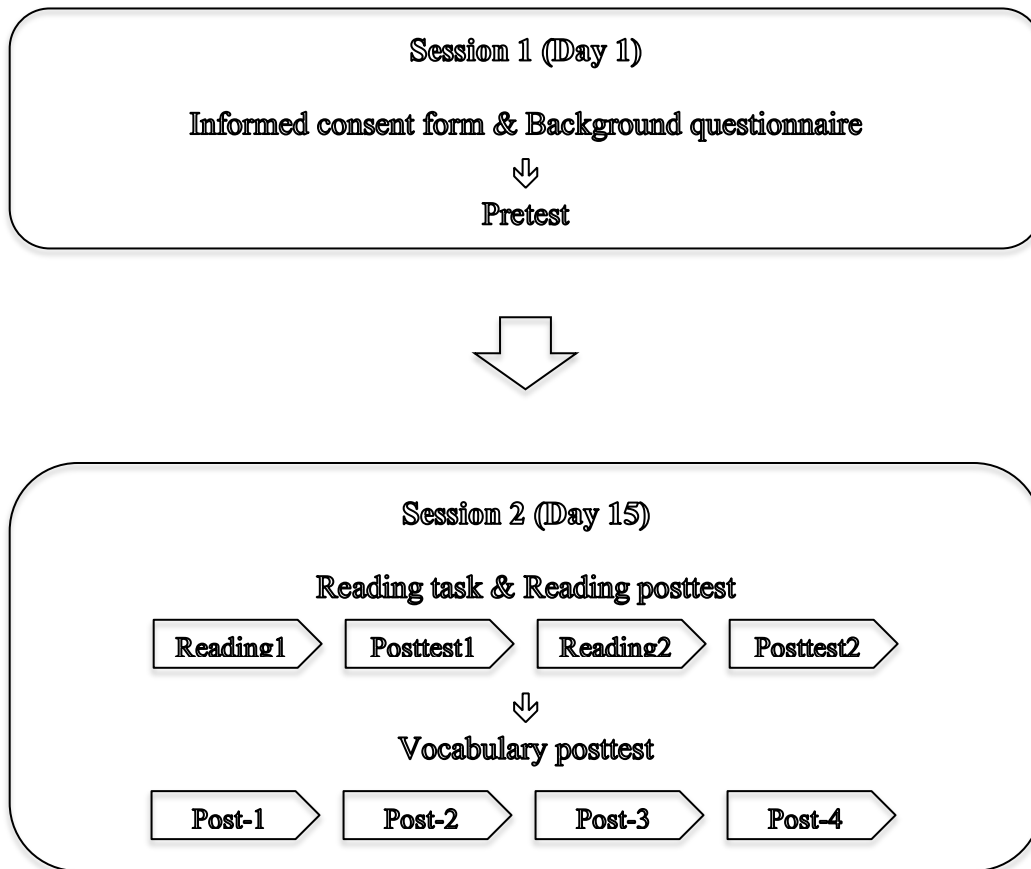


Figure 5.6. Procedure of the study

Upon completion of the reading task, participants moved on to a series of vocabulary posttests, from the Post-1 to the Post-4 in order, and they could not go back to a previous

³¹ The author guided participants one by one; if one participant took the reading text-1/test-1 first, then the author guided the next participant to take the reading text-2/test-2 first, and so on.

test. Using the same list of the words, participants were asked to circle the words that they had seen while reading the text in the Post-1, and to describe their definitions in the Post-2. They then went on to the Post-3 (whose format was the same as the pretest's) to fill in the blank using a word from a word bank. Finally they took the Post-4, which was the same as the Post-3 but with the additional information of the definition of the words in the word bank. Participants took approximately 30 minutes to complete all vocabulary tests. The second session was thus completed within about one hour.

5.2.4 Scoring

5.2.4.1 Reading comprehension test

This study had two reading comprehension tests corresponding to each text. Each test had 10 questions with a total possible score of 21 (42 for both tests). The test consisted of four multiple-choice questions (MCQ) and six true/false questions. For the MCQs, participants had to choose *all* and *only* the choices that were correct,³² and the total possible score MCQs was 15 points. One MCQ had only one answer and was worth a score of 1 for a correct answer; whereas the other three MCQs had more than one correct answers so every choice in each of those three MCQs was treated as a separate item and was worth a score of 1 or 0 (i.e., one question had four choices (4 points); two questions had five choices ($2*5 = 10$ points)). As one can see in the example below, if a question has five choices, then the maximum point for the question is 5 points. Accordingly, if a participant marked one correct choice and one wrong choice out of five, he would receive 3 points for the question, as shown in the following.

³² The participants were informed that there were more than one correct answer for MCQs (see Appendix H).

Question & correct answer	Subject answer	Scoring
a.	a. → chose	a. wrong → 0
b. → correct choice	b. → chose	b. correct → 1
c. → correct choice	c.	c. wrong → 0
d.	d.	d. correct → 1
e.	e.	e. correct → 1

The section of true/false questions consisted of six questions. 1 point was awarded for each correct judgment and 0 points for an incorrect judgment or a blank answer. The total points possible for the true/false questions was thus 6 points.

5.2.4.2 Vocabulary test

Each vocabulary test consisted of 30 items, which included 20 target word items and 10 distractor word item. Each item was worth 1 point for a correct answer, so the maximum possible score was 30. In scoring the pretest, the Post-3, and the Post-4, 1 point was awarded for a correct answer and 0 points for an incorrect answer or for leaving it blank. For Post-1, participants got 1 point for correctly identifying words presented in the reading texts that they had read and 0 points for leaving an item unmarked. For the Post-2, participants were awarded 1 point for describing/writing a definition for each of the words in the list. There were two ways of scoring the Post-2; restricted and released. In the restricted way of scoring, participants were awarded 1 point for a correct definition that was used in the context of the reading text; otherwise, they received 0 points. In the released way of scoring, participants got a point not only for the definition used in the reading text but also for other possible definitions of the word. For instance, for the word

cradle, if participants described it as *a baby's bed with high sides*, they received 0 points in the restricted way because it was not the meaning used in the reading text, whereas they received 1 point in the released way because they had used one of the possible meanings of *cradle*.

5.2.5 Data analysis

As presented earlier, the two research questions of the present study asked if sense-specific lexical information contributed to better vocabulary acquisition and reading comprehension. In order to examine these two research questions, the data was first statistically analyzed using SPSS, version 20.0 (<http://www-01.ibm.com/software/analytics/spss/>).

As different groups (e.g., GS, SS, AS, and NS) were being compared, for every statistical test, the Test of Homogeneity of Variances (i.e., Levene's Test) was first run to see if the variances of the error between groups are equal at the outset of the study. Without this prior test, it would not be clear to judge whether the significant differences that resulted from the main statistical tests were caused by the actual effect of the variables of interest (e.g. reading comprehension test scores, vocabulary test scores, etc.) or by inherent differences between groups (Larson-Hall, 2010). The significance level³³ was set at 0.05 for all statistical tests in this study.

³³ The significance level (i.e., alpha value) is a chance that the results happened by chance. So if the alpha value is 0.001, it means that there is 0.001% that the results happened by chance, meaning that the results are significantly different. An alpha level of 0.05 is a standard level for statistical analysis.

5.2.5.1 Vocabulary acquisition

The first research question was regarding whether participants show better vocabulary acquisition when given sense-specific lexical information while reading. In order to answer this question, an analysis was conducted to examine if there were statistically significant differences among the groups, each of which was provided with different lexical information (e.g., GS, SS, AS, and NS), in terms of the amount of vocabulary acquisition (e.g., pre-post gain). Accordingly, it is necessary to examine the participants' performance on the pretest and the posttest (i.e., Post-3).

A repeated-measures analysis of variance (RM ANOVA) is applicable when the same participants are tested more than once. A RM ANOVA was therefore utilized because the data in this analysis were collected from the same participants at two different time periods (e.g. pretest, Post-3). A RM ANOVA for this analysis included all participants, with Time (pretest, Post-3) as the within-subject variable, and Group (GS, SS, AS, NS) as the between-subject variable. The within-subject variable (Time) was used to explore how an individual participant, their assignment in a group aside, differed by Time. The between-subject variable (Group) was used to explore how each group was different, time aside, in the pretest and the posttest (Post-3), respectively. In case significant results were found, post-hoc pairwise comparisons were accompanied to determine which groups differed from each other.

It was necessary to check if there were any potential intervening variables in the results above. One possible variable that might have affected the results would be learners' prior vocabulary knowledge. For example, if some learners' prior vocabulary knowledge is too high (e.g., if they already knew 18 of the 20 target words), they have a

limit to show their improvement (e.g., they have *room* for only two words to be learned) from the treatment. Thus, their gain scores may have suppressed the average gain scores for their group. In this sense, comparing the performance of the *High* and *Low* learners³⁴ from each group would address the variable of prior vocabulary knowledge. Accordingly, the data was analyzed with the variable of learners' prior vocabulary knowledge.

Thus, the subjects in each group (i.e., GS, SS, AS, NS) were divided into a High group and a Low group based on their prior knowledge of the target words (i.e., their pretest scores). The subjects above the mean pretest score were assigned into the High group and the subjects below the mean pretest score were assigned into the Low group. Subjects in between +1 and -1 standard error of the mean (S.E.M.) were excluded from the group assignment, which gives us a 95% chance (i.e., confidence) that the true score of subjects in the Low group are below those subjects in the High group, and vice versa. After excluding subjects in between ± 1 S.E.M., for the GS group, the High group had six subjects and the Low group had five; in the SS group, eight were in the High group and five in the Low group; in the AS group, eight were in the High group and six in the Low group; in the NS group, six were in the High group and six in the Low group.

Two-way repeated-measures ANOVA was run for this analysis on pretest and posttest scores as a dependent variable; group (GS, SS, AS, NS) and prior vocabulary knowledge (High and Low) are two independent variables for this analysis. In case significant results were found, post-hoc pairwise comparisons were conducted to compare groups in pair.

Looking at the data in a more fine-grained way, there were the data of the pretest and the posttest in terms of only the target words that had been *clicked* by participants (GS,

³⁴ The *High* learners were those who knew more of the target words before the treatment, whereas the *Low* learners knew fewer words to start with.

SS, AS) while reading. A RM ANOVA was applied to analyze the pretest and posttest (pre-post gain) among *clicked* words.

In addition to the major analysis of participants' vocabulary acquisition, a series of further analyses was performed on scores of additional vocabulary posttests (e.g., Post-1, Post-2, Post-4) to investigate if there were any statistically significant differences in different kinds of vocabulary knowledge development³⁵ (i.e. Post-1, Post-2, Post-4) that participants demonstrated among groups. Since there was no comparable pretest corresponding to each of those posttests, a one-way ANOVA was performed on the scores of each of the Post-1, Post-2, and Post-4 as dependent variables, with group as an independent variable. Post-hoc comparisons were run if there were significant differences among groups on scores of the Post-1, Post-2, and Post-4, respectively.

5.2.5.2 Reading comprehension

The second research question was regarding whether sense-specific lexical information provided in reading would facilitate participants' reading comprehension. In an attempt to answer this question, the data was analyzed to investigate if there were statistically significant differences among groups in terms of their reading comprehension scores.

When the study examines three or more groups to see the effect of groups, ANOVA is used. When the experiment analyzes only one dependent variable and one independent variable, a one-way ANOVA is utilized. A one-way ANOVA was therefore performed with reading comprehension scores as a dependent variable and the four groups as an

³⁵ For example, some learners show more development in production knowledge while some show more in recognition knowledge.

independent variable in order to explore if there was any significant main effect of group on reading comprehension scores (i.e., differences in reading comprehension scores generated from the four groups). In case there were significant differences found, post-hoc analyses were conducted to determine any significant differences by comparing groups in pairs.

Reading comprehension scores of the High group and the Low group were also compared to see if learning new words affects reading scores differently than already knowing the words. For this analysis, two-way ANOVA was run on reading comprehension scores as a dependent variable, with group (GS, SS, AS, NS) and prior vocabulary knowledge (High/Low) as two independent variables. Post-hoc comparisons were run if there were significant differences among groups on reading comprehension scores.

5.2.5.3 Effects of the system errors

Among the four groups of participants in this study, there were two groups (e.g. the GS and SS groups) assigned as sense-specific groups, which was due to a computational reason in that the system built in this research did not perform completely correctly. Computational researchers have tried to implement a system as close as to the gold standard in performance; however, every system has some performance errors. The current system built in this research would also possibly provide sense-specific information that was not appropriate to its context in the reading text.

Thus, this study tried to investigate if system errors affect learners' vocabulary acquisition by analyzing the performance of the SS group. Among the groups, the SS

group alone had the chance to see an inappropriate target-sense (; the GS group always received correct information, the AS group received information for all senses, and the NS group did not receive any information). Thus, only the SS group was related to this analysis.

The author first discovered all target words that were presented with inappropriate sense-specific information (= inappropriate target-sense words). Among the inappropriate target-sense words, the author located all words that participants in the SS group clicked from log files in the user database (=inappropriate target-sense words *clicked*). Among the inappropriate target-sense words clicked, participants who clicked these words and who were wrong in the pretest were separated from those who were correct in the pretest. Finally, among participants who were wrong in the pretest on inappropriate target words clicked, participants who were wrong in the posttest (Post-3) and who were correct in Post-3 were further divided. The complete division of the cases is shown in Table 5.2. All cases were tallied and the relative frequencies of all of the final eight cases (the right-most column) were calculated. Vocabulary acquisition of the eight cases was finally analyzed.

Clicked while reading	Pretest	Post-3
inappropriate target-sense words	incorrect	correct
		incorrect
	correct	correct
		incorrect
appropriate target-sense words	incorrect	correct
		incorrect
	correct	correct
		incorrect

Table 5.2. All cases of participants' performance in the SS group

5.3 Results and discussion

This section presents the results of the data analyses and discusses the findings in detail.

5.3.1 Vocabulary acquisition

As stated previously, it is necessary to test the participants' homogeneity of variance prior to the experiment. Since the first research question is to examine the improvement between the pretest and the posttest (i.e., Post-3), the test of homogeneity of variance was carried out to ensure that the pretest/Post-3 scores of the participants across the four groups showed similar variances, indicating that all participants are homogeneous at the beginning of the experiment. Levene's test of homogeneity of variance was used to check equal variances on the pretest/Post-3 of all participants across the four groups. As shown in Table 5.3, the four groups can be considered to have similar variances on both the pretest ($F(3, 56) = 0.49, p = .69$) and the Post-3 ($F(3, 56) = 0.13, p = .94$), meaning that

this assumption underlying the use of ANOVA was met (Larson-Hall, 2010). The error variances of all four groups were therefore considered equivalent before they were compared, and it can be said that any significant differences found in the analysis are attributable to the actual effect of the variables (e.g. Time, Group, Time*Group), and not to inherent differences in the error variances of the groups.

Levene's Test of Homogeneity of Variances				
Test	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i> -value
Pretest	.49	3	56	0.69
Post-3	.13	3	56	0.94

Table 5.3. Test of Homogeneity on pretest/Post-3 for the four groups

The data collected from all four groups over two test sessions (pretest and posttest) were analyzed, and the descriptive statistics for the participants' performance are shown in Table 5.4. The mean scores (converted to a percentage out of 100) of participants' pretest/Post-3 and their gains from the pretest to the Post-3 are displayed graphically in Figure 5.7 and Figure 5.8, respectively.

	Pretest				Post-3				Gain	
	<i>M</i>	<i>Min</i>	<i>Max</i>	<i>SD</i>	<i>M</i>	<i>Min</i>	<i>Max</i>	<i>SD</i>	<i>M</i>	%
GS (n=15)	10.73	2	15	3.43	15.93	4	20	3.96	5.2	26
SS (n=15)	10.93	3	14	2.82	15.47	4	19	3.80	4.54	23
AS (n=15)	10.87	5	15	3.34	13.47	5	18	3.83	2.6	13
NS (n=15)	10.87	3	15	3.25	11.27	4	16	3.39	0.4	2

Table 5.4. Descriptive statistics of pretest and posttest (Post-3) scores across the four groups

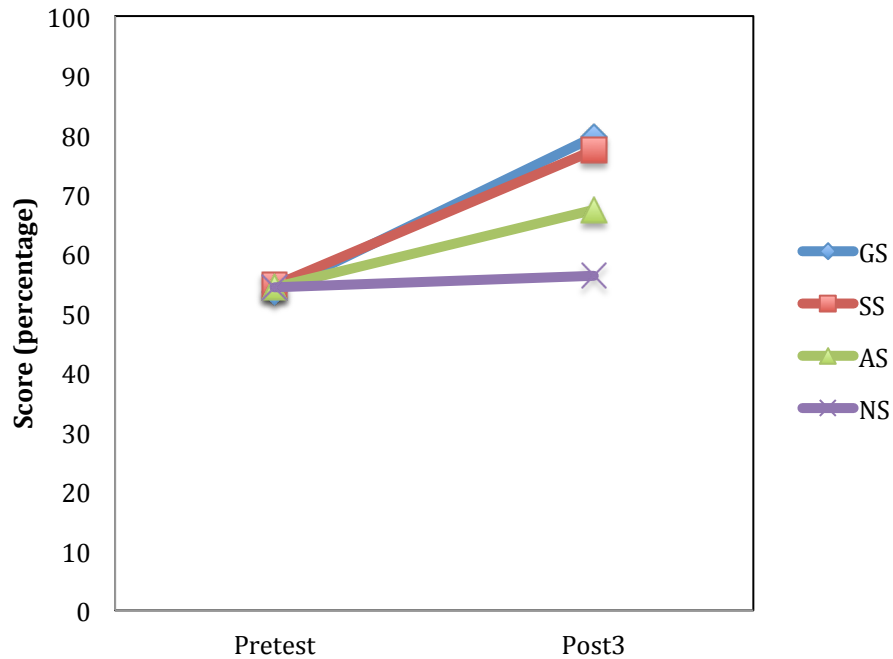


Figure 5.7. Scores on the pretest and the post-3 (in percentage)

As shown in Table 5.4 and Figure 5.7, the difference among the pretest mean scores of the four groups was within less than 1 point (SS: 10.93, GS: 10.73), which is about a 1 percent difference (SS: 54.65%, GS: 53.65%). This confirmed the results of the group homogeneity test, which suggested that the four groups were comparable. On the other hand, the four groups showed larger differences in their Post-3 results, as presented in Figure 5.7; the GS and SS groups showed the clearest gains, suggesting greater vocabulary acquisition than the AS and NS groups, which was expected (GS: 79.65%, SS: 77.35%, AS: 67.35%, NS: 54.35%). Vocabulary acquisition of each of the groups from the pretest to the Post-3 is presented in Figure 5.8, which shows that the GS group and the SS group gained 26% and 23%, respectively, while the AS group gained only 13% and the NS group 2%. Based on their mean scores on the pretest and the Post-3, the trend of their performances demonstrates what this study expected for the first research

question: the GS and SS groups show more vocabulary acquisition than do the AS and the NS groups.

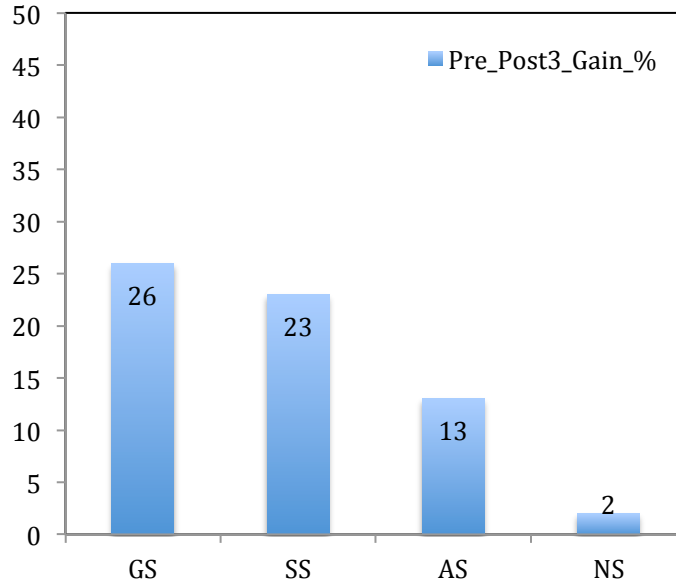


Figure 5.8. Pre-post-3 gain in percentage

In order to examine if the above differences among the groups were statistically significant, a repeated-measures ANOVA was run on the pretest and Post-3 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 5.5.

Source	<i>df1</i>	<i>df2</i>	Mean Square	<i>F</i>	<i>P</i> -value	Partial Eta ²	Obs. Power
Tests of Within-Subjects Effects							
Time	1	56	304.01	62.67	.00	.53	1.00
Time*Group	3	56	34.94	7.20	.00	.28	.98
Test of Between-Subjects Effects							
Group	3	56	33.36	1.71	.18	.08	.42

Table 5.5. Results of RM ANOVA comparing vocabulary test scores across the four groups over time

With respect to the within-subject variable (Time), the effect of Time showed a statistically significant difference ($F(1, 56) = 62.67, p < .001, \text{partial } \eta^2 = 0.53$). In other words, not considering Group, there is evidence of improvement from the pretest to the Post-3.

Most crucially related to the first research question of whether the groups would have different amounts of vocabulary acquisition over time, the analysis manifested a statistically significant difference on the effect of Time by Group interaction (Time*Group; $F(3, 56) = 7.20, p < 0.001, \text{partial } \eta^2 = 0.28$). The partial Eta² (η_p^2) for Time (0.53) and Time*Group (0.28) in Table 5.5 represented a large and medium effect size, which thus provided relatively strong evidence for the differences³⁶.

In order to locate where significant differences exist on the effect of Time*Group, two sets of post-hoc comparisons were conducted³⁷. The first comparisons were run to find if there was a significant difference on the mean differences (= improvement from the pretest to the Post-3) of each group. As illustrated in Table 5.6, three groups (GS, SS, AS) showed significant mean differences between their the pretest and the Post-3 ($p < 0.05$). Additionally, the GS group and the SS group had higher improvement than the AS group. No significant difference was observed on the mean difference of the NS group ($p = .62$). This was already pointed out in Figure 5.7, which displays very little difference in the trajectory of the NS group from the pretest to the Post-3. The results indicated that the

³⁶ According to Cohen (1992), the effect sizes of 0.10, 0.25, and 0.40 are considered small, medium, and large respectively F-Test (ANOVA).

³⁷ Instead of running six t-tests (GS-SS, GS-AS, GS-NS, SS-AS, SS-NS, AS-NS) to locate difference of Time*Group, people usually run multiple t-tests, but the study approached this differently, being more efficient in time/labor. Also less statistical tests are less error-prone on the results (Larson-Hall, 2010).

three groups who received lexical information showed improvement whereas the group who received no information did not.

Group	(I)Time	(J)Time	Mean	Std. Error	<i>P</i> -value
			Difference(I-J)		
GS	pretest	Post-3	-5.20*	.80	.00
SS	pretest	Post-3	-4.53*	.80	.00
AS	pretest	Post-3	-2.60*	.80	.00
NS	pretest	Post-3	-.40	.80	.62

* The mean difference is significant at the 0.05 level

Table 5.6. Mean difference between pre-post3 for each group

A second set of post-hoc tests were then run to compare the groups in pair in terms of the amount of the mean pre-post gains. In Table 5.7, the Contrast Estimate looked at the differences in the mean pre-post gains. According to it, the GS group is significantly different from the AS group ($p = 0.02$) and from the NS group ($p < 0.001$), and the SS group is significantly different from the NS group ($p = 0.001$). All the other group comparisons in pair (i.e., GS-SS, SS-AS, AS-NS) showed non-significant differences in their mean gains.

In conclusion, these post-hoc comparisons on the Time*Group interaction effect found a significant difference between the GS and AS groups and between the GS and NS groups in their vocabulary learning over time, with the GS group showing greater pre-post improvement.

Group Special Contrast		Dependent Variable
		Mean Difference
GS-AS	Contrast Estimate	2.60
	Sig.	0.02
GS-SS	Contrast Estimate	0.67
	Sig.	0.56
GS-NS	Contrast Estimate	4.80
	Sig.	0.00
SS-AS	Contrast Estimate	1.93
	Sig.	0.09
SS-NS	Contrast Estimate	4.13
	Sig.	0.001
AS-NS	Contrast Estimate	2.20
	Sig.	0.06

* The mean difference is significant at the 0.05 level

Table 5.7. Contrast Results for the amount of the mean pre-post gains

While the analysis showed significant effects for Time and Time*Group (=within-subject effect), it revealed that there was no statistically significant Group effect ($p = 0.18$). This means the effect of individual variation (i.e., within-subject) is more than the effect of what group they are in (i.e., between-subject). This was anticipated to answer the research question about if there are differences among groups with respect to time.

The data was further analyzed with a new variable of subjects' prior vocabulary knowledge; the subjects of each group were divided into High (H) and Low (L) groups based on their pretest scores, and the performances of the High group and the Low group from each of four groups were compared.

At first, Levene's test of homogeneity of variance was run to make sure that the scores of the pretest and the Post-3 of all subjects across groups (i.e., H/L in GS, H/L in SS, H/L in AS, H/L in NS) showed equal variances. As shown in Table 5.8, all groups have similar variance on both the pretest ($F(7, 42) = 2.17, p = .06$) and the Post-3 ($F(7, 42) =$

1.15, $p = .35$), indicating that all subjects were homogeneous at the beginning of the experiment.

Test	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i> -value
Pretest	2.17	7	42	0.06
Post-3	1.15	7	42	0.35

Table 5.8. Test of Homogeneity of variances on pretest/Post-3 for all groups

The descriptive statistics on the pretest and the Post-3 for all eight groups are shown in Table 5.9, and the subjects' gain scores (in percentage) from the pretest to the Post-3 are displayed in Figure 5.9.

		Pretest				Post-3				Gain
		<i>M</i>	<i>Min</i>	<i>Max</i>	<i>SD</i>	<i>M</i>	<i>Min</i>	<i>Max</i>	<i>SD</i>	<i>M</i>
GS	H (n=6)	13.67	12	15	1.37	17.83	14	20	2.14	4.17
	L (n=5)	7.00	2	9	2.92	14.00	4	20	5.96	7.00
SS	H (n=8)	12.75	12	14	0.71	15.75	13	19	1.83	3.00
	L (n=5)	8.00	3	10	3.08	14.80	4	19	6.26	6.80
AS	H (n=8)	13.50	12	15	0.93	14.50	7	18	3.51	1.00
	L (n=6)	7.33	5	10	1.97	12.83	5	17	4.22	5.50
NS	H (n=6)	13.67	12	15	1.97	13.17	7	16	3.55	-0.50
	L (n=6)	8.00	3	10	2.53	8.83	4	12	2.71	0.83
Total	H (n=28)	13.36	12	15	1.25	15.29	7	20	3.15	1.93
	L (n=22)	7.59	2	10	2.46	12.45	4	20	5.09	4.86

Table 5.9. Descriptive statistics of pretest and posttest (Post-3) scores across the eight groups

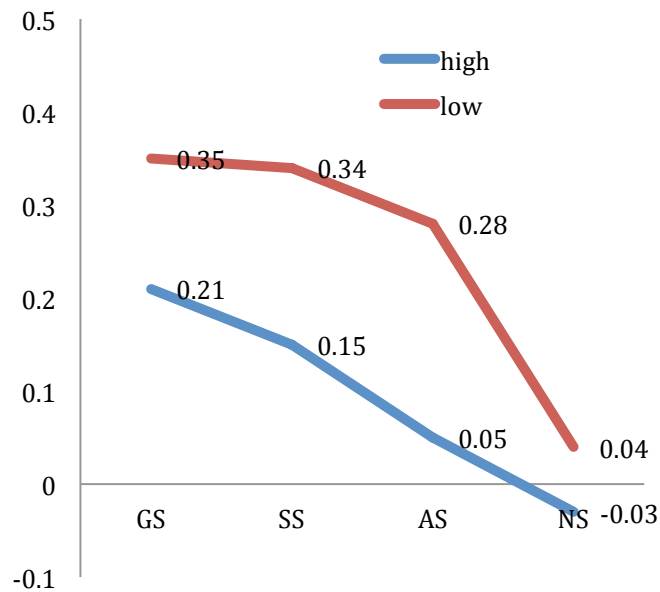


Figure 5.9. Pre-post gain scores for H and L across the four groups

As one can see in Table 5.9 and Figure 5.9, the Low group showed more improvement than did the High groups because they had more room for improvement. This indicates that the cut-off score (i.e., 16/20) from the first analysis (see Table 5.4 and Figure 5.7. for the average gain scores from the pretest to the Post-3 across the four groups) had a ceiling effect and the gain scores of the High group seemed to affect (suppress) the average gain scores for their respective group. At the same time, the Low group and the High group both showed the same pattern of average gain scores – i.e., $GS > SS > AS > NS$. Since the High and Low groups are a subset of their respective groups and therefore contributed to the average gain scores of each group, the similar patterns are as expected.

Two-way repeated-measure ANOVA was run to see if the results have statistically significant differences, and results from the test are in Table 5.10.

Source	<i>df1</i>	<i>df2</i>	Mean Square	<i>F</i>	<i>P</i> -value	Partial Eta ²	Obs. Power
Tests of Within-Subjects Effects							
Time	1	42	293.48	72.20	.00	.63	1.00
Time*Group	3	42	34.24	8.42	.00	.38	.99
Time*HL	1	42	59.02	14.52	.00	.26	.96
Test of Between-Subjects Effects							
Group	3	42	22.83	1.57	.21	.10	.38
HL	1	42	439.85	30.27	.00	.42	1.00

Table 5.10. Results of two-way RM ANOVA comparing vocabulary test scores across the four groups and the H/L groups over time

As one can see in Table 5.10, the effect of Time showed a statistically significant difference ($F(1, 42) = 72.20, p < .001, \text{partial } \eta^2 = 0.63$). More crucially, the effect of Time by Group interaction (Time*Group) showed a significant difference ($F(3, 42) = 8.42, p < .001, \text{partial } \eta^2 = 0.38$). As stated earlier, because H and L are subsets of their respective groups, the similar results (i.e., significant difference on the effect of Time*Group) are as expected. With a new variable, the High (H) and Low (L) groups, the effect of Time by HL interaction (Time*HL) also showed a significant difference ($F(1, 42) = 14.52, p < .001, \text{partial } \eta^2 = 0.26$). This indicated that H and L showed significant difference in terms of their pre-post improvement. In addition, H and L also showed a significant difference ($F(1, 42) = 30.27, p < .001, \text{partial } \eta^2 = 0.42$), regardless of Time, whereas Group did not show a significant difference ($F(3, 42) = 1.571, p = .21, \text{observed power} = 0.38$).

In sum, the further analysis (with the new variable of prior vocabulary knowledge – H and L in each of four groups) confirmed the results of the first analysis (based on the four

groups) on vocabulary learning. That is, the High and the Low groups across the four groups respectively showed a ranked order as GS<SS<AS<NS in vocabulary learning, which is the same pattern as that of overall gain scores across the four groups. Additionally, the Low group showed more improvement than did the High group, because they have more room for improvement/learning.

In addition to the analysis of learners' performance on the overall scores of their pretest and Post-3, a fine-grained analysis was conducted on their performance on the pretest and the Post-3, looking only at words they *clicked* while reading, as well as how much they clicked. That is, the focus of analysis was restricted to the words that participants *clicked* during reading. First, all words clicked and all *target* words clicked during reading were investigated, as presented in Figure 5.10. Since the participants (n=15) in NS group were not allowed to click on words and get lexical information, they were excluded in this analysis, which was pertinent only to clicked target words.

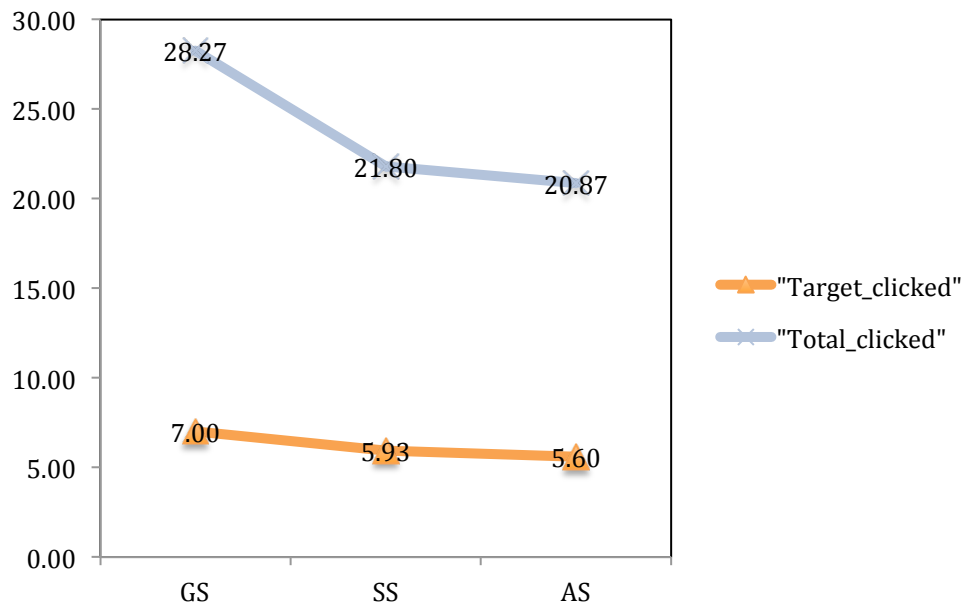


Figure 5.10. All words clicked vs. all target words clicked

As shown in Figure 5.10, the GS group clicked 28.27 words on average (7.00 target words), the SS group clicked 21.80 words (5.93), and the AS group clicked 20.87 words (5.60). The apparent trend may suggest that the GS group realized that they could get high-quality lexical information from clicking on words and thus clicked more often. Although the SS group was provided with the same kind of information (i.e. sense-specific information appropriate for the context of the reading text) as the GS group's, they showed a tendency of clicking words less frequently. They may have perceived that the current system presenting lexical information to the SS group could give inappropriate or empty information, which may have hindered them from clicking words. Meanwhile, the participants in the AS group demonstrated a descriptively lower trend of clicking than the GS and SS groups. They seemed to have realized that the lexical information yielded from clicking required them to do more work to receive helpful information for reading comprehension. That is, they had to determine which word sense was appropriate for the context of the reading text among multiple sense entries with which they were provided. Overall, the provision of convenient and helpful information seemed to induce participants to click more.

While the results descriptively showed the GS group clicked more than the SS and AS groups, the statistical analysis indicated that group differences were not statistically significant. A one-way ANOVA was run on all words clicked and all *target* words clicked, respectively, in order to examine group differences. The effect of Group on both all words clicked ($F(2, 44) = .91, p = .41$) and on all target words clicked ($F(2, 44) = .60, p = .55$) did not show a statistically significant difference.

In particular, for the target words, each of the four groups was further divided into the H and the L groups, and their clicking patterns were analyzed. In general, the High group showed less clicking than did the Low group, as shown in Figure 5.11. On average, the Low group clicked nearly twice as many words as did the High group across the three groups. At the same time, the Low group showed a different pattern ($GS > AS \cong SS$) from that of the average clicking pattern on target words across three groups ($GS > SS \cong AS$, Figure 5.10), while the High group showed another pattern ($GS \cong SS > AS$) that also differed from that of the average clicking pattern on target words.

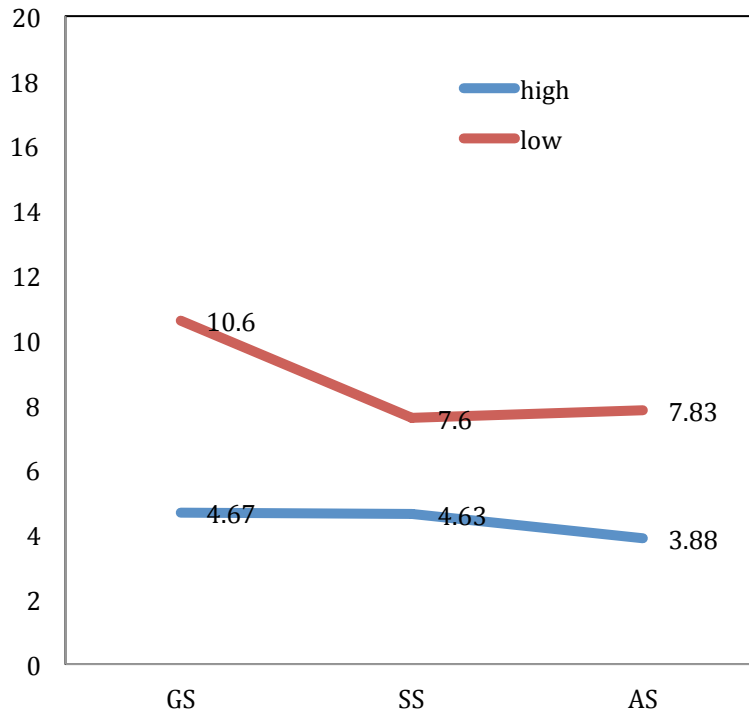


Figure 5.11. Target words clicked for H and L across three groups

In order to see if the results had statistically significant differences, two-way ANOVA was run with a dependent variable (i.e., the number of clicked target words) and two independent variables (i.e. Group & High/Low). While Group did not show a statistically

significant difference ($F(2, 32) = 0.998, p = .38$, observed power = 0.21), H/L showed a statistically significant difference ($F(1, 32) = 15.60, p < .001$, partial $\eta^2 = 0.33$). This suggests that for the tendency of clicking target words, the Low group clearly clicked more target words than did the High group did.

Next, among the target words that the participants clicked during the reading, the percentage correct for the set of those words on the pretest and the Post-3 was computed and analyzed.

As shown in Table 5.11, Levene’s test of homogeneity of variances suggested that the three groups could be considered to have similar variances on both the pretest (%_correct_target_clicked; $F(2, 42) = 1.54, p = .23$) and the posttest (Post3 %_correct_target_clicked; $F(2, 42) = 2.23, p = .12$), confirming the error variance of all participants were equivalent at the outset of the study.

Test	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i> -value
Pretest %_correct_target_clicked	1.54	2	42	.23
Post3 %_correct_target_clicked	2.23	2	42	.12

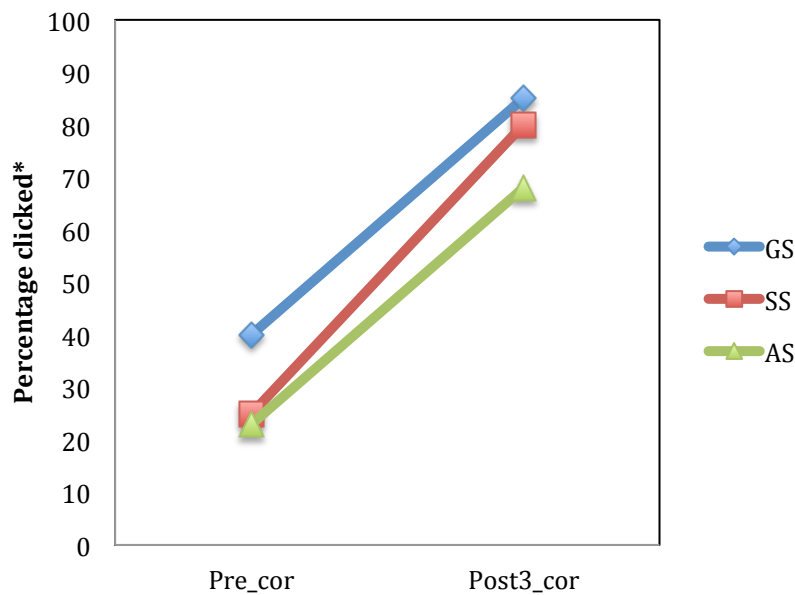
Table 5.11. Levene’s Test of Homogeneity of Error Variances

In computing the percentage, the participants’ raw scores were first calculated and then converted to a percentage. For example, if the participant clicked on five target words while reading (e.g. *resilient, trivial, obedience, trigger, indicate*), and he/she was correct with three of them on the pretest (e.g. *trivial, obedience, trigger*) and with four of them on the Post-3 (e.g. *resilient, trivial, obedience, trigger*), the pretest percentage correct out of target words clicked would be 60% (=3/5), the Post-3 would be 80%

(=4/5), and the gain between the pretest and the Post-3 would be 20% ($4/5 - 3/5 = 1/5$). The percentage figures were computed for all three groups in this manner. The descriptive statistics of the percentage correct of the clicked words in the pretest and the Post-3 are presented in Table 5.12 and graphically displayed in Figure 5.12.

	Pretest		Post-3	
	Mean	SD	Mean	SD
GS (n=15)	.40	.32	.85	.22
SS (n=15)	.25	.18	.81	.25
AS (n=15)	.23	.25	.68	.32

Table 5.12. Descriptive statistics for vocabulary acquisition for clicked words (percentage correct)



* percentage correct of the words that were clicked

Figure 5.12. Percentage correct of the clicked target words over the tests

As one can see in Figure 5.12, the SS group showed the largest gain from the pretest to the Post-3. In addition, as shown in Figure 5.13, the SS group gained more than the GS

and the AS group, which showed equal gains. This pattern (i.e., $SS > GS = AS$) is descriptively different from the average overall gain scores (i.e., $GS > SS > AS$).

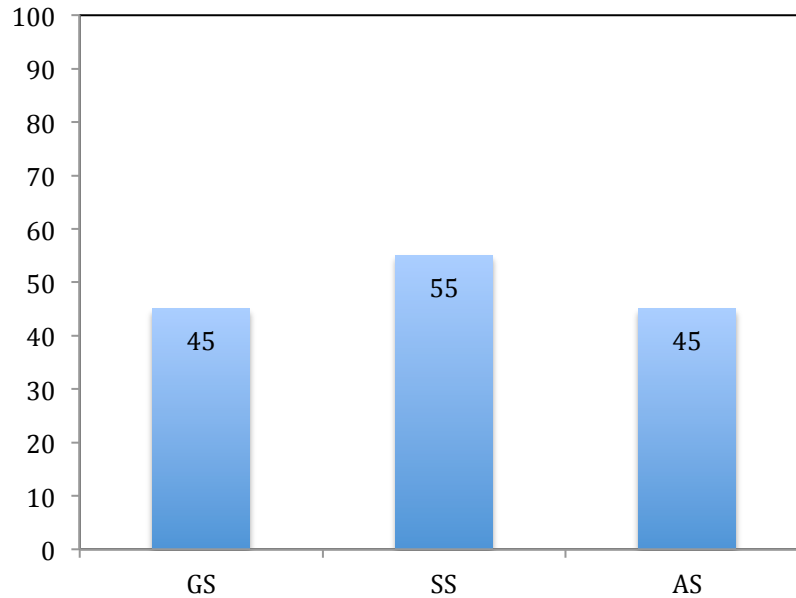


Figure 5.13. Pre-Post3 gain among target words clicked (percentage)

In order to examine if these results (i.e., group differences on pre-post gains on clicked target words) were statistically significant, the RM ANOVA was performed. In the RM ANOVA, the effect of Time showed a statistically significant difference ($F(1, 42) = 96.20, p < 0.001$). However, the effect of Time*Group showed no significant difference in this case ($F(2, 42) = 0.60, p = 0.55$), indicating Group differences in terms of their gains from the pretest to the Post-3 were not statistically reliable. Although none of these differences were significant, two potentially interesting points emerged: descriptively speaking, 1) as shown in Figure 5.13, the SS group showed the largest gain between pretest and Post-3 (55%) and 2) the AS group showed as much improvement as the GS group (45%). These results may have 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, or the

task of determining the relevant sense given context among sense entries could possibly have attracted the learner's attention enough to facilitate more acquisition.

Lastly, participants' vocabulary acquisition through additional vocabulary posttests (i.e., Post-1, Post-2, Post-4) was analyzed to investigate if there were any statistically significant differences on various kinds of vocabulary knowledge development (e.g., some learners show more development on production and some show better recognition). The descriptive statistics for the participants' performance on each of the three posttests are shown in Table 5.13.

Test		Mean (highest= 20)	SD
Post-1	GS (n=15)	13.87	3.94
	SS (n=15)	14.80	2.76
	AS (n=15)	13.73	3.99
	NS (n=15)	10.53	4.49
Post-2(target (target only)	GS (n=15)	10.00	3.02
	SS (n=15)	9.87	2.64
	AS (n=15)	9.47	2.39
	NS (n=15)	7.73	3.37
Post-2 (accept other senses)	GS (n=15)	13.33	3.74
	SS (n=15)	13.53	3.07
	AS (n=15)	13.33	3.22
	NS (n=15)	10.87	4.17
Post-4	GS (n=15)	16.93	3.41
	SS (n=15)	16.60	3.56
	AS (n=15)	15.73	3.92
	NS (n=15)	13.87	3.82

Table 5.13. Descriptive statistics of Post-1, Post-2, and Post-4 scores across the four groups

As one can see in Table 5.13, though the differences were very small, participants generally showed a similar trend on the three posttests; the sense-specific groups (GS and SS) appear to outperform the AS and NS groups. As the most superficial level, the Post-1 tested the participants on how much they could recognize target words, and descriptively, the sense-specific groups (GS and SS) showed more recognition than did the AS and NS groups. The Post-2 tested learners' production knowledge, and this test also showed that the sense-specific groups outperformed the AS and NS groups. For describing words in the Post-2, learners' scores were analyzed in two ways; one way that allowed only the sense of the word used in the reading text (*target only*), and another way that allowed any sense of the word (*accepting other senses*). Participants showed a better performance on the Post-2 when they were allowed to answer with other possible senses of the target words (accepting other senses), which may indicate that they know different meanings of target words than those used in the reading text. Comparing the participants' performance on the Post-2 for *target only* with the Post-2 for *accepting other senses*, when they were required to answer with the sense of the words used in the reading text (*target only*), the sense-specific groups (GS and SS) showed better performance, which may suggest that sense-specific information helps participants' learning. Since participants were provided with more cues to answer the questions in the Post-4, the average scores across the four groups were higher than the scores of the other posttests (i.e., Post-1 and Post-2). Again, the results showed a similar pattern to those of the other posttests; GS>SS>AS>NS.

These descriptive results need to be statistically tested to see if the differences of the groups for each test were significant. Prior to the experiment, the test of homogeneity of error variances confirmed that the error variances were equivalent (Post-1: $F(3, 56) =$

1.14, $p = 0.34$; Post-2 (target): $F(3, 56) = 0.71$, $p = 0.55$; Post-2 (others): $F(3, 56) = 0.79$, $p = 0.50$; Post-4: $F(3, 56) = 0.65$, $p = 0.59$, indicating that the groups are homogeneous for each of the posttests.

Test	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i> -value
Post-1	1.14	3	56	0.34
Post-2 (target)	.71	3	56	0.55
Post-2 (others)	.79	3	56	0.50
Post-4	.65	3	56	0.59

Table 5.14. Levene's Test of Homogeneity of Variances

Since there was no comparable pretest corresponding to each of these tests, a one-way ANOVA was performed on the scores of each test to examine if the group differences for each test were statistically significant. As shown in Table 5.15, the effect of Group only in the Post-1 showed a statistically significant difference, which indicated that the groups were different in their vocabulary development on recognition ($F(3, 56) = 3.51$, $p = 0.02$).

Test	<i>df1</i>	<i>df2</i>	Mean Square	<i>F</i>	<i>P</i> -value
Post-1	3	56	51.98	3.51	.02
Post-2 (target)	3	56	16.44	1.98	.13
Post-2 (others)	3	56	24.20	1.89	.14
Post-4	3	56	28.33	2.10	.11

Table 5.15. Results of ANOVA analysis

Since there were significant differences among the groups on the scores of the Post-1, post-hoc comparisons were performed to locate where the groups differ. Tukey post-hoc

test compared all four groups in pair and found a significant difference between the SS group and the NS group ($p = .02$), as shown in Table 5.16. All other groups in pair did not show a significant difference. To some extent, this result also supported the argument that sense-specific lexical information helps learners' vocabulary recognition.

(I)Group	(J)Group	Mean Difference (I-J)	Std. Error	Sig.	95% confidence interval	
					Lower Bound	Upper Bound
GS	SS	-.93	1.41	.91	-4.65	2.79
	AS	.13	1.41	1.00	-3.59	3.85
	NS	3.33	1.41	.09	-.39	7.05
SS	GS	.93	1.41	.91	-2.79	4.65
	AS	1.07	1.41	.87	-2.65	4.79
	NS	4.27*	1.41	.02	.55	7.99
AS	GS	-.13	1.41	1.00	-3.85	3.59
	SS	-1.07	1.41	.87	-4.79	2.65
	NS	3.20	1.41	.12	-.52	6.92
NS	GS	-3.33	1.41	.09	-7.05	.39
	SS	-4.27*	1.41	.02	-7.99	-.55
	AS	-3.20	1.41	.12	-6.92	.52

Table 5.16. Tukey post-hoc comparisons for Group differences

In conclusion, the overall results suggest a positive answer to the first research question about whether sense-specific lexical information leads to better vocabulary acquisition. The relatively consistent results from several different analyses suggested that 1) learners provided with sense-specific lexical information during reading have greater vocabulary gains from the pretest to the Post-3, 2) learners in the Low group generally showed more vocabulary learning than did learners in the High group, as the Low group had more room for improvement, 3) the Low group clicked on nearly twice as

many target words than did the High group, 4) learners provided with sense-specific information also showed better recognition of the target words.

5.3.2 Reading comprehension

The second research question explored whether sense-specific lexical information facilitates learners' reading comprehension. The data collected from all four groups on their reading comprehension task were analyzed and the descriptive statistics for reading comprehension mean scores of the four groups are summarized in Table 5.17 and displayed graphically in Figure 5.14.

	<i>Mean (highest=42)</i>	<i>SD</i>
GS (n=15)	35.80	2.98
SS (n=15)	37.07	2.46
AS (n=15)	34.93	3.08
NS (n=15)	33.27	3.69

Table 5.17. Descriptive statistics for reading comprehension mean scores across the four groups

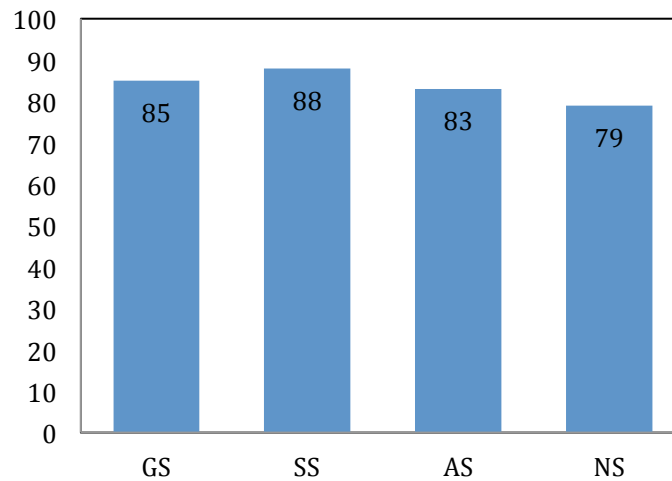


Figure 5.14. Reading comprehension scores of the four groups (percentage)

As shown in Table 5.17 and Figure 5.14, the difference among the reading comprehension mean scores of the four groups was within 4 points (SS, 37.07; NS, 33.27), 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 these differences among the groups were statistically significant, a one-way ANOVA was run on the reading comprehension scores to determine whether sense-specific lexical information had a significant effect on learners' reading comprehension. The test of homogeneity of error variances confirmed that the error variances were equivalent ($F(3, 56) = 0.96, p = 0.42$) prior to the experiment.

Test	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i> -value
RC	0.96	3	56	.42

Table 5.18. Levene's Test of Equality of Error Variances

The results of the one-way ANOVA for learners' reading comprehension scores are displayed in Table 5.19. As shown, the effect of Group showed a statistically significant difference, indicating that the groups were different in their reading comprehension ($F(3, 56) = 4.01, p = 0.01$). Since the four groups were shown to be significantly different in their reading comprehension performance, it was necessary to locate where the differences existed among the groups.

Source	<i>df1</i>	<i>df2</i>	Mean Square	<i>F</i>	<i>P</i> -value	Partial Eta ²
Group	3	56	38.18	4.01	.01	.18

Table 5.19. Results of one-way ANOVA for reading comprehension scores of the four groups

Tukey post-hoc tests (Table 5.20) compared all four groups in pair and revealed a significant difference between the SS group and the NS group ($P = .01$). On the other hand there were no significant differences between the GS group and the SS group ($P = .67$), the GS and the AS ($P = .87$), the GS and the NS ($P = .12$), the SS and the AS ($P = .24$), and the AS and the NS ($P = .46$).

(I)Group	(J)Group	Mean Difference (I-J)	Std. Error	Sig.	95% confidence interval	
					Lower Bound	Upper Bound
GS	SS	-1.27	1.12	.68	-4.25	1.72
	AS	.87	1.12	.87	-2.12	3.85
	NS	2.53	1.12	.12	-.45	5.52
SS	GS	1.27	1.12	.68	-1.72	4.25
	AS	2.13	1.12	.24	-.85	5.12
	NS	3.80*	1.12	.01	.82	6.78
AS	GS	-.87	1.13	.87	-3.85	2.12
	SS	-2.13	1.13	.24	-5.12	.85
	NS	1.67	1.13	.46	-1.32	4.65
NS	GS	-2.53	1.13	.12	-5.52	.45
	SS	-3.80*	1.13	.01	-6.78	-.82
	AS	-1.67	1.13	.46	-4.65	1.32

* The mean difference is significant at the 0.05 level

Table 5.20. Tukey post-hoc comparisons for Group differences

The participants' reading comprehension scores in the High and Low groups were also compared to see if learning new words (with different prior vocabulary knowledge) affects reading comprehension scores. Table 5.21 shows the descriptive statistics for the participants' performance, which is also graphically displayed in Figure 5.15. Both the High group and the Low group showed a very similar pattern; the sense-specific groups showed a better performance (i.e., $GS \cong SS > AS > NS$). At the same time, it appeared that

there was very little difference between the Low and High groups, which may indicate that learning new words seemed not to affect the reading comprehension scores.

Group	LowHigh	Mean (highest=42)	SD
GS	H (n = 6)	37.83	1.722
	L (n = 5)	35.20	3.194
SS	H (n = 8)	37.38	2.875
	L (n = 5)	35.80	1.483
AS	H (n = 8)	34.75	2.964
	L (n = 6)	34.83	3.656
NS	H (n = 6)	33.67	3.445
	L (n = 6)	31.17	3.601

Table 5.21. Descriptive statistics for reading comprehension mean scores for the High (H) and Low (L) groups across the four groups

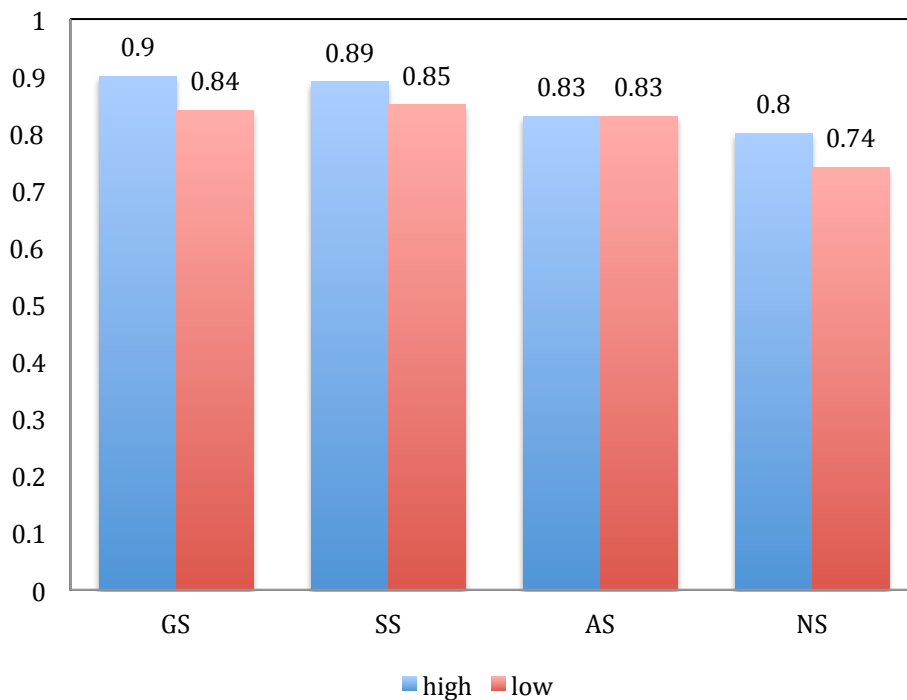


Figure 5.15. Total RC mean scores for the High/Low groups across the four groups

Two-way ANOVA was run to see if the differences were statistically significant. As shown in Table 5.22, Group showed a statistically significant difference ($F(3, 42) = 5.11$, $p = .004$, partial $\eta^2 = 0.27$), whereas H/L did not show a statistically significant difference ($F(1, 42) = 3.74$, $p = .06$, observed power = 0.47). Since Group showed a statistically significant difference, all groups needed to be compared in pair to locate where the differences existed among the groups.

Source	<i>df1</i>	<i>df2</i>	Mean Square	<i>F</i>	<i>P</i> -value	Partial Eta ²	Obs. Power
Group	3	42	45.56	5.11	.004	.27	.90
HL	1	42	33.34	3.74	.06	.08	.47

Table 5.22. Results of two-way ANOVA for reading comprehension scores for the four groups and HL groups

Tukey post-hoc comparisons revealed a significant difference between the GS and NS groups ($p = .01$) and between the SS and NS groups ($p < .001$). All other groups in pair did not show a significant difference (i.e., GS-SS: $p = 1.00$, GS-AS: $p = .43$, SS-AS: $p = .32$, AS-NS: $p = .20$). This confirmed that the sense-specific groups (GS and SS) outperformed the NS group in terms of reading comprehension and to some extent, supported the argument that sense-specific lexical information facilitates learners' reading comprehension.

(I)Group	(J)Group	Mean Difference (I-J)	Std. Error	Sig.	95% confidence interval	
					Lower Bound	Upper Bound
GS	SS	-.13	1.22	1.00	-3.41	3.14
	AS	1.85	1.20	.43	-1.37	5.07
	NS	4.22*	1.25	.01	.88	7.56
SS	GS	.13	1.22	1.00	-3.14	3.41
	AS	1.98	1.15	.32	-1.09	5.06
	NS	4.35*	1.20	.00	1.15	7.55
AS	GS	-1.85	1.20	.43	-5.07	1.37
	SS	-1.98	1.15	.32	-5.06	1.09
	NS	2.37	1.18	.20	-.77	5.51
NS	GS	-4.22*	1.25	.01	-7.56	-.88
	SS	-4.35*	1.20	.00	-7.55	-1.15
	AS	-2.37	1.18	.20	-5.51	.77

Table 5.23 Tukey post-hoc comparisons for Group differences

In sum, the results showed that the various kinds of lexical information (i.e., sense-specific information) affected the learners' reading comprehension scores, and the differences were statistically reliable. On the other hand, descriptively, the learners' previous knowledge of target words (i.e., High and Low) seemed not to affect their reading comprehension scores, but this result did not show statistical significance. In addition, it is worth noting that the learners' performance on reading comprehension showed a similar pattern to their performance on vocabulary acquisition; the groups were ranked similarly in terms of performance (i.e., the sense-specific groups (GS, SS) performed better than the AS/NS groups).

5.3.3 Effects of the system errors

In the present study, some differences between the Gold Senses (GS) condition and the System Senses (SS) condition were observed, but it has not been explored to what extent the learners in the SS group were impacted specifically by words that were incorrectly disambiguated.

Nine words revealed to be presented on the system with different sense-specific lexical information (because they were incorrectly disambiguated by the system, see section 4.5.3.4) from its context of the reading text (= *inappropriate target-sense words*). The nine words of this *inappropriate target-sense words* group were *agitate*, *aspect*, *banish*, *resist*, *indicate*, *expenditure*, *sector*, *chain*, and *conscience*.

After examining all words that learners in the SS group *clicked*, the author determined cases that learners got wrong in the pretest. Among the cases of the nine words for which the system provided lexical information that was different from the context of the reading text, a total of 18 cases clicked while reading were found to be incorrect in the pretest. Among the 18 cases, a total of nine cases were shown to be correct in the Post-3 (case B). The other nine cases showed to be incorrect in the Post-3 (case A). Likewise, 12 cases clicked while reading were found to be correct in the pretest. Among these 12 cases, two cases were incorrect in the Post-3 (case D). The other 10 cases were correct in the Post-3 (case C). The same procedure was applied to the other 11 words presented on the system with appropriate sense-specific information in context of the reading text (= *appropriate target-sense words*). One can see the different performance for words that learners *clicked* on for these two types in Table 5.24.

Based on the results in Table 5.24, when learners were wrong in the pretest and received appropriate sense information during reading, they showed 76% accuracy in the Post-3, whereas learners showed 50% improvement when provided with inappropriate sense information (comparing case *a* with case *A*). This, unsurprisingly, showed the value of correct sense information (case *a* = 76%).

Clicked while reading	Pretest	Post-3	Accuracy	
inappropriate target-sense words = 9 (word <i>types</i>)	Incorrect = 18 (<i>tokens</i>)	Correct = 9	9/18 = 0.50	<i>A</i>
		Incorrect = 9	9/18 = 0.50	<i>B</i>
	Correct = 12	Correct = 10	10/12 = 0.83	<i>C</i>
		Incorrect = 2	2/12 = 0.17	<i>D</i>
appropriate target-sense words = 11	Incorrect = 42	Correct = 32	32/42 = 0.76	<i>a</i>
		Incorrect = 10	10/42 = 0.24	<i>b</i>
	Correct = 16	Correct = 14	14/16 = 0.88	<i>c</i>
		Incorrect = 2	2/16 = 0.13	<i>d</i>

Table 5.24. Pre to Post-3 gain on clicked words for learners in the SS group

On the other hand, when learners were correct in the pretest and received appropriate sense information during reading, they showed 88% improvement whereas they showed 83% improvement when provided inappropriate sense information during the reading (comparing case *c* with case *C*). This is important, as it seems to indicate that wrong sense information was not leading learners astray.

5.3.4 Implications

This study intended to investigate if sense-specific lexical information would guide learners to better vocabulary acquisition and reading comprehension. The findings suggested the following important implications. First, the type of lexical information (GS, SS, AS, NS) impacts the amount of vocabulary learning and reading comprehension. One thing to note from the results was that the AS group demonstrated a high performance in vocabulary acquisition and reading comprehension that was similar to the performance shown by the GS and SS groups. The current study speculated that it could be attributed to a small number of sense entries of the target words (average = 2.95). As such, it would be interesting for a further study to explore with target words that have more sense entries. Furthermore, with target words that more sense entries, it would be an interesting future study to investigate how the number of sense entries of words affects learners' vocabulary learning (e.g., if words with more sense entries are harder for learners to acquire, or if more polysemous words are harder for learners to acquire, and so on.)

Second, learners demonstrated higher performance in vocabulary acquisition and, to some extent, reading comprehension when they received sense-specific information. This confirmed that the provision of sense-specific lexical information during reading is more helpful.

Third, learners with lower prior vocabulary knowledge showed more vocabulary learning (improvement) than learners who knew more target words before the task, as they knew fewer words at the beginning and thus had more room for learning (improvement). Also, on average, learners who knew fewer target words clicked nearly twice as many words while reading as did learners who knew more target words. This

indicates that learners with lower vocabulary knowledge seem to need more vocabulary assistance than do learners with higher vocabulary knowledge.

Lastly, for the effect of the automatic system errors, the results showed that learners demonstrated a similar amount of learning regardless of appropriateness of lexical information when they had already known the target word. However, they showed more learning provided with appropriate lexical information when they did not initially know the target word, which suggests that it is still worthwhile to provide sense-specific lexical information.

VI. SUMMARY AND OUTLOOK

In the present study, an intelligent learning system was developed and its positive effects on vocabulary learning and reading comprehension of second language learners were investigated. After presenting the rationale of the present study and related research (Chapters 1 and 2), the architecture of the system was described and the process of building the system was shown, providing the algorithm newly established in the study as well as an evaluation of the system, showing the effectiveness of the algorithm (Chapters 3 and 4). The present study then demonstrated the impact of the system on actual language learners of English, leading to improvement in their vocabulary learning and reading comprehension (Chapter 5). The following sections summarize the study results, discuss the implications and limitations of the study, and provide suggestions for future research.

6.1 The system from a computational perspective

To support vocabulary learning and reading comprehension for language learners, the online system was developed, allowing learners to upload or choose texts and click on any content word in order to obtain sense-appropriate lexical information for unfamiliar or unknown words during reading. The system consists of three components: 1) the system manager, 2) the NLP server, and 3) the lexical database. The system manager controls the interaction among each learner, the NLP server, and the lexical database. The NLP server contains several NLP modules for tokenizing, lemmatizing, POS tagging, collocation identification, and word sense disambiguation. Those modules take part in

converting a raw input text to a linguistically analyzed text. The lexical database is used to provide a sense-appropriate definition and example sentences of an input word to the learner. To obtain sense-appropriate information, the system first performs word sense disambiguation (WSD) on the input text. The system used SenseRelate::AllWords (SR::AW) (Pedersen and Kolhatkar, 2009) to perform WSD on input texts, as this WSD system has broad coverage of content words.

Pointing to appropriate examples, however, is complicated by the fact that the database of examples is from one repository (COBUILD), while automatic WSD systems generally rely on senses from another (WordNet). The lexical database, then, is indexed by WordNet senses, each of which points to an appropriate corresponding COBUILD sense. To make it feasible for the lexical database to redirect an input WordNet sense to a corresponding COBUILD sense, a word sense alignment (WSA) algorithm was developed. The WSA of the study works by first running SR::AW WSD system on COBUILD examples in order to induce a basic/initial alignment structure between WordNet and COBUILD, adjusting this structure according to a heuristic which favors flatter alignment structures. The best alignment structure generated from the WSA step is then sent to the lexical database for linking WSD output senses to corresponding COBUILD senses, allowing the system to finally present the appropriate COBUILD definition and examples.

There are several implications to this work: First, a new pair of inventories (WordNet and COBUILD) was employed, thus increasing the scope of WSA work. Second, the purpose for alignment in the study is unique: instead of increasing the size of a knowledge base, where issues such as the redundancy of senses are important, the study

finds the areas of commonality between two resources. Third, while previous studies mostly used a lot of information (extracted from inventories) as an input for more sophisticated WSA methods, this research simply uses information extracted from a state-of-the-art WSD classifier for the alignment work, which makes the present system fairly light but robust in performance. Lastly, the new method in the present study opens up the possibility of exploring alignment structures as a whole, and this is the first which focuses specifically on trends in alignment structure between two inventories.

The study examined the system performance with respect to accuracy of alignment between WordNet and COBUILD on an evaluation set constructed by pooling the judgments of semi-experts. To obtain these judgments, an online survey was used in which a sense of a target word from WordNet (as a question) with senses from COBUILD (as choices) was presented, asking for judgments of relatedness between them. With different weights of related meaning in estimating accuracy of the WSA system, the precision got higher with higher counts of related meaning. Also, the study examined the alignment results with and without accounting for the flatness of alignment structures; the finding was that system performance was enhanced by applying the heuristic favoring flatness in alignment structures. This validates the heuristic of favoring flat alignments.

Throughout the discussion of this evaluation process, the study has noted that 1) it was difficult for semi-experts to agree upon correct alignments, showing that the task itself was difficult even for human and that 2) despite this, such data could be used to gauge the accuracy of WSA systems, depending upon how much related meaning one wishes to capture in the alignments.

None of the previous systems developed for language learning, as discussed in section 2.2, is well-suited enough to accomplish the task. The findings of this research thus have significant implications for those who are trying to integrate NLP technology into educational systems (e.g., ICALL). First, this research demonstrated a successful application of NLP technologies (e.g. WSD, WSA) to the educational system, especially involved in meaning processing, which has been regarded as challenging to implement. Second, considering that the most ICALL systems provide practice on grammatical forms and functions, this research can foster awareness of the usefulness of ICALL systems for meaning processing, specifically for vocabulary practice, which can also enhance a learners' reading comprehension when combined with reading practice.

Despite the new insights on WSA that the present study suggests, there are still several directions to pursue. First, the study implemented a preference for flat alignments over skewed ones; however, this could be revised by using a sample of correct alignments or possibly from other resources, if available. The preliminary investigation into WSA between WordNet and Wiktionary, although they are entirely different resources, shows quite different behavior from the heuristic (i.e., favoring flat alignments) of the present study. Thus, this is an open question for future research. Also, other applications may require different assumptions. For further evaluation and development of the present system, more gold standard data needs to be collected; furthermore, the system needs to be tested with more words. Moreover, one may want to add human verification for the alignments in order to ensure quality mappings for a real-world vocabulary assistant system.

Regarding the WSA algorithm developed in the present study, the study found that the portion of the algorithm relying on the sense probability distribution seems relatively conservative and biased not to change the WSA structures from the initial alignment based on WSD outputs to the adjusted alignment. This may be due to two factors: 1) the sense distribution probability may be weighted relatively strongly whereas the alignment structure probability (reflecting the flatness of the structure) is relatively weak; or, 2) the words used for testing the system have a small number of senses to validate the results. Investigating these factors merits future work.

6.2 The system from a language learning perspective

Upon the completion of building the system, the study evaluated whether this set-up of providing sense-specific lexical information can lead learners to improve their vocabulary acquisition and reading comprehension; the system was examined with 60 intermediate Korean learners of English as a second language (ESL). Those 60 participants were randomly assigned to one of four groups: 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, and 4) No senses (NS) – reading without any support of lexical information. Since pre-determined input texts were used, in the study gold standard information was created, where each word in the text is manually given a link to the appropriate COBUILD information. This lets the present study gauge: 1) whether the gold-standard information is helpful to

learners, and 2) comparatively speaking, what the effects are of using the potentially noisy information provided by the authentic system.

To answer the question regarding if participants show better vocabulary acquisition with sense-specific lexical information, the study compared the performance of each of four groups and found that the GS group showed the most gain in vocabulary learning from the pretest to the posttest, whereas the NS group showed almost none. This demonstrated what the present study expected. The sense-specific groups (i.e., the GS and SS groups) showed more vocabulary acquisition than the AS and NS groups. To test if these results were statistically significant, a repeated-measures analysis of variance (RM ANOVA) was used. The analysis showed a statistically significant difference on the time by group interaction effect (Time*Group), which was most crucially related to the research question. This significant effect of time by group interaction was a reliable, positive answer to the question regarding if sense-specific information facilitates learners' vocabulary learning. Post-hoc analyses found significant differences between the GS-AS groups, the GS-NS groups, and the SS-NS groups on the improvement over time (Time*Group). The vocabulary improvement over time (i.e., pre-post gain) between learners above the mean gain score (High) and learners below the mean gain score (Low) were also analyzed. The learners in the Low group showed much more improvement than the learners in the High groups, indicating the Lower group could show more improvement because they have more room for improvement. Also the Low group clicked almost twice as many target words as did the High group. The study also analyzed the pre-post gain only for the target words which were clicked (GS, SS, AS) and found that all three groups showed more improvement on clicked words.

The overall results suggest a positive answer to the research question about whether sense-specific lexical information leads learners to better vocabulary learning. The results from several different analyses suggested that: 1) learners provided with lexical information during reading have more vocabulary learning, with sense-specific information having a descriptively greater increase; 2) learners in general appear to acquire more when they check the meaning during the task; and 3) they seem to check the meaning more when the meaning is disambiguated correctly.

To answer the question regarding whether learners improved in reading comprehension, the study compared reading comprehension test scores of the four groups and found a trend similar to that of vocabulary acquisition, though the differences between groups were relatively small (GS: 85%, SS: 88% > AS: 83%, NS: 79%). The analysis generated by a one-way ANOVA showed a statistically significant difference on the effect of Group. Post-hoc analyses located a significant difference between the SS group and the NS group. To some extent, the results support the idea that sense-specific lexical information facilitates learners' reading comprehension. Curiously, the GS group, which received more accurate sense information than the SS group, did not show significant differences in post-hoc comparisons with the NS group, despite descriptively showing slightly higher reading comprehension scores (6%). This issue warrants investigation in the future.

In order to gauge the effect of automatic system errors on vocabulary acquisition, distinguishing the SS from the GS conditions, the study also examined the target words for which the system gave incorrect information. The results showed that when learners were wrong in the pretest and received correct sense information, they did show 76%

accuracy. However, they showed 50% accuracy even when provided with incorrect information, which suggests that any information helps to some extent, though it is clear that correct sense information helped learners learn more (76% > 50%). On the other hand, when they already knew the sense of the word, the incorrect sense information did not impact their learning: there was 88% accuracy with correct sense information during the reading and 83% accuracy with incorrect sense information. This indicates that learners are able to distinguish the appropriateness of newly provided sense information during reading when they already know the sense of the word.

The findings of the present study showed some important implications: First, the types of lexical information (GS, SS, AS, NS) influenced learners' vocabulary acquisition and reading comprehension. In particular, learners demonstrated higher performance when they clicked words to get their sense information. Moreover, most results show a similar trend such that the GS and SS groups outperformed the AS and NS groups. These findings gave a reliably positive answer to the research questions about whether sense-specific lexical information would support learners' vocabulary acquisition and reading comprehension, which in turn justifies the rationale behind the design of the online system in the present study. Second, from the examination of the effect of automatic system errors, it can be concluded that although learners showed learning regardless of appropriateness of lexical information, they still showed relatively greater learning when given appropriate lexical information. Surprisingly, even at times inaccurate information was helpful, though how much degradation is allowed before it becomes harmful is an area to explore in the future.

Although a few studies have shown actual learning outcomes for an ICALL system (e.g., Heift, 2001, Kulkarni et al., 2008; Petersen, 2006), it is still scant. There have been even fewer studies trying to build systems to support vocabulary learning and reading comprehension. Among the few, the REAP tutor (Heilman et al., 2006; Kulkarni et al., 2008) is regarded as a good system, providing a various supports for language learners, but their system is limited in terms of vocabulary learning through reading (see 2.2.2 for more details). More importantly, the present study is the first to more seriously consider the sense inventory and examples that will be shown to learners, linking modern WSD with learner-appropriate examples. In that sense, the present study is the first to offer a tutoring system which more effectively utilizes the effectiveness of sense-specific information to support vocabulary acquisition and reading comprehension.

One of the interesting venues for future research would be developing a tutoring system which provides sense-specific information in a multimedia mode: such a system might extract corresponding multimedia sources of a word's sense from the resources and present the lexical information with multimedia sources. This might lead learners to accomplish more successful vocabulary acquisition and reading comprehension. Also, it would be interesting to explore if there are any differences between learners who learn COBUILD senses and those who learn WordNet senses. Focusing more on the polysemous nature of words, it would be an interesting future work to investigate if there are any differences between highly polysemous words and less polysemous in vocabulary learning and reading comprehension.

Appendix A

Code of computing a probability of alignment structure (=P(A))

probA.py

```
#!/bin/env python

import math
import itertools

##### change CB and WN

#=====
# function to generate type of WSA
#=====
def gen_atype(c,w):
    type = []
    type_list= list(itertools.combinations_with_replacement(range(w+1),c))

    for i in type_list:
        total = sum(i)

        if total == w:
            type.append(i)

    return type
#=====
c = 3 # number of CB senses
w = 14 # number of WN senses

dtpt = float(c)
numrtr = float(w)

### 1. generate average ####
m = numrtr/dtpt
print "Average = "
print m

### 2. generate alignment type given CB & WN senses ####
type = gen_atype (c,w)
print type
#type = [(0,0,6), (0,1,5), (0,2,4), (0,3,3), (1,1,4), (1,2,3), (2,2,2)]

### 3. generate standard deviation for each alignment type ####
dic={}
for each in type:
    list=[]
    for j in each:
        x= j-m
        t=pow(x, 2)
        list.append(t)
```

```

a=0
for k in list:
    a = a+k
a= float(a)
sd = math.sqrt(a/dtpt)

dic[each]=sd

### 4. find the highest score (bigest SD = most skewed) #####
print "\n"+ "each_alignment_type & SD"

higher=0
for i in dic:
    print i, dic[i]

    if higher < dic[i]:
        higher=dic[i]

print "\n" +"highest SD"
print higher

### 5. reverse sd to have flattest highest, skewedest lowest #####
### and calculate sum of reversed scores #####
print "\n"+ "each_alignment_type & SD & highestSD-SD(=score)"

dic2={} # dic for each alingment type and its reversed score (higest SD-SD)
sum=0 # sum of reversed score
for i in dic:
    rvrs= higher-dic[i]
    sum=sum+rvrs

    print i, dic[i], rvrs
    dic2[i]=rvrs

print "\n"+ "sum of score"
print sum

### 6. normalize : each score/sum ==> P(A) #####
print "\n"+ "each_alignment_type & score & normalized=score/sum=P(A)"

dic_P_A = {} # dic for P(A) for each alingment type
for i in dic2:
    P_A = dic2[i]/sum

    print i, dic2[i], P_A
    dic_P_A[i]=P_A

print "\n"+ "dictionary of P(A) for each type of alignment = "
print dic_P_A

print "\n"

```

Appendix B

Code of computing word sense alignment ($=P(A)*P(WN|CB)$)

(1) main_WSA.py

```
#!/bin/env python

from defs_WSA_20 import *
from copy import deepcopy
import sys

#=====
# extract all info for WSA from every file (i.e. CB, WN, ST, Avrf, P_A)
#=====

file = open(sys.argv[1], 'r')

line_list = file.readlines()

item=line_list[0].split()

cb = item[2]
wn = item[3]

CB = int(cb)
WN = int(wn)

word = line_list[1]

ST_list = line_list[2].strip()
SenseTable = ST_list.split('=')
ST = eval(SenseTable[1])

AV = line_list[3].strip()
Average = AV.split('=')
avrg = float(Average[1])

PofA = line_list[4].strip()
Prob_A = PofA.split('=')
P_A = eval(Prob_A[1])

#=====
# Main
#=====

items = ST.items()
items.sort()

#=====
# Step_1
#=====

# determine the number of wn senses
biggest=1
for i in items:

    if i[0][1] > biggest:
        biggest = i[0][1]

BEST=[]
```

```

LISTS=[]
a=1

while a <= biggest: ##Ndics:
    n=[a]
    a=a+1

    # call def bld_lists_wn
    n=bld_lists_wn(n,items)

    if len(n)==1:
        pass

    else:
        LISTS.append(n)

        higher = 0

        # call def best
        higheritem=best(higher, n)
        BEST.append(higheritem)

print word
print "Best Alignment = ", BEST

```

(2) defs_WSA.py

```

#!/bin/env python

from copy import deepcopy

#=====
# (step_1) build lists per wn sense
#=====
def bld_lists_wn(n, items):
    for i in items:
        if i[0][1] == n[0]:
            n.append(i)
    return n #print n # (i.e.) [1, ((c1,w1), val), ((c2,w1), val), ((c3,w1),val)]

#=====
# (step_1) build a list of best pairs
#=====
def best(higher, n):
    for i in n[1:]: #becuase n[0]=wn sense no tag
        if higher < i[1]:
            higher = i[1]
            higheritem = i
        elif higher == i[1]:
            pass
        else:
            pass
    return higheritem

```

```

=====
# (step_2) generate alternative alignments: list for each of WN with dis-selected WN_CB pairs in BEST
=====
def alt_algnmt(BEST, LISTS):
    new_LISTS = []
    a=0
    for wn_list in LISTS:
        for j in wn_list[1:]:
            if j[0] != BEST[a][0]:
                another = deepcopy(BEST)
                another[a] = j
                new_LISTS.append(another)
            another = []
        a=a+1
    return new_LISTS

=====
# (step_3_1) build dictionary of CB senses as key and its branches as value
=====
def gen_cb_brnch(cnd_alignment, CB):
    c_brnch = {}
    for i in cnd_alignment: #BEST:
        if c_brnch.has_key(i[0][0]):
            c_brnch[i[0][0]]+=1
        else:
            c_brnch[i[0][0]]=1

    c_list=[]
    a=1
    while a <= CB:
        c_list.append(a)
        a = a+1

    for key in c_brnch:
        for i in c_list:
            if key == i:
                c_list.remove(i)

    for i in c_list:
        c_brnch[i]=0

    return c_brnch

=====
# (step_3_2) generate alignment type for given candidate alingment
=====
def gen_aligmnt_type(c_brnch):
    type=[]
    for key in c_brnch:
        type.append(c_brnch[key])
    type.sort()
    return type

```



```

=====
# (step_3_3) # find cb branches which meet the average of branches
=====
def bigger_cb_brnch (avrg, c_brnch):
    fixed=[]
    for key in c_brnch:
        if c_brnch[key] >= avrg:
            fixed.append(key)
    return fixed

=====
# (step_3_4) # find cb branches which is bigger than the average of branches
# and from the BEST, find items which have the same cb sense as the above one
=====
def find_bigger_inBEST (c_brnch, avrg, BEST):
    pairs = []
    for key in c_brnch:
        if c_brnch[key] > avrg:
            for i in BEST:
                if i[0][0] == key: #and key not in fixed:
                    pairs.append(i[0])
    return pairs

=====
# (step_3_5)# from new_LISTS, find items which have the same WN sense and different CB sense from the BEST
# But exclude wn-cb whose cb is in the fixed (already meets/beyond the average)
=====
def gen_cnd_algnmt(new_LISTS, pairs, fixed):
    cnd_lists = []
    for i in new_LISTS:
        for j in i:
            for k in pairs:
                if j[0][1] == k[1] and j[0][0] != k[0] and j[0][0] not in fixed:
                    cnd_lists.append(i)
    return cnd_lists

=====
# (step_4) generate type of each candidate alignment
=====
def gen_type_cnd(cnd_lists, CB):
    for i in cnd_lists:
        c_brnch= gen_cb_brnch(i, CB)
        type = gen_algnmnt_type(c_brnch)
        i.append(type)
    return cnd_lists

=====
# (step_4) find alignment type for given candidate alignment and its P(A)
=====
def gen_P_A(i, P_A):
    for key in P_A:
        if list(key) == i[-1]:
            i[-1].append(P_A[key])
    return i

```

```

=====
# (step_4) calculate P(ST|A_k)P(A_k) ==> for NB, change prod to sume of log
=====
def gen_prob (BEST):
    prod=1
    for i in BEST[:-1]:
        prod = prod * i[1]  ##### log

    prod= prod*BEST[-1][-1]  ##### log
    BEST.append(prod)

    return BEST

```

Appendix C

Definitions and examples of nine words from WordNet and COBUILD

a. *area.n*

CB = 6	WN = 6
<p>[NOUN] An area is a particular part of a town, a country, a region, or the world. the large number of community groups in the area 60 years ago half the French population still lived in rural areas.</p>	<p>(n) area, country (a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography)) <i>"it was a mountainous area"; "Bible country"</i></p>
<p>[NOUN] Your area is the part of a town, country, or region where you live. An organization's area is the part of a town, country, or region that it is responsible for. Local authorities have been responsible for the running of schools in their areas If there is an election in your area, you should go and vote.</p>	<p>(n) area (a subject of study) <i>"it was his area of specialization"; "areas of interest include..."</i></p>
<p>[NOUN] A particular area is a piece of land or part of a building that is used for a particular activity. a picnic area. the main check-in area located in Terminal 1.</p>	<p>(n) area, region (a part of an animal that has a special function or is supplied by a given artery or nerve) <i>"in the abdominal region"</i></p>
<p>[NOUN] An area is a particular place on a surface or object, for example on your body. You will notice that your baby has two soft areas on the top of his head.</p>	<p>(n) sphere, domain, area, orbit, field, arena (a particular environment or walk of life) <i>"his social sphere is limited"; "it was a closed area of employment"; "he's out of my orbit"</i></p>
<p>[NOUN] The area of a surface such as a piece of land is the amount of flat space or ground that it covers, measured in square units. The islands cover a total area of 625.6 square kilometers</p>	<p>(n) area (a part of a structure having some specific characteristic or function) <i>"the spacious cooking area provided plenty of room for servants"</i></p>
<p>[NOUN] You can use area to refer to a particular subject or topic, or to a particular part of a larger, more general situation or activity. the politically sensitive area of old age pensions.</p>	<p>(n) area, expanse, surface area (the extent of a 2-dimensional surface enclosed within a boundary) <i>"the area of a rectangle"; "it was about 500 square feet in area"</i></p>

b. *community.n*

CB = 3	WN = 6
<p>[NOUN] The community is all the people who live in a particular area or place. He's well liked by people in the community The growth of such vigilante gangs has worried community leaders, police and politicians.</p>	<p>(n) community (a group of people living in a particular local area) <i>"the team is drawn from all parts of the community"</i></p>

[NOUN] A particular community is a group of people who are similar in some way. The police haven't really done anything for the black community in particular. the business community.	(n) community (common ownership) <i>"they shared a community of possessions"</i>
[NOUN] Community is friendship between different people or groups, and a sense of having something in common. Two of our greatest strengths are diversity and community.	(n) community (a group of nations having common interests) <i>"they hoped to join the NATO community"</i>
	(n) community , community of interests (agreement as to goals) <i>"the preachers and the bootleggers found they had a community of interests"</i>
	(n) residential district, residential area, community (a district where people live; occupied primarily by private residences)
	(n) community , biotic community ((ecology) a group of interdependent organisms inhabiting the same region and interacting with each other)

c. *indicate.v*

CB = 6	WN = 5
[VERB] If one thing indicates another, the first thing shows that the second is true or exists. A survey of retired people has indicated that most are independent and enjoying life Our vote today indicates a change in United States policy This indicates whether remedies are suitable for children.	(v) bespeak, betoken, indicate , point, signal (be a signal for or a symptom of) <i>"These symptoms indicate a serious illness"; "Her behavior points to a severe neurosis"; "The economic indicators signal that the euro is undervalued"</i>
[VERB] If you indicate an opinion, an intention, or a fact, you mention it in an indirect way. Mr. Rivers has indicated that he may resign U.S. authorities have not yet indicated their monetary policy plans.	(v) indicate , point, designate, show (indicate a place, direction, person, or thing; either spatially or figuratively) <i>"I showed the customer the glove section"; "He pointed to the empty parking space"; "he indicated his opponents"</i>
[VERB] [FORMAL] If you indicate something to someone, you show them where it is, especially by pointing to it. He indicated a chair. 'Sit down.'	(v) indicate (to state or express briefly) <i>"indicated his wishes in a letter"</i>
[VERB] If one thing indicates something else, it is a sign of that thing. Dreams can help indicate your true feelings	(v) argue, indicate (give evidence of) <i>"The evidence argues for your claim"; "The results indicate the need for more work"</i>
[VERB] If a technical instrument indicates something, it shows a measurement or reading. The needles that indicate your height are at the top right-hand corner	(v) indicate , suggest (suggest the necessity of an intervention; in medicine) <i>"Tetracycline is indicated in such cases"</i>

The temperature gauge indicated that it was boiling.	
[VERB] [mainly BRIT] When drivers indicate , they make lights flash on one side of their vehicle to show that they are going to turn in that direction. in AM, use signal He told us when to indicate and when to change gear.	

d. *involve.v*

CB = 5	WN = 7
[VERB] If a situation or activity involves something, that thing is a necessary part or consequence of it. Running a kitchen involves a great deal of discipline and speed Nicky's job as a public relations director involves spending quite a lot of time with other people.	(v) involve , affect, regard (connect closely and often incriminatingly) <i>"This new ruling affects your business"</i>
[VERB] If a situation or activity involves someone, they are taking part in it. If there was a cover-up, it involved people at the very highest levels of government.	(v) involve (engage as a participant) <i>"Don't involve me in your family affairs!"</i>
[VERB] If you say that someone involves themselves in something, you mean that they take part in it, often in a way that is unnecessary or unwanted. I seem to have involved myself in something I don't understand.	(v) imply, involve (have as a necessary feature) <i>"This decision involves many changes"</i>
[VERB] If you involve someone else in something, you get them to take part in it. Noel and I do everything together, he involves me in everything	(v) necessitate, ask, postulate, need, require, take, involve , call for, demand (require as useful, just, or proper) <i>"It takes nerve to do what she did"; "success usually requires hard work"; "This job asks a lot of patience and skill"; "This position demands a lot of personal sacrifice"; "This dinner calls for a spectacular dessert"; "This intervention does not postulate a patient's consent"</i>
[VERB] If one thing involves you in another thing, especially something unpleasant or inconvenient, the first thing causes you to do or deal with the second. A late booking may involve you in extra cost	(v) involve (contain as a part) <i>"Dinner at Joe's always involves at least six courses"</i>
	(v) involve (occupy or engage the interest of) <i>"His story completely involved me during the entire afternoon"</i>
	(v) involve (make complex or intricate or complicated) <i>"The situation was rather involved"</i>

e. *job.n*

CB = 5	WN = 12
<p>[NOUN] A job is the work that someone does to earn money. Once I'm in America I can get a job Thousands have lost their jobs I felt the pressure of being the first woman in the job. overseas job vacancies.</p>	<p>(n) occupation, business, job, line of work, line (the principal activity in your life that you do to earn money) <i>"he's not in my line of business"</i></p>
<p>[NOUN] A job is a particular task. He said he hoped that the job of putting together a coalition wouldn't take too much time</p>	<p>(n) job, task, chore (a specific piece of work required to be done as a duty or for a specific fee) <i>"estimates of the city's loss on that job ranged as high as a million dollars"; "the job of repairing the engine took several hours"; "the endless task of classifying the samples"; "the farmer's morning chores"</i></p>
<p>[NOUN] The job of a particular person or thing is their duty or function. Their main job is to preserve health rather than treat illness Drinking a lot helps the kidneys do their job.</p>	<p>(n) job (an object worked on; a result produced by working) <i>"he held the job in his left hand and worked on it with his right"</i></p>
<p>[NOUN] If you say that someone is doing a good job, you mean that they are doing something well. In British English, you can also say that they are making a good job of something. We could do a far better job of managing it than they have</p>	<p>(n) job (the responsibility to do something) <i>"it is their job to print the truth"</i></p>
<p>[NOUN] If you say that you have a job doing something, you are emphasizing how difficult it is. He may have a hard job selling that argument to investors</p>	<p>(n) job (the performance of a piece of work) <i>"she did an outstanding job as Ophelia"; "he gave it up as a bad job"</i></p>
	<p>(n) job (a damaging piece of work) <i>"dry rot did the job of destroying the barn"; "the barber did a real job on my hair"</i></p>
	<p>(n) problem, job (a state of difficulty that needs to be resolved) <i>"she and her husband are having problems"; "it is always a job to contact him"; "urban problems such as traffic congestion and smog"</i></p>
	<p>(n) Job (a Jewish hero in the Old Testament who maintained his faith in God in spite of afflictions that tested him)</p>
	<p>(n) Job (any long-suffering person who withstands affliction without despairing)</p>
	<p>(n) job ((computer science) a program application</p>

	that may consist of several steps but is a single logical unit)
	(n) Job , Book of Job (a book in the Old Testament containing Job's pleas to God about his afflictions and God's reply)
	(n) caper, job (a crime (especially a robbery)) <i>"the gang pulled off a bank job in St. Louis"</i>

f. *policy.n*

CB = 3	WN = 3
[NOUN] A policy is a set of ideas or plans that is used as a basis for making decisions, especially in politics, economics, or business. plans which include changes in foreign policy and economic reforms. the UN's policy-making body.	(n) policy (a plan of action adopted by an individual or social group) <i>"it was a policy of retribution"; "a politician keeps changing his policies"</i>
[NOUN] An official organization's policy on a particular issue or towards a country is their attitude and actions regarding that issue or country. the government's policy on repatriation. the corporation's policy of forbidding building on common land.	(n) policy (a line of argument rationalizing the course of action of a government) <i>"they debated the policy or impolicy of the proposed legislation"</i>
[NOUN] An insurance policy is a document which shows the agreement that you have made with an insurance company. You are advised to read the small print of household and motor insurance policies.	(n) policy , insurance policy, insurance (written contract or certificate of insurance) <i>"you should have read the small print on your policy"</i>

g. *process.n*

CB = 2	WN = 6
[NOUN] A process is a series of actions which are carried out in order to achieve a particular result. There was total agreement to start the peace process as soon as possible The best way to proceed is by a process of elimination.	(n) procedure, process (a particular course of action intended to achieve a result) <i>"the procedure of obtaining a driver's license"; "it was a process of trial and error"</i>
[NOUN] A process is a series of things which happen naturally and result in a biological or chemical change. It occurs in elderly men, apparently as part of the aging process	(n) process , cognitive process, mental process, operation, cognitive operation ((psychology) the performance of some composite cognitive activity; an operation that affects mental contents) <i>"the process of thinking"; "the cognitive operation of"</i>

	<i>remembering"</i>
	(n) summons, process (a writ issued by authority of law; usually compels the defendant's attendance in a civil suit; failure to appear results in a default judgment against the defendant)
	(n) process , unconscious process (a mental process that you are not directly aware of) <i>"the process of denial"</i>
	(n) process , outgrowth, appendage (a natural prolongation or projection from a part of an organism either animal or plant) <i>"a bony process"</i>
	(n) process , physical process (a sustained phenomenon or one marked by gradual changes through a series of states) <i>"events now in process"; "the process of calcification begins later for boys than for girls"</i>

h. *require.v*

CB = 2	WN = 4
[VERB] If you require something or if something is required , you need it or it is necessary. If you require further information, you should consult the registrar This isn't the kind of crisis that requires us to drop everything else Some of the materials required for this technique may be difficult to obtain.	(v) necessitate, ask, postulate, need, require , take, involve, call for, demand (require as useful, just, or proper) <i>"It takes nerve to do what she did"; "success usually requires hard work"; "This job asks a lot of patience and skill"; "This position demands a lot of personal sacrifice"; "This dinner calls for a spectacular dessert"; "This intervention does not postulate a patient's consent"</i>
[VERB] If a law or rule requires you to do something, you have to do it. The rules also require employers to provide safety training At least 35 manufacturers have flouted a law requiring prompt reporting of such malfunctions The law now requires that parents serve on the committees that plan and evaluate school programs Then he'll know exactly what's required of him.	(v) ask, require , expect (consider obligatory; request and expect) <i>"We require our secretary to be on time"; "Aren't we asking too much of these children?"; "I expect my students to arrive in time for their lessons"</i>
	(v) command, require (make someone do something)
	(v) want, need, require (have need of) <i>"This piano wants the attention of a competent tuner"</i>

i. *section.n*

CB = 3	WN = 14
[NOUN] A section of something is one of the parts into which it is divided or from which it is formed. He said it was wrong to single out any section of society for Aids testing They moulded a complete new bow section for the boat. a large orchestra, with a vast percussion section. the Georgetown section of Washington, D.C.	(n) section , subdivision (a self-contained part of a larger composition (written or musical)) <i>"he always turns first to the business section"; "the history of this work is discussed in the next section"</i>
[NOUN] A section of an official document such as a report, law, or constitution is one of the parts into which it is divided. section 14 of the Trade Descriptions Act 1968.	(n) section (a very thin slice (of tissue or mineral or other substance) for examination under a microscope) <i>"sections from the left ventricle showed diseased tissue"</i>
[NOUN] A section is a diagram of something such as a building or a part of the body. It shows how the object would appear to you if it were cut from top to bottom and looked at from the side. For some buildings a vertical section is more informative than a plan.	(n) section (a distinct region or subdivision of a territorial or political area or community or group of people) <i>"no section of the nation is more ardent than the South"; "there are three synagogues in the Jewish section"</i>
	(n) section , segment (one of several parts or pieces that fit with others to constitute a whole object) <i>"a section of a fishing rod"; "metal sections were used below ground"; "finished the final segment of the road"</i>
	(n) section (a small team of policemen working as part of a police platoon)
	(n) part, section , division (one of the portions into which something is regarded as divided and which together constitute a whole) <i>"the written part of the exam"; "the finance section of the company"; "the BBC's engineering division"</i>
	(n) section (a land unit equal to 1 square mile)
	(n) section , plane section ((geometry) the area created by a plane cutting through a solid)
	(n) section , discussion section (a small class of students who are part of a larger course but are taught separately) <i>"a graduate student taught sections for the professor's lecture course"</i>
	(n) section (a division of an orchestra containing all instruments of the same class)
	(n) section (a small army unit usually having a special function)
	(n) department, section (a specialized division of a

	large organization) <i>"you'll find it in the hardware department"; "she got a job in the historical section of the Treasury"</i>
	(n) section (a segment of a citrus fruit) <i>"he ate a section of the orange"</i>
	(n) incision, section , surgical incision (the cutting of or into body tissues or organs (especially by a surgeon as part of an operation))

Appendix D

Scores for nine words

a. area.n

area	CB1	CB2	CB3	CB4	CB5	CB6
WN1	0.82	0.54	0.54	0.14	0.04	-0.54
WN2	-0.45	-0.50	-0.45	-0.45	-0.59	0.86
WN3	-0.07	-0.07	-0.07	0.71	-0.21	0.29
WN4	-0.42	-0.17	0.00	-0.58	-0.83	0.08
WN5	-0.25	-0.19	0.88	-0.06	-0.50	-0.50
WN6	-0.36	-0.21	-0.36	-0.14	0.71	-0.50

b. community.n

community	CB1	CB2	CB3
WN1	0.94	0.22	-0.28
WN2	-0.07	0.00	0.14
WN3	-0.23	0.64	0.36
WN4	-0.50	0.29	0.50
WN5	0.67	-0.50	-0.50
WN6	0.38	-0.06	-0.06

c. indicate.v

indicate	CB1	CB2	CB3	CB4	CB5	CB6
WN1	0.29	-0.07	-0.07	1.00	0.21	-0.07
WN2	-0.22	-0.11	0.94	-0.33	-0.17	0.06
WN3	-0.11	0.46	-0.11	-0.11	0.18	-0.04
WN4	1.00	-0.18	0.09	0.45	0.50	-0.27
WN5	-0.36	-0.36	-0.36	-0.57	-0.64	-0.43

d. involve.v

involve	CB1	CB2	CB3	CB4	CB5
WN1	0.08	0.50	0.33	0.08	0.25
WN2	-0.06	0.75	0.69	0.50	0.13
WN3	1.00	-0.14	-0.29	-0.14	-0.43
WN4	0.61	-0.11	-0.44	-0.06	-0.39
WN5	0.89	0.68	0.25	0.36	-0.07
WN6	-0.45	0.23	0.23	0.41	-0.32
WN7	-0.57	-0.14	0.29	-0.21	0.00

e. job.n

job	CB1	CB2	CB3	CB4	CB5
WN1	1.00	-0.17	0.42	-0.83	-0.67
WN2	0.19	0.56	0.75	-0.31	-0.31
WN3	-0.14	0.00	-0.50	-0.36	-0.29
WN4	0.11	0.22	0.72	-0.33	-0.44
WN5	0.12	0.38	0.42	0.58	-0.08
WN6	-0.32	-0.05	-0.36	0.05	0.05
WN7	-0.36	0.36	-0.07	-0.14	0.93
WN8	-1.00	-1.00	-1.00	-1.00	-0.83
WN9	-0.75	-1.00	-1.00	-0.88	-0.88
WN10	-0.50	0.21	-0.36	-0.71	-0.86
WN11	-1.00	-1.00	-1.00	-1.00	-1.00
WN12	-0.19	0.19	-0.73	-0.81	-0.65

f. policy.n

policy	CB1	CB2	CB3
WN1	0.95	0.64	0.09
WN2	0.64	0.50	-0.64
WN3	0.08	0.00	1.00

g. process.v

process	CB1	CB2
WN1	1.00	0.19
WN2	-0.14	0.50
WN3	-0.50	-0.83
WN4	-0.23	0.19
WN5	-0.41	0.32
WN6	0.57	0.57

h. require.v

require	CB1	CB2
WN1	0.67	0.08
WN2	0.50	0.88
WN3	-0.07	0.79
WN4	1.00	0.11

i. section.n

section	CB1	CB2	CB3
WN1	0.65	0.77	-0.35
WN2	0.23	-0.32	0.32
WN3	0.64	-0.36	-0.64
WN4	1.00	0.17	-0.25
WN5	-0.06	-0.56	-0.56
WN6	1.00	0.36	-0.14
WN7	0.00	0.00	-0.78
WN8	-0.15	-0.46	0.54
WN9	0.59	-0.23	-0.32
WN10	0.14	-0.36	-0.71
WN11	0.25	-0.50	-0.50
WN12	0.63	-0.25	-0.56
WN13	0.14	-0.36	-0.43
WN14	-0.11	-0.33	0.22

Appendix E

Results of the initial alignment (*based on SR::AW*) and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *all* positive; improvements in bold)

word	Human alignment (<i>all</i> positive, $\alpha=0.5$)	Initial alignment (based on SR::AW)		Adjusted alignment			
			P	R		P	R
area	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c6) (w3,c4), (w3,c6), (w4,c3) (w4,c6), (w5,c3), (w6,c5)	(w1,c1),(w2,c4),(w3,c5), (w4,c6), (w5,c2), (w6,c4)	0.33	0.18	(w1,c1),(w2,c4),(w3,c5), (w4,c6), (w5,c2), (w6,c3)	0.33	0.18
commu- nity	(w1,c1), (w1,c2), (w2,c2) (w2,c3), (w3,c2), (w3,c3) (w4,c2), (w4,c3), (w5,c1) (w6,c1)	(w1,c1),(w2,c2),(w3,c2), (w4,c2), (w5,c3), (w6,c1)	0.67	0.44	(w1,c1),(w2,c3),(w3,c2), (w4,c2), (w5,c3), (w6,c1)	0.83 ↑	0.56 ↑
indicate	(w1,c1), (w1,c4), (w1,c5) (w2,c3), (w2,c6), (w3,c2) (w3,c5), (w4,c1), (w4,c3) (w4,c4), (w4,c5)	(w1,c2),(w2,c3),(w3,c6), (w4,c2), (w5,c6)	0.20	0.09	(w1,c2),(w2,c3),(w3,c1), (w4,c4), (w5,c6)	0.40 ↑	0.18 ↑
involve	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c2) (w2,c3), (w2,c4), (w2,c5) (w3,c1), (w4,c1), (w5,c1) (w5,c2), (w5,c3), (w5,c4) (w6,c2), (w6,c3), (w6,c4) (w7,c3), (w7,c6)	(w1,c1),(w2,c2),(w3,c3), (w4,c5),(w5,c1), (w6,c2), (w7,c4)	0.57	0.21	(w1,c1),(w2,c2),(w3,c3), (w4,c5),(w5,c1), (w6,c2), (w7,c4)	0.57	0.21
job	(w1,c1),(w1,c3), (w2,c1) (w2,c2), (w2,c3), (w3,c2) (w4,c1), (w4,c2), (w4,c3) (w5,c1), (w5,c2), (w5,c3) (w5,c4), (w6,c4), (w6,c5) (w7,c2),(w7,c5),(w10,c2) (w12,c2)	(w1,c5),(w2,c3),(w3,c2), (w4,c2),(w5,c1),(w6,c4) (w7,c3), (w8,c5),(w9,c5) (w10,c5), (w11,c3), (w12,c4)	0.33	0.22	(w1,c5),(w2,c3),(w3,c2), (w4,c2),(w5,c1),(w6,c4) (w7,c3), (w8,c5),(w9,c1) (w10,c5), (w11,c3), (w12,c4)	0.33	0.22
policy	(w1,c1), (w1,c2), (w1,c3) (w2,c1), (w2,c2), (w3,c1) (w3,c2), (w3,c3)	(w1,c1),(w2,c2),(w3,c3)	1.00	0.43	(w1,c1),(w2,c2),(w3,c3)	1.00	0.43
process	(w1,c1), (w1,c2), (w2,c2) (w4,c2), (w5,c2), (w6,c1) (w6,c2)	(w1,c1),(w2,c1),(w3,c1), (w4,c1), (w5,c2), (w6,c2)	0.50	0.43	(w1,c1),(w2,c2),(w3,c1), (w4,c1), (w5,c2), (w6,c2)	0.67 ↑	0.57 ↑
require	(w1,c1), (w1,c2), (w2,c1) (w2,c2),(w3,c2), (w4,c1) (w4,c2)	(w1,c2),(w2,c2),(w3,c2), (w4,c1)	1.00	0.57	(w1,c2),(w2,c2),(w3,c1), (w4,c1)	0.75 ↓	0.43 ↓
section	(w1,c1), (w1,c2), (w2,c1) (w2,c3),(w3,c1), (w4,c1) (w4,c2), (w6,c1), (w6,c2) (w7,c1), (w7,c2), (w8,c3) (w9,c1),(w10,c1)(w11c1) (w12c1)(w13c1)(w14c3)	(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c2), (w7,c1),(w8,c1),(w9,c1), (w10,c2)(w11c2)(w12c1) (w13,c1),(w14,c1)	0.64	0.56	(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c2), (w7,c1),(w8,c1),(w9,c1), (w10,c2)(w11c2)(w12c1) (w13,c1),(w14,c1)	0.57 ↓	0.50 ↓

Appendix F

- a. Results of the initial alignment (*based on SL*) and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *all* positive; improvements in bold)

word	Human alignment (<i>all</i> positive, $\alpha=0.5$)	Initial alignment (based on SR::AW)			Adjusted alignment		
			P	R		P	R
area	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c6) (w3,c4), (w3,c6), (w4,c3) (w4,c6), (w5,c3), (w6,c5)	(w1,c1),(w2,c5),(w3,c5), (w4,c4), (w5,c3), (w6,c5)	0.50	0.27	(w1,c1),(w2,c5),(w3,c5), (w4,c4), (w5,c3), (w6,c5)	0.50	0.27
commu- nity	(w1,c1), (w1,c2), (w2,c2) (w2,c3), (w3,c2), (w3,c3) (w4,c2), (w4,c3), (w5,c1) (w6,c1)	(w1,c3),(w2,c2),(w3,c1), (w4,c2), (w5,c1), (w6,c1)	0.50	0.33	(w1,c3),(w2,c2),(w3,c1), (w4,c2), (w5,c1), (w6,c1)	0.50	0.33
indicate	(w1,c1), (w1,c4), (w1,c5) (w2,c3), (w2,c6), (w3,c2) (w3,c5), (w4,c1), (w4,c3) (w4,c4), (w4,c5)	(w1,c3),(w2,c5),(w3,c2), (w4,c4), (w5,c2)	0.40	0.18	(w1,c3),(w2,c5),(w3,c2), (w4,c4), (w5,c2)	0.40	0.18
involve	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c2) (w2,c3), (w2,c4), (w2,c5) (w3,c1), (w4,c1), (w5,c1) (w5,c2), (w5,c3), (w5,c4) (w6,c2), (w6,c3), (w6,c4) (w7,c3), (w7,c6)	(w1,c5),(w2,c5),(w3,c1), (w4,c1),(w5,c1)	1.00	0.26	(w1,c5),(w2,c5),(w3,c4), (w4,c1),(w5,c1)	0.80	0.21
job	(w1,c1),(w1,c3), (w2,c1) (w2,c2), (w2,c3), (w3,c2) (w4,c1), (w4,c2), (w4,c3) (w5,c1), (w5,c2), (w5,c3) (w5,c4), (w6,c4), (w6,c5) (w7,c2),(w7,c5),(w10,c2) (w12,c2)	(w1,c5),(w2,c2),(w3,c1), (w4,c3),(w5,c1),(w6,c1)	0.50	0.17	(w1,c5),(w2,c2),(w3,c1), (w4,c3),(w5,c1),(w6,c1)	0.50	0.17
policy	(w1,c1), (w1,c2), (w1,c3) (w2,c1), (w2,c2), (w3,c1) (w3,c2), (w3,c3)	(w1,c1),(w2,c2),(w3,c1)	1.00	0.43	(w1,c1),(w2,c2),(w3,c1)	1.00	0.43
process	(w1,c1), (w1,c2), (w2,c2) (w4,c2), (w5,c2), (w6,c1) (w6,c2)	(w1,c2),(w2,c1),(w3,c1)	0.33	0.14	(w1,c2),(w2,c1),(w3,c1)	0.33	0.14
require	(w1,c1), (w1,c2), (w2,c1) (w2,c2),(w3,c2), (w4,c1) (w4,c2)	(w1,c2),(w2,c1),(w3,c1), (w4,c1)	0.75	0.43	(w1,c2),(w2,c2),(w3,c1), (w4,c1)	0.75	0.43
section	(w1,c1), (w1,c2), (w2,c1) (w2,c3),(w3,c1), (w4,c1) (w4,c2), (w6,c1), (w6,c2) (w7,c1), (w7,c2), (w8,c3) (w9,c1),(w10,c1)(w11c1) (w12c1)(w13c1)(w14c3)	(w1,c2),(w2,c1),(w3,c1), (w4,c1),(w5,c1),(w6,c1)	0.83	0.31	(w1,c2),(w2,c1),(w3,c1), (w4,c1),(w5,c1),(w6,c1)	0.83	0.31

b. Results of the initial alignment (*based on SL*) and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *top* positive; improvements in bold)

word	Human alignment (<i>top</i> positive, $\alpha=0.5$)	Initial alignment (based on <i>SL</i>)			Adjusted alignment		
			P	R		P	R
area	(w1,c1),(w2,c6),(w3,c4), (w4,c6), (w5,c3), (w6,c5)	(w1,c1),(w2,c5),(w3,c5), (w4,c4), (w5,c3), (w6,c5)	0.50	0.50	(w1,c1),(w2,c5),(w3,c5), (w4,c4), (w5,c3), (w6,c5)	0.50	0.50
commu- nity	(w1,c1),(w2,c3),(w3,c2), (w4,c3), (w5,c1), (w6,c1)	(w1,c3),(w2,c2),(w3,c1), (w4,c2), (w5,c1), (w6,c1)	0.33	0.33	(w1,c3),(w2,c2),(w3,c1), (w4,c2), (w5,c1), (w6,c1)	0.33	0.33
indicate	(w1,c4),(w2,c3),(w3,c2), (w4,c1), (w5,c1)	(w1,c3),(w2,c5),(w3,c2), (w4,c4), (w5,c2)	0.20	0.25	(w1,c3),(w2,c5),(w3,c2), (w4,c4), (w5,c2)	0.20	0.25
involve	(w1,c2),(w2,c2),(w3,c1), (w4,c1),(w5, c1), (w6,c4), (w7,c3)	(w1,c5),(w2,c5),(w3,c1), (w4,c1),(w5, c1)	0.60	0.43	(w1,c5),(w2,c5),(w3,c4), (w4,c1),(w5, c1)	0.40	0.29
job	(w1,c1),(w2,c3),(w3,c2), (w4,c3),(w5, c4),(w6, c4) (w7,c5), (w8,c5),(w9, c1) (w10,c2), (w11,c1), (w12, c2)	(w1,c5),(w2,c2),(w3,c1), (w4,c3),(w5, c1),(w6, c1)	0.17	0.11	(w1,c5),(w2,c2),(w3,c1), (w4,c3),(w5, c1),(w6, c1)	0.17	0.11
policy	(w1,c1),(w2,c1),(w3,c3)	(w1,c1),(w2,c2),(w3,c1)	0.33	0.33	(w1,c1),(w2,c2),(w3,c1)	0.33	0.33
process	(w1,c1),(w2,c2),(w3,c1), (w4,c2), (w5,c2), (w6,c1)	(w1,c2),(w2,c1),(w3,c1)	0.00	0.00	(w1,c2),(w2,c1),(w3,c1)	0.00	0.00
require	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	(w1,c2),(w2,c1),(w3,c1), (w4,c1)	0.25	0.25	(w1,c2),(w2,c2),(w3,c1), (w4,c1)	0.50	0.50
section	(w1,c2),(w2,c3),(w3,c1), (w4,c1),(w5,c1),(w6,c1), (w7,c1),(w8,c3),(w9,c1), (w10,c1)(w11c1)(w12c1) (w13,c1),(w14,c3)	(w1,c2),(w2,c1),(w3,c1), (w4,c1),(w5,c1),(w6,c1)	0.67	0.33	(w1,c2),(w2,c1),(w3,c1), (w4,c1),(w5,c1),(w6,c1)	0.67	0.33

c. Results of the initial alignment (*based on NB*) and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *all* positive; improvements in bold)

word	Human alignment (<i>all</i> positive, $\alpha=0.5$)	Initial alignment (based on SR::AW)			Adjusted alignment		
			P	R		P	R
area	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c6) (w3,c4), (w3,c6), (w4,c3) (w4,c6), (w5,c3), (w6,c5)	(w1,c5),(w2,c4),(w3,c1), (w4,c6), (w5,c5)	0.40	0.18	(w1,c5),(w2,c4),(w3,c1), (w4,c6), (w5,c5)	0.40	0.18
commu- nity	(w1,c1), (w1,c2), (w2,c2) (w2,c3), (w3,c2), (w3,c3) (w4,c2), (w4,c3), (w5,c1) (w6,c1)	(w1,c2),(w2,c1),(w3,c1), (w4,c2)	0.50	0.22	(w1,c2),(w2,c3),(w3,c1), (w4,c2)	0.75	0.33
indicate	(w1,c1), (w1,c4), (w1,c5) (w2,c3), (w2,c6), (w3,c2) (w3,c5), (w4,c1), (w4,c3) (w4,c4), (w4,c5)	(w1,c2),(w2,c2),(w3,c1), (w4,c1)	0.25	0.09	(w1,c2),(w2,c2),(w3,c3), (w4,c1)	0.25	0.09
involve	(w1,c1), (w1,c2), (w1,c3) (w1,c4), (w1,c5), (w2,c2) (w2,c3), (w2,c4), (w2,c5) (w3,c1), (w4,c1), (w5,c1) (w5,c2), (w5,c3), (w5,c4) (w6,c2), (w6,c3), (w6,c4) (w7,c3), (w7,c6)	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3), (w6,c5)	0.67	0.21	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3),(w6,c5)	0.67	0.21
job	(w1,c1),(w1,c3), (w2,c1) (w2,c2), (w2,c3), (w3,c2) (w4,c1), (w4,c2), (w4,c3) (w5,c1), (w5,c2), (w5,c3) (w5,c4), (w6,c4), (w6,c5) (w7,c2),(w7,c5),(w10,c2) (w12,c2)	(w1,c2),(w2,c4),(w3,c5), (w4,c4),(w5,c5),(w6,c4) (w7,c4), (w8,c1)	0.13	0.06	(w1,c2),(w2,c4),(w3,c5), (w4,c4),(w5,c5),(w6,c4) (w7,c4), (w8,c1)	0.13	0.06
policy	(w1,c1), (w1,c2), (w1,c3) (w2,c1), (w2,c2), (w3,c1) (w3,c2), (w3,c3)	(w1,c1),(w2,c1)	1.00	0.29	(w1,c3),(w2,c1)	1.00	0.29
process	(w1,c1), (w1,c2), (w2,c2) (w4,c2), (w5,c2), (w6,c1) (w6,c2)	(w1,c1),(w2,c1),(w3,c1), (w4,c1)	0.25	0.14	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	0.50	0.29
require	(w1,c1), (w1,c2), (w2,c1) (w2,c2),(w3,c2), (w4,c1) (w4,c2)	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	1.00	0.57	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	1.00	0.57
section	(w1,c1), (w1,c2), (w2,c1) (w2,c3),(w3,c1), (w4,c1) (w4,c2), (w6,c1), (w6,c2) (w7,c1), (w7,c2), (w8,c3) (w9,c1),(w10,c1)(w11c1) (w12c1)(w13c1)(w14c3)	(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c1), (w7,c1)	0.71	0.31	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3),(w6,c1), (w7,c1)	0.71	0.31

d. Results of the initial alignment (*based on NB*) and adjusted alignment over nine words ($\alpha=0.5$, gold standard = *top* positive; improvements in bold)

word	Human alignment (top positive, $\alpha=0.5$)	Initial alignment (based on <i>NB</i>)			Adjusted alignment		
			P	R		P	R
area	(w1,c1),(w2,c6),(w3,c4), (w4,c6), (w5,c3), (w6,c5)	(w1,c5),(w2,c4),(w3,c1), (w4,c6), (w5,c5)	0.20	0.17	(w1,c5),(w2,c4),(w3,c1), (w4,c6), (w5,c5)	0.20	0.17
commu- nity	(w1,c1),(w2,c3),(w3,c2), (w4,c3), (w5,c1), (w6,c1)	(w1,c2),(w2,c1),(w3,c1), (w4,c2)	0.00	0.00	(w1,c2),(w2,c3),(w3,c1), (w4,c2)	0.25	0.17
Indicate	(w1,c4),(w2,c3),(w3,c2), (w4,c1), (w5,c1)	(w1,c2),(w2,c2),(w3,c1), (w4,c1)	0.25	0.25	(w1,c2),(w2,c2),(w3,c3), (w4,c1)	0.25	0.25
involve	(w1,c2),(w2,c2),(w3,c1), (w4,c1),(w5,c1), (w6,c4), (w7,c3)	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3), (w6,c5)	0.33	0.29	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3),(w6,c5)	0.33	0.29
job	(w1,c1),(w2,c3),(w3,c2), (w4,c3),(w5,c4),(w6,c4) (w7,c5), (w8,c5),(w9,c1) (w10,c2), (w11,c1), (w12,c2)	(w1,c2),(w2,c4),(w3,c5), (w4,c4),(w5,c5),(w6,c4) (w7,c4), (w8,c1)	0.13	0.11	(w1,c2),(w2,c4),(w3,c5), (w4,c4),(w5,c5),(w6,c4) (w7,c4), (w8,c1)	0.13	0.11
policy	(w1,c1),(w2,c1),(w3,c3)	(w1,c1),(w2,c1)	1.00	0.67	(w1,c3),(w2,c1)	0.50	0.33
process	(w1,c1),(w2,c2),(w3,c1), (w4,c2), (w5,c2), (w6,c1)	(w1,c1),(w2,c1),(w3,c1), (w4,c1)	0.25	0.17	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	0.50	0.33
require	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	1.00	1.00	(w1,c1),(w2,c2),(w3,c2), (w4,c1)	1.00	1.00
section	(w1,c2),(w2,c3),(w3,c1), (w4,c1),(w5,c1),(w6,c1), (w7,c1),(w8,c3),(w9,c1), (w10,c1)(w11,c1)(w12,c1) (w13,c1),(w14,c3)	(w1,c2),(w2,c1),(w3,c1), (w4,c2),(w5,c1),(w6,c1), (w7,c1)	0.43	0.25	(w1,c2),(w2,c3),(w3,c1), (w4,c2),(w5,c3),(w6,c1), (w7,c1)	0.57	0.33

Appendix G

Precision/Recall of WSA based on SenseLearner (SL) and Naïve Bayes (NB)

a. Precision & recall of nine words processed by WSA based on SL

WSA (SL)		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 0$	
		AP	TP	AP	TP	AP	TP
Area	P	0.67	0.50	0.500	0.500	0.50	0.50
	R	0.21	0.50	0.273	0.500	0.43	0.60
community	P	0.67	0.33	0.500	0.333	0.50	0.33
	R	0.31	0.33	0.333	0.333	0.43	0.40
indicate	P	0.80	0.20	0.400	0.200	0.40	0.20
	R	0.21	0.25	0.182	0.250	0.29	0.25
involve	P	0.60	0.40	0.800	0.400	0.40	0.40
	R	0.12	0.29	0.211	0.286	0.18	0.33
job	P	0.50	0.17	0.500	0.167	0.33	0.33
	R	0.13	0.10	0.167	0.111	0.17	0.29
policy	P	1.00	0.33	1.000	0.333	0.67	0.33
	R	0.38	0.33	0.429	0.333	0.40	0.33
process	P	0.33	0.00	0.333	0.000	0.00	0.00
	R	0.14	0.00	0.143	0.000	0.00	0.00
require	P	1.00	0.75	0.750	0.500	0.50	0.50
	R	0.50	0.75	0.429	0.500	0.40	0.50
section	P	1.00	0.83	0.833	0.667	0.67	0.50
	R	0.29	0.36	0.313	0.333	0.36	0.38

b. Precision & recall of nine words processed by WSA based on NB

WSA (NB)		$\alpha = 1$		$\alpha = 0.5$		$\alpha = 0$	
		AP	TP	AP	TP	AP	TP
area	P	0.60	0.20	0.400	0.200	0.00	0.00
	R	0.16	0.17	0.182	0.167	0.00	0.00
community	P	0.75	0.25	0.750	0.250	0.25	0.00
	R	0.23	0.17	0.333	0.167	0.14	0.00
indicate	P	0.75	0.25	0.250	0.250	0.25	0.25
	R	0.16	0.25	0.091	0.250	0.14	0.25
involve	P	0.83	0.50	0.667	0.333	0.50	0.50
	R	0.20	0.43	0.211	0.286	0.27	0.50
job	P	0.38	0.13	0.125	0.125	0.00	0.00
	R	0.13	0.10	0.056	0.111	0.00	0.00
policy	P	1.00	0.50	1.000	0.500	0.50	0.50
	R	0.25	0.33	0.286	0.333	0.20	0.33
process	P	0.50	0.50	0.500	0.500	0.50	0.50
	R	0.29	0.33	0.286	0.333	0.40	0.40
require	P	1.00	1.00	1.000	1.000	1.00	1.00
	R	0.50	1.00	0.571	1.000	0.80	1.00
section	P	0.86	0.43	0.714	0.571	0.71	0.43
	R	0.29	0.21	0.313	0.333	0.45	0.38

Appendix H

Two reading texts for the empirical study

a. Fashion Victim

Fashion Victim or Environmental Victory?

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 led to 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?

Clothing and textiles are the fastest growing waste product in Britain. A full 63 percent of unwanted clothes end up in landfills, while only 16 percent are recycled. This has grave consequences for the environment because many of today's clothes are made from synthetic fibers, which do not decompose naturally. This can create problems; for example, water cannot flow into the soil, and chemicals from the fabrics release poisons into the surrounding air and water.

There are other options for our unwanted clothing. Many of us ease our conscience about all this waste by donating our surplus clothing to charities. However, even charities cannot keep up with the turnover of clothing. They often end up selling the excess for recycling or export. Since 1989, U.S. exports of used clothing and other worn textiles have more than tripled and now exceed 7 billion pounds per year. Many of these clothes end up in the flourishing markets for secondhand clothes in Africa and Eastern Europe. Clothing in good condition is highly sought after and provides consumers there with the opportunity to be smartly dressed.

However, this cycle of developed countries providing clothes for those in poorer countries is only sustainable if the clothing is durable and resilient. Unfortunately, the primary reason that fast fashion clothes are so economical is because of a decline in the quality of materials and manufacturing. This presents less reason for developing countries to buy the poorly made clothing, and so it will eventually lead to a collapse in demand for Western castoffs.

So we need to take a closer look at the economic and ecological impact of fast fashion. Both government bodies and the fashion federations are demanding environmentally friendly approaches to fashion. For example, the British government used the occasion of London Fashion Week to launch its sustainable clothing action plan. The plan encourages everyone to consider the impact of a fashion item, from the cradle (its design) to the grave (its disposal).

b. Sleep Research

Sleep Research

To conduct some forms of sleep research, we have to find a way to track sleepiness during the day. Some people might believe that measuring sleepiness is a fairly trivial task. For instance, couldn't you simply count the number of times a person yawns each hour or so?

Yawning is a slow, exaggerated opening of the mouth with a long, deep inhalation of air, and then a brief exhalation. In most people's minds, yawning is the most obvious sign of sleepiness. It is a common behavior among many animals, including dogs, cats, crocodiles, snakes, birds, and even some fish. It is certainly true that sleepy people tend to yawn more than people who are fully awake. It is also true that when people say that they are bored, they yawn more frequently. However, we do not know whether yawning is a sign that you are getting ready for sleep or that you are successfully resisting sleep. In fact, simply stretching your body will often trigger a yawn.

Unfortunately, yawns don't just indicate sleepiness. In some animals, yawning is a sign of stress. When a trainer sees a dog yawning in a dog obedience class, it is usually a sign that the animal is under pressure. Perhaps the handler is pushing too hard or moving too fast and the dog feels agitated. Playing for a moment and then turning to another activity is usually enough to banish yawning for a while.

Yawning can also be a sign of stress in humans. For example, military officers report that soldiers often yawn before they jump out of an airplane with a parachute for the first time, even if they have recently drunk a lot of coffee. This is not because they are bored, obviously; it is probably because they are tense.

There is also a social aspect to yawning. Psychologists have done experiments that involve asking actors to yawn deliberately in crowded rooms and auditoriums. Within moments, there is usually an increase in yawning by everyone else in the room. Similarly, when people watch films or videos that show other people yawning, they are more likely to yawn themselves. Even just reading about yawning tends to stimulate people to yawn.

The truth of the matter is that we really don't know what purpose yawning serves. Scientists originally thought that the purpose of yawning was to increase the amount of oxygen in the blood or to release some accumulated carbon dioxide (CO₂). We now know that this is not true. If the concentration of CO₂ in the air increases, it doesn't make people more likely to yawn; it makes them breathe faster to try to get more oxygen.

In conclusion, yawning seems to be associated with a lot more than the need for sleep. Therefore, we obviously have to find some other measure of sleepiness. Some researchers have simply tried to ask people how sleepy they feel at particular times. However, there are problems with getting people to make these types of judgments. Sometimes people simply lie to the researchers. This occurs because admitting fatigue is considered a sign of weakness or a lack of ambition. Other times, people do not realize how sleepy they are because drink so much coffee. That is why many researchers have developed an alternate method to determine how sleepy a person is. It is based upon a simple definition of sleep need: The sleepier you are, the faster you will fall asleep if you are given the opportunity to do so.

Appendix I

Reading comprehension tests

a. Fashion Victim

Multiple Choice

1. What does the phrase “fast fashion” refer to? *(Circle only the best answer.)*
 - a. cheap and simple clothing that can be produced in factories very quickly
 - b. new styles of clothing that appear in stores quickly so that people can buy the latest designs
 - c. clothing that catches people’s attention quickly and makes them want to buy it right away
 - d. styles of clothing that match each other so that outfits can be created quickly

2. Which of the following are positive results of fast fashion? *(Circle all of the benefits that were mentioned in the article.)*
 - a. More and more people are learning how to design and sew their own clothes.
 - b. People can have the latest styles without spending a lot of money.
 - c. The economy is growing because people are buying a lot of clothes.
 - d. Fashion designers are becoming more and more famous around the world.
 - e. Dry cleaners and other laundry services are making more money.

3. Which of the following are negative results of fast fashion? *(Circle all of the problems that were mentioned in the article.)*
 - a. People are throwing out their old clothes.
 - b. People who enjoy shopping for clothes are almost going bankrupt.
 - c. Clothing prices are falling too much.
 - d. The quality of clothing is decreasing.
 - e. Unwanted clothing is causing environmental pollution.

4. According to the article, what might happen in the future in relation to fast fashion? *(Circle all of the future possibilities that were mentioned in the article.)*
 - a. People in developing countries (e.g., in Africa) might not want to buy secondhand Western clothing anymore.
 - b. The British government’s plan will make clothing companies pay for the environmental problems that they have caused.
 - c. Charities in the U.S. will need more donations of clothing because people are exporting so much clothing to other countries.
 - d. The British government’s plan will make people more aware of the consequences of their fashion choices.

True or false?

According to the article...

1. _____ Fast fashion clothing is usually inexpensive.
2. _____ Fast fashion is for people who want shopping to be convenient and don't care much about having the latest styles.
3. _____ Some old clothing is recycled, but most of it is thrown away.
4. _____ The U.S. has been exporting a lot of clothing to Africa and Europe.
5. _____ These days, more and more clothing is being made with natural materials.
6. _____ The main environmental problem with fast fashion is the fact that clothing factories are polluting the air.

b. Sleep Research

Multiple Choice

1. Based specifically on the information in the article, which of the following can you infer to be correct? *(Circle only the best answer.)*
 - a. Some people feel even more tired after they yawn.
 - b. Some people have difficulty falling asleep when they yawn a lot.
 - c. Some people yawn on purpose to show other people that they are tired.
 - d. Some people yawn whether they are tired or not.

2. Which of the following feelings make a person more likely to yawn? *(Circle all of the feelings that were mentioned in the article as being associated with yawning.)*
 - a. boredom
 - b. sleepiness
 - c. annoyance
 - d. stress
 - e. weakness

3. Which of the following experiences make a person more likely to yawn? *(Circle all of the experiences that were mentioned in the article as being associated with yawning.)*
 - a. not having enough O₂ (oxygen)
 - b. reading about yawning
 - c. seeing another person yawn
 - d. stretching his or her body
 - e. breathing too much CO₂ (carbon dioxide)

4. Why don't scientists trust people's reports of how sleepy they are? *(Circle all of the reasons that were mentioned in the article.)*
 - a. People sometimes lie because they do not want to admit that they are addicted to coffee.
 - b. People sometimes lie because they do not want to seem weak.
 - c. People sometimes do not realize how sleepy they are because they drink a lot of coffee.
 - a. People sometimes do not realize how sleepy they are because stress makes them feel awake.

True or false?

According to the article...

1. _____ Scientists agree that yawning doesn't mean that a person is getting ready for sleep; instead, it means that the person is trying not to fall asleep.
2. _____ Humans, dogs, and cats sometimes yawn, but other animals usually do not yawn.
3. _____ When a dog yawns in obedience class, the trainer will often push him a little bit harder in order to help him wake up.
4. _____ Scientists believe that the most reliable method of measuring sleepiness involves actually watching people's behavior.
5. _____ To measure sleepiness accurately, scientists agree that one good method is counting the number of times a person yawns.
6. _____ In an experiment, psychologists asked actors to yawn in a crowded room to see if that would make other people more likely to fall asleep soon after.

Appendix J

Pretest

On each of the following pages, you will see a word bank at the top of the page. Below that, you will see several sentences with blanks in them. Use the words from the word bank to fill in the blanks in the sentences. You may use the same word more than once, and some of the words will not be used. In some cases, we have helped you by adding grammatical information to the blanks (e.g., [-d/-ed] for past tense).

각 페이지마다 맨위에 word bank 가 주어지고, 각 word bank 아래 빈칸이 들어있는 문장이 주어집니다. 여러분은 word bank 에 주어진 단어들을 이용하여 각 문장의 빈칸을 채우면 됩니다. 주어진 단어들 중 두번 사용되거나, 사용되지 않는 단어도 있으니 주의하십시오. 단어를 채울때 문법이 변화될 경우(예: 동사의 시제 변화), 이미 문제에 변화된 문법의 형태가 주어져 있으니 참고하세요.

[Word Bank]

chain agitated unveil deed obscure sector alternate conscience cradle resilient delectable
expenditure mend outfit trigger

1. A lot of bad things have happened to that woman, but she is a very _____ person, so she is still strong, happy, and optimistic.
2. The young couple had reached the limit on their credit card, so they decided that they couldn't afford the _____ of money on a new TV set.
3. When my grandmother was young, her mother taught her how to _____ old socks that had holes in them so that she could keep wearing them.
4. In a TV interview last night, the owner of the sports team _____ (-d/-ed) his plans to build a new stadium, and everyone was surprised to hear the unexpected news.
5. There aren't many factory jobs these days, but people can still find jobs in the financial _____.
6. When people have been angry with each other many times in the past, sometimes even a small disagreement can suddenly _____ a large argument, and they start fighting.
7. Some people who care about supporting small businesses refuse to eat at large _____ restaurants that are owned by huge companies.
8. The small boy wondered what would happen if he kicked a frog, so he tried it, but then he had a guilty _____ because he was worried that he had hurt the frog.
9. Many scientists consider oceans and lakes to be the _____ of life because that is where the first living things developed.
10. Everyone was surprised by the man's colorful _____ because no one had seen him wear such bright clothing before.

[Word Bank]

mend deliberate resist trivial obedience intervention agitated fatigue consumerism
indicate banish inherit alternate discipline aspect

11. The woman was too _____ to think clearly because her son had just left home and joined the army.

12. After running 30 miles without stopping, the man was hospitalized for extreme _____ and then had to rest for the next 2 weeks.

13. In some cultures, it is important to show _____ and respect to older people, but in other cultures, independence is more valued.

14. People shouldn't worry about _____ things like missing their favorite TV show when there are more serious problems in the world.

15. It may be easy to forgive people when they have hurt you by accident, but it is more difficult to forgive them when they have planned to hurt you _____(-ly).

16. If traffic is really bad on the highway, it might be possible to save time by using (a/an) _____ road that takes you to the same place by a different route.

17. When people found out that their leader had stolen their tax money and put innocent men in prison, they _____(-d/-ed) him to another country and said that he could never come back.

18. Whenever the teenager saw a new video game advertised on TV, he couldn't _____ buying it; he was addicted to playing!

19. Unfortunately, the doctor told the man that his coughing and sneezing might _____ that he was suffering from a serious illness.

20. The professor thought about every _____ of the student's question before giving an answer because it was a very complicated issue.

[Word Bank]

facilitate unveil deliberate resist discipline sector correlation mend intervention delectable
fatigue indicate chain outfit alternate obscure

21. Sometimes when authors write in a very _____ and unclear way, readers get frustrated and bored because they can't understand what is happening in the story.

22. Most people know about the _____ between car accidents and talking on the phone while driving, but they talk on their phones while driving anyway.

23. After the man ended up in jail for being drunk and yelling at people in public, his friends decided that they should organize (a/an) _____ to get him to stop drinking so much.

24. Although the boss _____(-d/-ed) his employees for checking their personal email at work, they still often check their email while working.

25. It can be very difficult for students to understand fast speech in a foreign language, so teachers sometimes speak slowly in order to _____ their understanding and learning.

26. The old woman had broken her leg the previous year and didn't want to hurt it again, so she walked down the icy street very slowly and _____(-ly).

27. In the past before there were firefighters, if someone's house caught on fire, the people in the town would line up in a long _____ and pass buckets of water from person to person until they could put out the flames.

28. At night, the military _____ moved silently across the field under the cover of darkness.

29. It's easier for people to _____ their friendship after a serious disagreement if they have known each other for a long time and care a lot about each other.

30. The mother _____(-d/-ed) her anger with a frown and a sharp look so that she wouldn't attract attention by yelling at her daughter in public.

Appendix K

Vocabulary posttest-3 (Post-3)

3. We are going to show you the same words another time. You will see a word bank at the top of each page. Below each list of words, you will see a set of sentences with blanks in them. Some of the sentences are very similar to sentences that appeared in the reading passages, and some of them are not. For every sentence, please try to fill in the blank with a word from the list. (Sometimes you will see that the blank provides additional grammatical information—for example, to make a verb past tense so that it can fit.) Please try to make sure that the meaning of the word fits with the meaning of the sentence. Some of the words have more than one meaning, so they might fit in 2 different sentences. Some other words might not fit in any sentence, so you will not use them.

이번에는 같은 단어들 이 다른 방식으로 보여 집니다. 각 페이지 마다 맨 위 에 **word bank** 가 주어 지고, 각 **word bank** 아래 빈 칸 이 들어 있는 문장 이 주어 집니다. 어떤 단어 들은 여러분 이 독해 시 보 았 던 문장 들 과 매우 비슷 하고, 어떤 단어 들은 그렇지 않 습니다. **Word Bank** 에 주어 진 단어 들 을 이용 하여 각 문장 에 있는 빈 칸 을 채우 십시요. 빈 칸 에 들어 갈 단어 의 의미 가 전체 문장 의미 에 맞아 야 합니 다. (또한, 가끔 여러분 은 빈 칸 옆 에 추가 적인 문법 변화 를 볼 수 있 습니다- 예를 들 면, 동사 의 과거 형태 가 들어 가야 하는 경우, 그 문법 변화 가 자동 적으로 적용 될 수 있 도록 문법 변화 가 이미 주어 져 있 습니다) 마지막으로, 주어 진 단어 들 중 두 번 사용 되거나, 사용 되지 않는 단어 도 있으 니 주의 하 십시요.

[Word Bank]

intervention deliberate resist trivial obedience deed agitated fatigue indicate
commitment banish inherit alternate obscure aspect

1. Perhaps the dog trainer is pushing too hard or moving too fast and the dog feels nervous and _____.
2. Sometimes people simply lie to the researchers about how sleepy they are. This occurs because admitting _____ is considered a sign of weakness or a lack of ambition.
3. When a trainer sees a dog yawning in a dog _____ class, it is usually a sign that the animal is under pressure while he is learning to follow the rules of being a good dog.
4. Some people might believe that measuring sleepiness is a fairly _____ or easy task. For instance, couldn't you simply count the number of times a person yawns each hour or so?
5. Usually, people don't plan to yawn, but psychologists have done experiments that involve asking actors to yawn _____[-ly] in crowded rooms.
6. Certainly, there are physical reasons for yawning, but there is also a social _____ of yawning.
7. Playing for a moment and then turning to another more interesting or fun activity is usually enough to _____ yawning so that the feeling doesn't come back for a while.
8. We do not know whether yawning is a sign that you are getting ready for sleep or that you are successfully _____[-ing] sleep because you are trying not to go to sleep.
9. Unfortunately, yawns don't just _____ sleepiness. In some animals, yawning is a sign of stress.
10. One commonly used method didn't work. That is why many researchers have developed [a/an] _____ method to determine how sleepy a person is.

[Word Bank]

commitment facilitate unveil cradle sector correlation conscience resilient
delectable expenditure mend consumerism chain outfit trigger

11. This cycle of developed countries providing clothes for those in poorer countries is only sustainable if the clothing is durable and _____ .

12. Over the past ten years, falling prices have led to exceptional growth in _____[-s] on clothing as people are buying more and more.

13. They don't think of _____[-ing] a hole or sewing on a new button when a new shirt can be bought for six or seven dollars.

14. Fashion retailers are _____[-ing] new lines of clothing monthly or even weekly, and shoppers are excited to see the new fashions for the first time.

15. In Britain, shoppers spend over \$37 billion per year on clothes, and the fast fashion _____ comprises one-fifth of this market.

16. It might seem surprising that your own actions can make you start to yawn, but it is true: In fact, simply stretching your body will often _____ a yawn.

17. There's a chance that there will be new items of clothing if you shop at any of the retail _____[-s] of stores that use the "fast fashion" model of business.

18. Many of us ease our _____ about all this waste by donating our surplus clothing to charities so that we will feel less guilty.

19. The plan encourages everyone to consider the impact of a fashion item, from the _____ (its design) to the grave (its disposal)

20. 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 _____[-s] in clothing stores.

[Word Bank]

banish deliberate resist discipline correlation intervention agitated mend
inherit indicate chain trivial outfit obscure facilitate

21. Government documents are sometimes written in very _____ and confusing language that doesn't make sense to people who are not politicians or lawyer

22. There is a direct _____ between the best-known brands and the best-selling brands; the ones that most people know are also the ones that most people buy.

23. The politicians were fighting over the idea of a government _____ to regulate prices; some of them argued that the government should make rules that companies would be required to follow, while others argued that companies should be free to set prices.

24. The boy kept getting in trouble for yawning and sleeping at school until his teacher finally started to _____ him by lowering his grade every time he fell asleep.

25. People are more likely to donate their old clothing when church organizations and other charities _____ the process by sending a truck to people's homes and picking up the clothes directly.

26. The scientist spoke about his research very carefully and _____ [-ly] so that people in the audience would catch the importance of every word.

27. The group of environmentalists wanted to protest peacefully, so they linked their arms together and tried to form a human _____ around the clothing factory instead of fighting the police.

28. Surprisingly, the soldiers in a military _____ will sometimes yawn as a response to stress in dangerous situations.

29. If you yawn while your wife is talking to you about something important, you may need to buy her flowers to _____ your relationship.

30. When people don't want to say directly how boring something is, they sometimes _____ it indirectly by yawning instead.

Appendix L

Vocabulary posttests-1, -2, and -4

a. Post-1

1. We are going to show you a list of words. Some of the words appeared in the reading passages that you just completed (“Fashion Victim or Environmental Victory?” and “Sleep Research”), but some of the words did not appear in those reading passages. Please mark all of the words that you remember appearing in at least one of the reading passages. If a word did not appear in at least one of the reading passages, please do not mark it even if you know the word.

아래에 단어가 주어졌습니다. 어떤 단어는 여러분이 방금 끝낸 독해 지문 (“Fashion Victim or Environmental Victory?” 와 “Sleep Research”)에 나왔던 단어이고 어떤 단어는 나오지 않았던 단어입니다. 주어진 단어들 중 여러분이 이전에 독해시 보았던 단어들이 있으면 동그라미 치세요. 비록 여러분이 아는 단어라도 독해시 나왔던 단어가 아니면 표시하지 마세요.

[Word Bank]

commitment , facilitate , unveil , cradle , deliberate , resist , trivial , obedience , deed , discipline , sector , correlation , conscience , intervention , resilient , delectable , expenditure , agitated , fatigue , mend , consumerism , indicate , chain , banish , inherit , outfit , stimulate , alternate , obscure , aspect

b. Post-2

2. Now we are going to show you the same list of words. You might know the meanings of some of them, but you might not know the meanings of others. If you have any idea about the meaning of a word, please type a definition for it or describe its meaning as well as you can.

이번 역시 같은 단어가 아래에 주어졌습니다. 어떤 단어는 여러분이 그 의미를 아는 단어일 것이고, 어떤 단어는 모르는 단어일 것입니다. 여러분에 단어의 의미에 대해 조금이라도 알고 있으면 여러분이 할수 있는 한 그 단어의 의미를 표현해 보세요.

[Word Bank]

commitment , facilitate , unveil , cradle , deliberate , resist , trivial , obedience , deed , discipline , sector , correlation , conscience , intervention , resilient , delectable , expenditure , agitated , fatigue , mend , consumerism , indicate , chain , banish , inherit , outfit , stimulate , alternate , obscure , aspect

c. Post-4

4. We are going to show you the same lists of words another time. Below each list, you will see a set of sentences with blanks in them. Some of the sentences appeared in the reading passages, and some of them did not. For every sentence, please try to fill in the blank with a word from the list. (Sometimes you will see that the blank provides additional grammatical information—for example, to make a verb past tense so that it can fit.) Please try to make sure that the meaning of the word fits with the meaning of the sentence. Some of the words have more than one meaning, so they might fit in 2 different sentences. Some other words might not fit in any sentence, so you will not use them.

이번에도 같은 단어들이 다른 방식으로 보여집니다. 각 페이지마다 맨위에 **word bank** 가 주어지고, 각 **word bank** 아래 빈칸이 들어있는 문장이 주어집니다. 어떤 단어들은 여러분이 독해시 보았던 문장들이거나 또 매우 비슷하고, 어떤 단어들은 그렇지 않습니다. **Word Bank** 에 주어진 단어들을 이용하여 각 문장에 있는 빈칸을 채우십시오, 빈칸에 들어갈 단어의 의미가 전체 문장 의미에 맞아야 합니다. (또한, 가끔 여러분은 빈칸 옆에 추가적인 문법 변화를 볼수 있습니다- 예를들면, 동사의 과거 형태가 들어가야 하는 경우, 그 문법 변화가 자동적으로 적용될수 있도록 문법 변화가 이미 주어져 있습니다) 마지막으로, 주어진 단어들 중 두 번 사용되거나, 사용되지 않는 단어도 있으니 주의하십시오.

[Word Bank]

Intervention deliberate resist trivial obedience deed agitated fatigue indicate
commitment banish inherit alternate obscure aspect

1. Perhaps the dog trainer is pushing too hard or moving too fast and the dog feels nervous and _____.

→ (DEF) worried and unable to think clearly or calmly

2. Sometimes people simply lie to the researchers about how sleepy they are. This occurs because admitting _____ is considered a sign of weakness or a lack of ambition.

→ (DEF) a feeling of extreme physical or mental tiredness

3. When a trainer sees a dog yawning in a dog _____ class, it is usually a sign that the animal is under pressure while he is learning to follow the rules of being a good dog.

→ (DEF) the showing of respect for a person's authority by following an order, request, or law

4. Some people might believe that measuring sleepiness is a fairly _____ or easy task. For instance, couldn't you simply count the number of times a person yawns each hour or so?

→ (DEF) unimportant and not serious

5. Usually, people don't plan to yawn, but psychologists have done experiments that involve asking actors to yawn _____ [-ly] in crowded rooms.

→ (DEF) on purpose rather than by chance; planned or decided upon beforehand

6. Certainly, there are physical reasons for yawning, but there is also a social _____ of yawning.

→ (DEF) one of the parts of something's character or nature

7. Playing for a moment and then turning to another more interesting or fun activity is usually enough to _____ yawning so that the feeling doesn't come back for a while.

→ (DEF) to send something away and prevent it from returning

8. We do not know whether yawning is a sign that you are getting ready for sleep or that you are successfully _____ [-ing] sleep because you are trying not to go to sleep.

→ (DEF) to stop yourself from doing something even though you want to do it

9. Unfortunately, yawns don't just _____ sleepiness. In some animals, yawning is a sign of stress.

→ (DEF) to be a sign of something

10. One commonly used method didn't work. That is why many researchers have developed [a/an] _____ method to determine how sleepy a person is.

→ (DEF) different from something that is already being used

[Word Bank]

commitment facilitate unveil cradle sector correlation conscience resilient delectable
expenditure mend consumerism chain outfit trigger

11. This cycle of developed countries providing clothes for those in poorer countries is only sustainable if the clothing is durable and _____.

→ (DEF) strong and not easily damaged by being hit, stretched, or squeezed

12. Over the past ten years, falling prices have led to exceptional growth in _____[-s] on clothing as people are buying more and more.

→ (DEF) the spending of money on something, or the money that is spent on something

13. They don't think of _____[-ing] a hole or sewing on a new button when a new shirt can be bought for six or seven dollars.

→ (DEF) to repair something that is broken or not working, so that it works properly or can be used

14. Fashion retailers are _____[-ing] new lines of clothing monthly or even weekly, and shoppers are excited to see the new fashions for the first time.

→ (DEF) to introduce a plan, new product, or some other thing that has been kept secret to the public

15. In Britain, shoppers spend over \$37 billion per year on clothes, and the fast fashion _____ comprises one-fifth of this market.

→ (DEF) a smaller group which is a part of a larger group

16. It might seem surprising that your own actions can make you start to yawn, but it is true: In fact, simply stretching your body will often _____ a yawn.

→ (DEF) to cause an event to happen, or to cause a situation to exist

17. There's a chance that there will be new items of clothing if you shop at any of the retail _____[-s] of stores that use the "fast fashion" model of business.

→ (DEF) a group of several shops, hotels, or other businesses which are owned by the same person or company

18. Many of us ease our _____ about all this waste by donating our surplus clothing to charities so that we will feel less guilty.

→ (DEF) a feeling of guilt because you know you have done something that is wrong

19. The plan encourages everyone to consider the impact of a fashion item, from the _____ (its design) to the grave (its disposal)

→ (DEF) the place where something began

20. 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 _____[-s] in clothing stores.

→ (DEF) a set of clothes

[Word Bank]

banish deliberate resist discipline correlation intervention agitated mend inherit
indicate chain trivial outfit obscure facilitate

21. Government documents are sometimes written in very _____ and confusing language that doesn't make sense to people who are not politicians or lawyers.

→ (DEF) difficult to understand or deal with, usually because it involves so many parts or details

22. There is a direct _____ between the best-known brands and the best-selling brands; the ones that most people know are also the ones that most people buy.

→ (DEF) a connection or a link between things

23. The politicians were fighting over the idea of a government _____ to regulate prices; some of them argued that the government should make rules that companies would be required to follow, while others argued that companies should be free to set prices.

→ (DEF) the act of becoming involved in an argument, fight, or other difficult situation in order to change what happens

24. The boy kept getting in trouble for yawning and sleeping at school until his teacher finally started to _____ him by lowering his grade every time he fell asleep.

→ (DEF) to punish someone for something that they have done wrong

25. People are more likely to donate their old clothing when church organizations and other charities _____ the process by sending a truck to people's homes and picking up the clothes directly.

→ (DEF) to make an action or process easier or more likely to happen

26. The scientist spoke about his research very carefully and _____[-ly] so that people in the audience would catch the importance of every word.

→ (DEF) slowly and carefully

27. The group of environmentalists wanted to protest peacefully, so they linked their arms together and tried to form a human _____ around the clothing factory instead of fighting the police.

→ (DEF) a group of things arranged in a line

28. Surprisingly, the soldiers in a military _____ will sometimes yawn as a response to stress in dangerous situations.

→ (DEF) an organization

29. If you yawn while your wife is talking to you about something important, you may need to buy her flowers to _____ your relationship.

→ (DEF) to repair or resolve something, like a disagreement or quarrel between people

30. When people don't want to say directly how boring something is, they sometimes _____ it indirectly by yawning instead.

→ (DEF) to mention an opinion, an intention, or a fact in an indirect way

Appendix M

Informed consent documents

a. English version

IRB # 2010-056
Page 1 of 4

GEORGETOWN UNIVERSITY CONSENT TO PARTICIPATE IN RESEARCH INVOLVING TREATMENT

PROJECT TITLE

Automatic presentation of sense-specific lexical information in intelligent computer assisted language learning

PROJECT DIRECTOR

Dr. Graham Katz, Georgetown University

PRINCIPAL INVESTIGATOR

Soojeong Eom

TELEPHONE

301-646-9668

The Georgetown University Institutional Review Board (IRB) has approved this research project. For information on your rights as a research subject, call the Institutional Review Board office at 202-687-1506.

INTRODUCTION

You are invited to consider participating in a research study involving the development, implementation, and evaluation of intelligent computer-assisted language learning technology for students in intermediate-level ESL/EFL classes. This form will describe the purpose and nature of the research, its possible risks and benefits, other options available to you, and your rights as a participant in the study. Please take whatever time you need to discuss the study with your family and friends, or anyone else you wish to. The decision to participate, or not to participate, is yours. If you decide to participate, please be sure to sign and date the last page of this form.

WHY IS THIS RESEARCH STUDY BEING DONE?

In this research study, we are designing an intelligent computer tutoring system for intermediate learners of English as a second/foreign language. In the process of developing and evaluating this system, we will be investigating:

- (1) how effective sense-specific lexical information are facilitating learners' vocabulary acquisition and reading comprehension while reading

Developers of language learning technology often emphasize the importance of providing learners sense-specific semantic information on unknown words while reading; however, computer-based systems that are able to provide sense-specific semantic information on words are often restricted to a small fixed number of words, unable to cover any words learners look for. We hope to improve upon this.

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD
APPROVED NOV 10 2011
(DATE)
EXPIRATION OCT 13 2012
(DATE)

HOW MANY PEOPLE WILL TAKE PART IN THE STUDY?

About 45 people will take part in this study. Participants in the study are referred to as “subjects.” Subjects will be recruited from ESL/EFL classes at universities in the Washington, D.C., metropolitan area, including Georgetown University, George Mason University, the George Washington University, and the University of Maryland at College Park.

WHAT IS INVOLVED IN THE STUDY?

The study will be conducted activities as described below. Each subject can expect to engage in the following activities.

- Vocabulary test to provide the researchers/system developers with a picture of your English vocabulary ability (< 10 min).
- Online reading in English, provided by sense-specific lexical database tool. For example, while you are reading, if you are not sure or do not know the meaning of the word, you can click it and the system will provide you its definition and examples, which are appropriate in the context of your reading text (< 20 min).
- Post reading and vocabulary tests so that the researcher can analyze how much you improve in your ability of vocabulary and reading comprehension when you use the system for reading (< 20 min).
- A bio-data questionnaire so that the researchers can analyze results according to a variety of (summarized) learner characteristics, such as age, gender, major, previous experience with foreign language study, etc. (< 10 min).

When you are participating in the study, you will be asked to do all the above activities. Once you are done all the activities, you will be invited to return after 2 weeks for another post-test of your vocabulary acquisition.

HOW LONG WILL I BE IN THE STUDY?

We expect that your participation in the study will last approximately 1 hour. You can stop participating at any time. However, if you decide to stop participating in the study, we encourage you to talk with one of the researchers first.

WHAT ARE THE RISKS OF THE STUDY?

No known risks are involved in this research study. However, you should feel free to discuss any concerns you might have with one of the researchers, or with anyone else that you wish to.

ARE THERE BENEFITS TO TAKING PART IN THE STUDY?

It is reasonable to expect that participating in this study may assist you in your learning of English vocabulary and reading. However, participants will not experience any direct benefits as a result of participating in this study. Others may benefit in the future from the information we obtain in this study.

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD

APPROVED NOV 10 2011
(DATE)

EXPIRATION OCT 13 2012
(DATE)

WHAT ABOUT CONFIDENTIALITY?

Your name will not be used when data from this study are published. Every effort will be made to keep your research records and other personal information confidential. However, we cannot guarantee absolute confidentiality. Your name and any material that could identify you will remain confidential except as may be required by law.

Individuals from the Georgetown University IRB, other Georgetown University offices, Federal regulatory agencies, and the researchers involved in this study may look at records related to the study, both to assure quality control and to analyze data. The researchers conducting this study are a graduate student in computational linguistics at Georgetown University and one professor of computational linguistics at Georgetown University. Our names and contact information are given below:

Graham Katz	egk7@georgetown.edu	301-687-7939
Soojeong Eom	se48@georgetown.edu	301-646-9668

We will take the following steps to keep information about you confidential, and to protect it from unauthorized disclosure, tampering, or damage: When you first login to the system, you will be assigned a numerical code (e.g., Participant #42) which will be used to label your activity transcripts for the purposes of data analysis. Any personal information you provide will be associated with that code, and number-name correspondences will be available, if necessary, only to the researchers/entities listed above. All data will be stored in MySQL, a secure, password-protected database, and efforts will be made to ensure that identifying personal characteristics are not apparent in any of the examples used to illustrate trends in the data when the results of this study are published.

WHAT ARE MY RIGHTS AS A RESEARCH PARTICIPANT?

Participation in this study is entirely voluntary at all times. You have the right not to participate at all or to leave the study at any time. Deciding not to participate or choosing to leave the study will not result in any penalty or loss of benefits to which you are entitled, and it will not harm your relationship with Georgetown University or any of its employees.

If you decide to leave the study, please simply inform the researcher present of your intention to withdraw. You will not be obligated in any sense to complete any further part of the study and have the right to request that your data not be used. Throughout this study, a researcher will be available to discuss any information that may affect your interest in remaining in the study.

WHOM DO I CONTACT IF I HAVE QUESTIONS OR PROBLEMS?

Please contact Soojeong Eom (Phone: 301-646-9668, Email: se48@georgetown.edu), or the other researcher listed above on page 4, if you have any questions about the study.

Call the Georgetown University IRB Office at 202-687-1506 if you have any questions about your rights as a research participant.

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD
 APPROVED NOV 10 2011
 (DATE)
 EXPIRATION OCT 13 2012
 (DATE)

Statement of Person Obtaining Informed Consent

I have fully explained this study to the subject. I have discussed the study's purpose, its experimental and nonexperimental procedures and interventions, the possible risks and benefits, the standard and research aspects of the study, the alternatives to participation, and the voluntary nature of participation. I have invited the subject to ask questions and have answered any questions that the subject has asked.

Signature of Person Obtaining Informed Consent

Date

Consent of Subject

I have read the information provided in this Informed Consent Document. My questions were answered to my satisfaction. I voluntarily agree to participate in this study.

Printed name of Subject

Signature of Subject

Date

Upon signing, the subject will receive a copy of this form, and the original will become part of the subject's clinical record. If there is no relevant clinical record, the original will be held in the subject's research record.

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD

APPROVED NOV 10 2011
(DATE)

EXPIRATION OCT 13 2012
(DATE)

**GEORGETOWN UNIVERSITY
CONSENT TO PARTICIPATE IN RESEARCH STUDY
INFORMED CONSENT (KOREAN VERSION)**

영어를 구사하지 않는 피험자를 위한 서면 동의서

연구 참여 동의

연구제목

Automatic presentation of sense-specific lexical information in intelligent computer assisted language learning

지도교수

Graham Katz, 조지타운 대학교

연구자 이름

엄수정

연락처

1-301-646-9668 (미국)

82-10-3719-8765 (한국)

서문

귀하께서는 중급 영어 학습자의 컴퓨터를 이용한 영어 습득에 관련한 연구에 참여하도록 요청 받으셨습니다. 본 동의서를 충분한 시간을 가지고 읽으신 후, 연구 참여에 동의하여 주시기 바랍니다. 동의서를 읽으시는 중 궁금한 점이나 문의 사항이 있으시면 언제든지 연구자에게 물어보시면 됩니다. 본 연구의 참여 여부는 전적으로 본인의 의사에 따라 결정이 됩니다. 만일 연구에 참여하기를 희망하신다면, 동의서 마지막 부분에 서명을 하시고 서명한 날짜를 기입해 주십시오.

본 연구에 대한 질문이 있으시면 망설이지 마시고 언제든지 연구자에게 알려주십시오.

연구 배경과 목적

본 연구는 컴퓨터를 이용한 언어 학습이, 영어를 제 2 외국어로 학습하는 중급 학습자의 영어 지식 습득에 도움이 되는 지를 조사합니다. 컴퓨터 습 시스템을 개발하고 평가하는 과정을 통해, 본 연구는 (1) 컴퓨터로부터 문맥에 맞는 단어 의미 제공을 받는 것이 학습자의 어휘 습득과 독해 능력 향상에 얼마나 효과적인가를 알아보려고 합니다.

언어 학습을 도와주 첨단 기술 개발자들은 문맥에 맞는 단어 의미 제공의 중요성을 자주 강조합니다. 그러나 그러한 목적을 지닌 컴퓨터 시스템들은 현재 미리 지정된 단어에 관해서만 정보를 제공할 뿐 학습자가 원하는 모든 어휘에 대해서는 아직까지 지원을 하지 못하고 있습니다. 따라서 본 연구에서 제공하는 연구 실험 결과 분석은 차후 영어

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD
APPROVED NOV 10 2011
(DATE)
EXPIRATION OCT 13 2012
(DATE)

학습을 한층 더 효과적으로 도울수 있는 보다 나은 최첨단 컴퓨터 시스템을 개발하는데 공헌을 할 것입니다.

연구 계획

귀하께서는 영어를 제 2 외국어로 배우는 성인이기에 본 연구의 참여자로 선정되었습니다. 이 연구에서는 약 45 명의 성인 영어 학습자들이 피실험자로 참여하게 됩니다. 본 연구의 포함된 실험 절차는 다음과 같습니다.

- 동의서 작성, 실험 전 평가 (약 20 분 소요).
- 온라인 독해 참여 (약 15 분 소요).
- 독해 참여 후 즉시 어휘 평가와 독해 평가 (약 20 분 소요).
- 독해 참여 후 2 주일이 지난후 어휘 평가 (약 10 분 소요).

참여 시간

본 연구는 한 명의 피실험자 당 최대 약 1 시간의 참여로 이루어집니다. 귀하께서는 실험 중 언제든지 참여를 중단하실 수 있습니다. 그러나 만일 중단을 원하실 경우, 먼저 연구자에게 알려주시기를 권고합니다.

위험 노출 가능성

본 연구와 연계된 위험은 없습니다.

참여시 잠재적 혜택

실험에 참여하실 때 기대되는 직접적인 혜택은 없습니다. 그러나 실험에 참여하시는 피 실험자께서는 컴퓨터화 된 영어 과제를 통해 영어 연습을 하실 수 있습니다. 이와같은 영어 연습은 피실험자들의 영어 능력 향상에 도움을 줍니다.

기밀 유지

본 연구자는 피실험자들의 개인적 정보에 대한 기밀을 지키기 위해 모든 노력을 다할 것입니다.

귀하의 정보를 안전하게 지키기 위하여, 본 연구자는 귀하의 이름을 암호화 할 것입니다. 또한 귀하와 관련된 모든 정보는 본 연구자의 암호화 된 컴퓨터에 안전하게 저장될 것입니다. 귀하의 사인이 된 동의서 역시 연구자의 캐비닛에 안전하게 보관될 것입니다. IRB 에서 승인한 기간이 지나면, 귀하의 보관된 정보는 모두 일괄 소멸될 것입니다.

귀하의 이름 및 관련 정보는 학술 발표나 학계 논문 등에 전혀 발표되지 않을 것입니다. 하지만 귀하의 이름 등이 발표되지 않을지라도, 연구자는 계속해서 귀하와 연계된 관련

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD
APPROVED NOV 10 2011
(DATE)
EXPIRATION OCT 13 2012
(DATE)

정보를 열람할 수 있음을 알려드립니다. 나아가, 조지타운 대학교의 IRB 또한 필요시 귀하의 정보에 대한 열람을 요청할 수 있습니다.

연구와 관련된 피실험자의 권리

연구 실험에 참여하는 것은 전적으로 피실험자의 자발적인 참여에 의해 이루어집니다. 귀하께서는 본인 스스로 참여, 불참 및 참여 중 중도 포기에 대한 의사 결정을 할 수 있습니다. 만약 불참 혹은 참여 중 중도 포기를 선택하신다고 하여도, 귀하와 연구자 간의 관계에는 아무런 영향이 없습니다.

만약 귀하께서 더 이상 실험에 참여하기를 원하지 않을 경우, 귀하께서는 연구자에게 본인의 의사를 명백히 전달할 것을 부탁드립니다. 이미 수집된 귀하의 정보는 연구의 최종 결과에 포함되지 않을 것입니다.

질문 사항

만일 연구에 대한 질문이나 문의 사항이 있으실 경우, 연구자의 아래 연락처로 언제든지 연락하여 주십시오.

연구자: 염수정

연락처: 1-301-646-9668 (미국), 82-10-3719-8765 (한국)

이메일 주소: se48@georgetown.edu

피실험자의 연구참여 동의를 받는 사람의 성명서

나는 (1) 연구의 목적, 절차 및 기간; (2) 모든 실험적 시술; (3) 합리적으로 예상되는 연구의 모든 위험, 불편함 및 혜택; (4) 잠재적으로 혜택을 줄 수 있는 모든 대체 시술 및 치료; 그리고 (5) 기밀 유지 방법에 대하여 충분히 설명을 하였습니다. 연구자는 피실험자가 궁금해 할 수 있는 모든 사항에 대해 완벽한 대답을 제공하여 주었습니다.

Signature of Person Obtaining Informed Consent

Date

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD

NOV 10 2011

APPROVED _____

(DATE)

EXPIRATION _____

OCT 13 2012

(DATE)

피실험자의 동의

나는 이 서면 동의서에 기술된 모든 내용을 이해하였습니다.

나는 내 질문과 문의에 대해 연구자로부터 적절한 설명을 들었습니다.

나는 이 연구에 자발적으로 참여하는 것을 희망합니다.

Printed name of Subject

Signature of Subject

Date

이 피실험자의 동의서는 피실험자와 연구자에게 각각 한 부씩 보관용으로 주어집니다.

GEORGETOWN UNIVERSITY INSTITUTIONAL REVIEW BOARD

APPROVED NOV 10 2011

(DATE)

EXPIRATION OCT 13 2012

(DATE)

c. Anonymous survey

INFORMED CONSENT SCRIPT

ANONYMOUS SURVEY

You are invited to participate in a research study titled “Automatic presentation of sense-specific lexical information in intelligent computer assisted language learning”. This study is being conducted by Soojeong Eom, a Ph.D. candidate in the Department of Linguistics at Georgetown University, and Markus Dickinson, an assistant professor in the Department of Linguistics at Indiana University, for building gold standard data to evaluate computational performance of a system which has been built to match senses between dictionaries.

Participation in this study is entirely voluntary at all times. You can choose not to participate at all or to leave the study at any time. Regardless of your decision, there will be no effect on your relationship with the researchers or any other consequences.

You are being asked to take part in this study because you have some training in linguistics.

If you agree to participate, you will be asked to fill out 3 surveys about choosing the same meaning as the given word/meaning. This survey should take around 15 minutes to complete (; each survey is to take 5 minutes). The survey will be collected during May through August 2011 and via online webpage; you are to follow the given links and answer the surveys.

All of your responses to this survey will remain anonymous and cannot be linked to you in any way. No identifying information about you will be collected at any point during the study, and your survey will be identified only with an arbitrarily assigned number. Once you submit your completed survey, there will be no way to withdraw your responses from the study because the survey contains no identifying information.

Study data will be kept in digital formats in the online webpage. Access to digital data will be protected automatically set by Soojeong Eom when it is first made. Only Soojeong Eom will have access to the data.

There are no risks associated with this study. While you will not experience any direct benefits from participation, information collected in this study may benefit others in the future by helping to build a comprehensive dictionary which has more lexical information than any other dictionaries. This kind of comprehensive dictionary would provide much more help to users.

If you have any questions regarding the survey or this research project in general, please contact the principal investigator, Soojeong Eom, at (301) 646-9668 or via email at se48@georgetown.edu or her co-mentor, Markus Dickinson, at (812) 856-2535 or via email at md7@indiana.edu. If you have any questions about your rights as a research participant, please contact the Georgetown University IRB at (202) 687-6553 or irboard@georgetown.edu.

By completing and submitting this survey, you are indicating your consent to participate in this study.

Soojeong Eom
Ph.D. candidate
Department of Linguistics
301-646-9668
se48@georgetown.edu

REFERENCES

- Amaral, L. & Meurers, D. (2006). Using foreign language tutoring systems for grammatical feedback. *EUROCALL 2006*. Granada, Spain.
- Atkins, S. (1993). Tools for computer-aided corpus lexicography: the hector project. *Acta Linguistica Hungarica*, 41, 5-72.
- Bailey, S. & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA3)* (pp. 107-115). Columbus, OH: Association for Computational Linguistics.
- Beheydt, L. (1987). The semantization of vocabulary in foreign language learning. *System*, 15, 55-67.
- Black, A. (1991). On-line consultation of definitions and examples: Implications for the design of interactive dictionaries. *Applied Cognitive Psychology*, 5, 149-166.
- Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7, 15-32.
- Bensoussan, M., Sim, D., & Weiss, R. (1984). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Readings in a Foreign Language*, 2, 262-276.
- Bogaards, P. (1998). Using dictionaries: Which words are looked up by foreign language learners? In B.T.S. Atkins & K. Varantola (Eds.), *Studies of dictionary use by language learners and translators* (pp. 151-157). Tubinger, Germany: Niemeyer.
- Chanier, T., & Selva, T. (1998). The ALEXIA system: the use of visual representations to enhance vocabulary learning. *Computer Assisted Language Learning*, 11(5), 489-521.
- Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 225-237). Cambridge: University Press.
- Coady, J., Magoto, J., Hubbard, P., Graney, J., & Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. In T. Huckin, M. Haynes, & J. Coady (Eds.) *Second Language Reading and Vocabulary Learning* (pp. 67-85), Norwood, NJ: Ablex.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

- Coombe, C. (2011). Assessing vocabulary in the language classroom. In D. Anderson & R. Sheehan (Eds.), *Focus on Vocabulary: Emerging Theory and Practice for Adult Arab Learners* (pp. 111-124), United Arab Emirates: HCT Press.
- Coombe, C., & Hubley, N. (2003). *Assessment practices*. Alexandria, VA: TESOL Publications.
- Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2010). Timbl: Tilburg memory based learner, version 6.3, reference guide. Technical report, ILK Research Group. Technical Report Series no. 10-01.
- Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (1998). TiMBL: Tilburg Memory Based Learner, version 1.0, Reference Guide. *ILK Technical Report*.
- Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2001). TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. *ILK Technical Report*.
- Delmonte, R. (2002). Feedback generation and linguistic knowledge in 'SLIM' automatic tutor. *ReCALL*, 14(2), 209-234.
- DeSmedt, W. (1995). Herr Kommissar: An ICALL conversation simulator for intermediate German. In V. M. Holland, J. Kaplan, & M. Sams (Eds.), *Intelligent language tutors: Theory shaping technology* (pp. 153-174). Hillsdale, NJ: Lawrence Erlbaum.
- Dela Rosa, K., & Eskenazi, M. (2011). Impact of word sense disambiguation on ordering dictionary definitions in vocabulary learning tutors. *Proceedings of the 24th International FLAIRS Conference*. Palm Beach, FL: AAAI Press.
- Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, 8(2), 103-128.
- Eom, S., Dickinson, M., & Katz, G. (2012). Using semi-experts to derive judgments on word sense alignment: a pilot study. *Proceedings of the Eight International Conference of Language Resources and Evaluation (LREC-12)*. Istanbul, Turkey: European Language Resources Association.
- Erk, K., & McCarthy, D. (2009). Graded word sense assignment. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 440-449), Singapore: Association for Computational Linguistics.
- Erk, K., & McCarthy, D., & Gaylord, N. (2009). Investigations on word senses and word usages. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*

- of the AFNLP* (pp. 10-18), Suntec, Singapore: Association for Computational Linguistics.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
- Fernando, S., & Stevenson, M. (2010). Aligning wordnet synsets and wikipedia articles. *Proceedings of the AAAI Workshop on Collaboratively-Built Knowledge Sources and Artificial Intelligence* (pp. 48-50), Athens, GA: AAAI Press.
- Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21, 225-240.
- Gamper, J., & Knapp, J. (2001). Adaptation in a language learning system. *Proceedings of ABIS-Workshop held in conjunction with LLWA '01*, Dortmund, Germany.
- Gamper, J., & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning*, 15(4), 329-342.
- Goodfellow, R. (1995). A review of the types of CALL programs for vocabulary instruction. *Computer Assisted Language Learning*, 8, 205-226.
- Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375-406.
- Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4), 39-52.
- Grefenstette, G., & Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenisation. *Proceedings of 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)*. Budapest, Hungary.
- Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning and Technology*, 4(1), 60-81.
- Harley, B. (1996). Introduction: Vocabulary learning and teaching in a second language. *The Canadian Modern Language Review*, 53(1), 3-11.
- Haynes, M. (1993). Patterns and perils of guessing in second language reading. In T. Huckin, M. Haynes, M., & J. Coady (Ed.), *Second language reading and vocabulary learning* (pp. 46-64). Norwood, N.J.: Ablex.
- Haynes, M., & Baker, I. (1993). American and Chinese readers learning from lexical familiarization in English texts. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second language reading and vocabulary acquisition* (pp.130-152). Norwood, NJ: Ablex

- Heift, T. & Nicholson, D. (2001). Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, 12(4), 310-325.
- Heift, T. (2001). Error-specific and individualized feedback in a web-based language tutoring system: Do they read it? *ReCALL*, 13(2): 129-142.
- Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. *Proceedings of the 9th International Conference on Spoken Language Processing*. Pittsburgh, PA.
- Heilman, M. & Eskenazi, M. (2006, December). Authentic, individualized practice for English as a second language vocabulary. Paper presented at *Interfaces of Intelligent Computer-Assisted Language Learning Workshop*, Ohio State University, Columbus, OH.
- Heilman, M., & Eskenazi, M. (2008). Self assessment in vocabulary tutoring. *Proceedings of the Young Researcher's Track, Ninth International Conference on Intelligent Tutoring Systems*. Montreal, Canada.
- Hirsh, D. (2012). *Current Perspectives in Second Language Vocabulary Research*. New York: Peter Lang AG.
- Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 21(2), 181-193.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Ide, N., & Veronis, J. (1990). Mapping dictionaries: A spreading activation approach. *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary* (pp. 52-64), Waterloo, Canada.
- Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition. *Applied Linguistics*, 19(3), 357–377.
- Jurafsky, D., & Martin, J. E. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing*. Upper Saddle River, NJ: Prentice Hall.
- Kilgarriff, A., Hus'ak, M., McAdam, K., Rundell, M., & Rychl'y, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of EURALEX-08*. Barcelona, Spain.
- Knight, K., & Luk, S. K. (1994). Building a large-scale knowledge base for machine

- translation. *Proceedings of AAAI-94*. Seattle, WA: AAAI Press.
- Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, 78(3), 285-299.
- Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, 58, 285–325.
- Koyama, T., & Takeuchi, O. (2004). How look-up frequency affects EFL learning: An empirical study on the use of handheld-electronic dictionaries. *Proceedings of CLaSIC 2004*, 1018-1024.
- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal*, 73, 440-464.
- Kwong, O. Y. (1998). Aligning WordNet with additional lexical resources. *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems* (pp. 73-79). Montreal, Canada.
- Kulkarni, A., Heilman, M., Eskenazi, M., & Callan, J. (2008). Word sense disambiguation for vocabulary learning. *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems*. Montreal, Canada: Springer.
- Landes, S., Leacock, C., & Teng, R. I. (1998). Building semantic concordances. In C. Fellbaum (Ed.), *WordNet: A Lexical Reference System and its Application*. Cambridge, MA: MIT Press.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.
- Laufer, B. (1997). The lexical plight in second language reading : Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 20-34). Cambridge: University Press.
- Laufer, B. (2001). Reading, word-focused activities and incidental vocabulary acquisition in a second language. *Prospect*, 16(3), 44-54.
- Laufer, B. (2003). Vocabulary acquisition in a second language: do learners really acquire most vocabulary by reading? *Canadian Modern Language Review*, 59(4), 565-585.
- Laufer, B. (2005). Instructed second language vocabulary learning: the fault in the 'default hypothesis'. In A. Housen & M. Pierrard (Eds.), *Investigations in Instructed Second Language Acquisition* (pp. 286-303), Mouton de Gruyter.

- Laufer, B., & Hill, M. (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.
- Laufer, B. and J. Hulstijn (2001). Incidental vocabulary acquisition in a second language: the construct of task-induced involvement. *Applied Linguistics* 22, 1-26.
- Laws, F., Scheible, C., & Schutze, H. (2011). Active learning with amazon mechanical turk. *Proceedings of EMNLP-11*, Edinburgh, UK: Association for Computational Linguistics.
- Leacock, C., Towell, G., & Voorhees, E. (1993). Corpus-based statistical sense resolution. *Proceedings of the ARPA Workshop on Human Language Technology*, Edinburgh, UK: Association for Computational Linguistics.
- Leffa, V. J. (1992). Making foreign language texts comprehensible for beginners: An experiment with an electronic glossary. *System*, 20(1), 63-73.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26). New York, NY: ACM.
- L'haire, S., & FaLtin, A.V.V. (2003). Error diagnosis in the FreeText project. *CALICO Journal*, 20(3), 481-495.
- Li, X. (1988). Effects of contextual clues on inferring and remembering meanings of new words. *Applied Linguistics*, 9(4), 402-413.
- Liu, H. (2004). MontyLingua: An end-to-end natural language processor with common sense, <http://web.media.mit.edu/~hugo/montylingua>.
- Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* (pp. 62-69), Somerset, NJ: Association for Computational Linguistics.
- Lyman-Hager, M., Davis, J. N., Burnett, J., & Chennault, R. (1993). Une Vie de Boy: Interactive reading in French. *Proceedings of the CALICO 1993 Annual Symposium on "Assessment"* (pp. 93-97), Durham, NC: Duke University.
- Lupescu, S., & Day, R. R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning*, 43(2), 263-287.
- Ma, Q. (2009). *Second Language Vocabulary Acquisition*. Germany: Peter Lang.

- Madhani, N., Chodorow, M., Tetreault, J., & Rozovskaya, A. (2011). They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 508–513), Portland, OR: Association for Computational Linguistics.
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Dickinson, M., Eom, S., Kang, Y., Lee, C. M., & Sachs, R. (2008). A balancing act: how can intelligent computer-generated feedback be provided in learner-to-learner interactions? *Computer Assisted Language Learning*, 21(5), 369-382.
- Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language*, 22(1), 56-70
- Meyer, C. M., & Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 883–892), Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Mihalcea, R., & Csomai, A. (2005). Sense-Learner: Word sense disambiguation for all words in unrestricted text. *Proceedings of the ACL Interactive Poster and Demonstration Sessions* (pp. 53-56), Ann Arbor, MI: Association for Computational Linguistics.
- Mihalcea, R., & Faruque, E. (2004). Senselearner: Minimally supervised word sense disambiguation for all words in open text. *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 155–158), Barcelona, Spain: Association for Computational Linguistics.
- Miller, G. A., Chodorow, M., Landes, S., Leacock, C., & Thomas, R. G. (1994). Using a semantic concordance for sense identification. *Proceedings of ARPA Human Language Technology Workshop* (pp. 240-243), Plainsboro, NJ: Morgan Kaufmann.
- Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. New York: Multilingual Matters.
- Mooney, R. J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 92-91), Philadelphia, PA.
- Muhonen, K., & Purtonen, T. (2011). Creating a dependency syntactic treebank: Towards

- intuitive language modeling. *Proceedings of International Conference on Dependency Linguistics (DepLing'11)* (pp. 155-164), Barcelona, Spain.
- Nagata, N. (2002). BANZAI, An application of natural language processing to web-based language learning. *CALICO Journal*, 19(3), 583-599.
- Nagy, W. (1997). On the role of context in first-and second-language vocabulary learning. In N. Schmitt & M. McCarty (Eds). *Vocabulary: Description, Acquisition and Pedagogy* (pp. 64-83). Cambridge, MA: Cambridge University Press.
- Nagy, W. E., Herman, P.A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233-253.
- Nastase, V., & Szpakowicz, S. (2001). Word-sense disambiguation in Roget's Thesaurus using WordNet. *Proceedings of the Workshop on WordNet and other lexical resources*, Pittsburgh, PA.
- Nation, I. S. P., & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 97–110), London, UK: Longman
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York, NY: Cambridge University Press.
- Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32 (4), 678-692
- Navigli, R., & Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. *Proceedings of ACL-10* (pp. 216–225), Uppsala, Sweden: Association for Computational Linguistics.
- Navigli, R., Litkowski, K. C., & Hargraves, O. (2007). Semeval-2007 task 07: Coarse-grained English all-words task. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 30-35), Prague, Czech Republic: Association for Computational Linguistics.
- Navigli, R. (2006). Reducing the granularity of a computational lexicon via an automatic mapping to a coarse grained sense inventory. *Proceedings of LREC 2006*, Genova, Italy.
- Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2), 1-69.
- Nerbonne, J., & Smit, P. (1996). GLOSSER-RuG: in support of reading. *Proceedings of*

- COLING-96 (pp. 830-835), Copenhagen, Denmark.
- Ng, H. T., & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics*, Santa Cruz, CA: Association for Computational Linguistics.
- Ng, H. T., & Lee, H. B. (1997). DSO corpus of sense-tagged English. Technical report, *Linguistic Data Consortium*, Philadelphia, PA.
- Niemann, E., & Gurevych, I. (2011). The peoples web meets linguistic knowledge: Automatic sense alignment of wikipedia and wordnet. *Proceedings of the 9th International Conference on Computational Semantics* (pp. 205-214), Oxford, UK.
- Palmer, M., Dang, H. T., & Rosenzweig, J. (2000). Sense tagging the Penn Treebank. *Proceedings of the Second Language Resources and Evaluation Conference (LREC-00)*, Athens, Greece.
- Paribakht, T. S., & Wesch, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 174-200). Cambridge, MA: Cambridge University Press.
- Paribakht, T. S., & Wesch, M. (1999). Reading and “incidental” L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21, 195-224.
- Pedersen, T., & Bruce, R. (1997). A new supervised learning algorithm for word sense disambiguation. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, Providence, RI.
- Pedersen, T., & Kolhatkar, V. (2009). WordNet::SenseRelate::AllWords - a broad coverage word sense tagger that maximizes semantic relatedness. *Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technology 2009 Conference*, Boulder, CO: Association for Computational Linguistics.
- Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity - measuring the relatedness of concepts. *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, Boston, MA: Association for Computational Linguistics.
- Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, 11(2), 36-58.

- Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language learning*, 59(1), 113-151.
- Prichard, C. (2008). Evaluating L2 readers' vocabulary strategies and dictionary use. *Reading in a Foreign Language*, 20(2), 216-231.
- Qian, D. D. (1999). Assessing the roles of depth and breath of vocabulary knowledge. *Canadian Modern Language Review*, 56, 282-307.
- Rapaport, W. J., & Kibby, M. W. (2002). Contextual vocabulary acquisition: A computational theory and educational curriculum. *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)*, 261-266.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10, 355-371.
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60), Mahwah, NJ: Erlbaum.
- Read, J. (2000). *Assessing vocabulary*. Cambridge, MA: Cambridge University Press.
- Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, 7(2), 105-125.
- Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 20, 509-42.
- Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. *Proceedings of 3rd Atlantic Web Intelligence Conference (AWIC-05)*, 3528, 380-386.
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.
- Schmitt, D., Schmitt, N., & Mann, D. (2011). *Focus on Vocabulary 1: Bridging Vocabulary*. New York: Pearson ESL.
- Schouten-van Parreren, C. (1989). Vocabulary learning through reading: which conditions should be met when presenting words in texts? *AILA Review*, 6, 75-85.
- Segler, T. (2007). *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Ph.D. dissertation, University of Edinburgh, City of Edinburgh, United Kingdom. <http://hdl.handle.net/1842/1750>.

- Segler, T., Pain, H., & Sorace, A. (2002). Second language vocabulary acquisition and learning strategies in ICALL environments. *Computer Assisted Language Learning*, 15(4), 409–422.
- Shei, C. (2001). FollowYou! an automatic language lesson generation system. *Computer Assisted Language Learning*, 14(2), 129-144.
- Sinclair, J. (2006). *Collins COBUILD Advanced Lerner's English Dictionary*. New York: Harper Collins.
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP-08* (pp. 254-263), Honolulu, Hawaii: Association for Computational Linguistics.
- Sternberg, R.J. (1987). Most vocabulary is learned from context. In M.G. McKeown & M.E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89-105), Hillsdale, NJ: Erlbaum.
- Summers, D. (1998). The role of dictionaries in language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 111-125). London: Longman.
- Tang, G. M. (1997). Pocket electronic dictionaries for second language learning: Help or hindrance?, *TESL Canada Journal*, 15, 39-57.
- The College Board. (2009). *The Official SAT Study Guide*. New York: College Board.
- Toral, A., Ferrández, O., Agirre, E., & Muñoz, R. (2009). A study on linking and disambiguating wikipedia categories to wordnet using text similarity. *Proceedings of the International Conference RANLP 09* (pp. 449-454), Borovets, Bulgaria: Association for Computational Linguistics.
- Toutanova, K., & Manning, C. D. (2000). Enriched the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VCL-2000)*, Hong Kong.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL 2003* (pp. 252–259), Edmonton, Canada.
- Tozcu, A., & Coady, J. (2004). Successful learning of frequent vocabulary through CALL also benefits reading comprehension and speed. *Computer Assisted Language Learning*, 17(5), 473-495.

- Wang, A., Hoang, C. D. V., & Kan, M. (2009). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 43(1), 1-23.
- Wesch, M., & Paribakht, T. S. (1996). Assessing L2 vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, 53(1), 13-40.
- Wode, H. (1999). Incidental vocabulary acquisition in the foreign language classroom. *Studies in Second Language Acquisition*, 21, 243-258.
- Wolf, E., & Gurevych, I. (2010). Aligning sense inventories in wikipedia and wordnet. *Proceedings of the First Workshop on Automated Knowledge Base Construction* (pp. 24-28), Grenoble, France.