Testing the Story Workbench:
A Faster and more Accurate Tool for Word Sense Disambiguation

Diana Moore
MIT CSAIL
May 2008

1. Introduction

The ability to quickly and accurately collect semantic data is necessary for many tasks in artificial intelligence. For example, any task that seeks to learn something by reading a text needs a method to collect data about the meaning of the words in the text.

The task of associating a given word in a text with a meaning (sense) that is different from other meanings associated with the word is known as word sense disambiguation. Some of these distinctions between meanings are coarse while others are subtle. For example, the word *bank*, in its noun form, has two coarse meanings, one as a sloping piece of land and the other as a holder of money. These coarse meanings are usually not too difficult to distinguish when assigning word senses. However, under the meaning of a holder of money, there are more subtle divisions. For instance, the word *bank* could refer to the financial institution of a bank, the physical bank building, or a personal money container such as a piggy bank. The need to distinguish between these subtle meanings is what makes the task of word sense disambiguation difficult.

As a step on the path towards quickly and accurately collecting semantic data, Mark Finlayson of the Genesis Group at MIT CSAIL has created a new tool for word sense disambiguation, the Story Workbench. This tool provides the ability to collect structured semantic representations of what people know quickly and accurately thereby enabling previously unrealizable psychology experiments that will provide insight into how humans understand language. I tested the workbench to determine if it is a viable and practical approach to word sense disambiguation and collecting this semantic data.

The outline of the paper is as follows. Section 2 discusses previous approaches to the problem of word sense disambiguation and finishes with a description the Story Workbench. Section 3 describes the Story Workbench in its current state. Section 4 details the experimental procedure and section 5 presents the results. Section 6 discusses future studies and improvements for the workbench and section 7 states my contributions to the problem of word sense disambiguation as well as the broader goals of quickly and accurately collecting semantic data.

2. Previous Approaches

Past approaches to word sense disambiguation have two basic forms: manual and automatic. In the manual approach, human annotators mark words in a text with the appropriate meaning, which provides the computer with the meaning of each word in the text. This approach is accurate at choosing the correct meaning of a word in context because humans are good at understanding language. However, the task of manually annotating each word is impractical because it is slow and tedious. A faster approach is automatic annotation. In this approach, the computer applies an algorithm to each word in the text to find the appropriate definition of the word in context. Although this method is much faster than the manual method, it is far too inaccurate. Automatic approaches on average choose the correct meaning of the word 60% of the time. Some automatic

approaches have performed at 80% to 90% accuracy levels but this performance was only on a small subset of data (Palmer 2006).

The Story Workbench is a tool for word sense disambiguation that seeks to take the best of both approaches by using a semi-automatic approach. The semi-automatic approach uses an algorithm to make a best guess at the correct meaning of a word in context at which point a human annotator either agrees with this guess or corrects it. The hope is that this tool will allow words to be semantically annotated with the accuracy of a human annotator while significantly reducing time required to perform the annotations. Using the accuracy measurements from above the human annotator would only have to perform a difficult annotation 40% of the time.

3. Story Workbench

Given a text the Story Workbench identifies the words whose meaning is ambiguous. A word is ambiguous if more than one definition exists for the word in some electronic dictionary. The Story Workbench currently uses WordNet 2.1 (Fellbaum 1998). Next, using a word sense disambiguation algorithm, the Story Workbench makes a best guess at the appropriate definition for each word in the text. The Story Workbench currently uses the Lesk Algorithm (Lesk 1986).

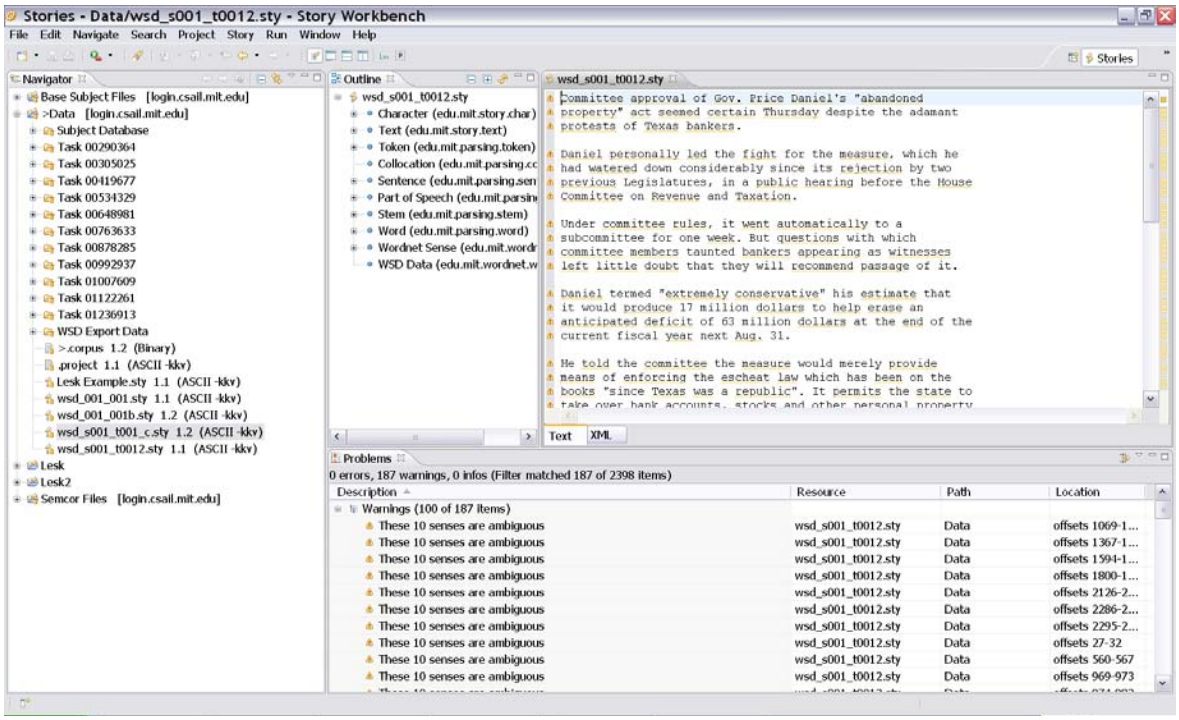


Figure 3.1 *the Story Workbench*

A human annotator then runs the Word Sense Disambiguation Wizard, shown in Figure 3.2, to check the work of the computer. The wizard presents the human annotator with a word to annotate surrounded by three sentences of context for the word. The wizard also supplies a list of possible definitions for the word with the computer's guess highlighted in green. When the user clicks on this definition the computer will display the contextual evidence that it used to make its guess. If

the guess seems correct the human annotator continues to the next word but if the guess seems incorrect the human annotator looks through the other definitions and chooses the correct one.

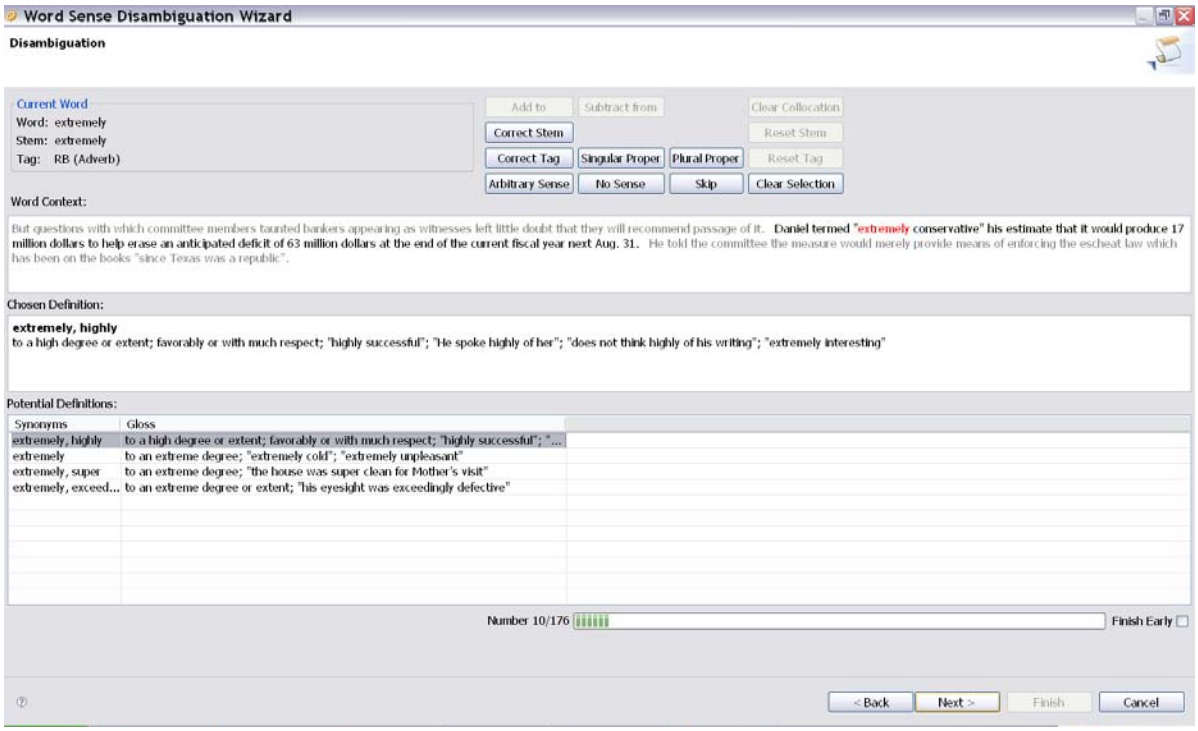


Figure 3.2 *Word Sense Disambiguation Wizard*

4. Experiments

In order to test the story workbench to see how well it performs, I ran experiments that aimed to show that the workbench is faster and more accurate than traditional approaches. In the experiment, subjects used the story workbench to annotate 217 words from a text using three different methods of annotation. Appendix A contains the text that subjects annotated. This text was split into subsections. Subjects performed one method on the first 72 words of the text, a second method on the next 73 words, and a third method on the last 72 words of the test. Table 4.1 indicates the subsections of the text. All subjects annotated the text in order, meaning that no matter which method a subject used the subject annotated the first subsection first, then the second and finished with the third subsection.

Section	Number of Words	Start Word	End Word	Lines
1	72	committee	means	1-12
2	73	enforcing	bank	12-22
3	72	customer	Monday	22-32

Table 4.1 *Text Subsections*

4.1 Random Method

As a baseline, one method, known as the *Random Method*, presents the possible definitions in a randomized order. This method is considered a baseline to compare other methods to because it accounts for any effect on speed or accuracy from using the workbench interface so it allows differentiation between the effect of the interface and the effect of the particular method invoke by the workbench.

4.2 Most Common First Method

A second method, know as the *Most Common First Method*, presents the possible definitions in a most common first order. Presumably this ordering should cut down on the amount of time need to perform the annotations because the annotator only has to read down the list of definition until finding the correct one and since the first one is the correct one most often the annotator only has to read one definition most of the time.

4.3 Lesk Method

The third method, known as the *Lesk Method*, uses an implementation of the Lesk algorithm that provides a guess by the computer to the human annotator (Lesk 1986). The computer's guess is highlighted and if the user clicks on the definition the workbench highlights the contextual evidence it used to make the guess. This method is designed to help the annotator by providing a likely answer so that the annotator can select that option if it seems correct quickly or search for another option in the less likely event that the option is wrong.

In the pilot studies, a fourth method was also tested. This method implemented a one sense per discourse heuristic that highlights the definition the annotator chose for the same word earlier. From the results of the pilot studies, the effect of this method seemed small so it was not used in the actual experiments.

In an attempt to control the effect of the order in which the methods were used on the overall speed and accuracy associated with the methods, I randomly assigned each subject into one of six groups. Each of the six groups represents a possible order of the methods. The group orderings are as presented in Table 4.2.

	First Method	Second Method	Third Method
Group 1	Random	Most Common First	Lesk
Group 2	Random	Lesk	Most Common First
Group 3	Most Common First	Random	Lesk
Group 4	Most Common First	Lesk	Random
Group 5	Lesk	Most Common First	Random
Group 6	Lesk	Random	Most Common First

Table 4.2 *Method Orderings by Group*

4.4 Experimental Setup

When the subject entered the room, the experimenter asked her to sit in the chair in front of the computer. The subject read and signed the consent form. After signing the consent form, the subject read the task instructions. The experimenter told the subject that after she read the instructions the experimenter would guide her through a training task. Once the subject finished reading the instructions, the experimenter gave the subject the training text and reading comprehension questions for the text. After reading the text and answering the questions, the subject used the workbench to perform annotations while the experimenter watched and provided suggestions. The training task, given in Appendix B, provided the subject with thirteen words to annotate. The training task contained two collocation examples and one instance where the subject needed to change the stem of a word. The experimenter made sure that the subject did not miss the collocations, assisted with using Choose Arbitrary Sense for the verb reelected, reminded the subject they could use the arrow keys to select the definition and answered any questions the subject had while performing the training task. Once the subject completed the task, the experimenter asked if she had any questions. The experimenter then gave the subject the text she would be annotating and reading comprehension questions for the task. After reading the text and answering the reading comprehension questions, the subject performed the annotation task. The experimenter told subjects that they would be using three different methods to annotate the text so they would be stopped twice during the annotation task so that the experimenter could change the method. When it was time for the subject to use the *Lesk Method* the experimenter provided the subject with special instructions for using that method. Once the task was complete, the subject completed a survey about their cultural background and their experience using the workbench.

5. Results

The results of the study suggest two main findings. One, the ordering of the methods had a significant effect on the speed and accuracy of the annotations. Subjects consistently performed the third annotation task faster and more accurately than the first two. This finding suggests that subjects perform better with more training. Second, both the *Most Common First Method* and the *Lesk Method* lead to more accurate annotations than the *Random Method*. However, the study did not reveal a significant difference between the accuracy produced by the *Most Common First Method* and the *Lesk Method*. The study also failed to find any significant difference in time to annotate between each of the methods. Most likely, the ordering effect overwhelmed any effect by the annotation methods.

5.1 Data Analysis

In analyzing the data, I looked at the time each subject took to perform the annotation and how accurate each subject's annotations were when compared to the annotations made on the same text by professional annotators, hereafter referred to as the gold standard. The text was chosen from the Semcor corpus because the text found in this corpus has been annotated using Word Net by professional annotators.

Two complications arose in the analysis of the dataset as a whole. First, some subjects had missing data for some of the annotations. For these subjects, some words appeared unannotated and did not have values for the definition chosen and the time taken to annotate the word. Three subjects had missing data for more than fifteen words in the text. In the final analysis, I excluded the data from

these three subjects. Twelve other subjects had missing data for one to fifteen words. For these subjects, I marked a word that was not annotated as if the subject had chosen the No Sense option and thus each of the unannotated words counted against the accuracy of the subject. As a result, the accuracy numbers presented in this analysis may be smaller than they actually should be. However, the subjects with missing data were distributed evenly among the task groups so any comparisons between groups should produce accurate results.

The second complication arose from difference in the words annotated for the gold standard and those given to the subjects to annotate. The gold standard contained four annotated words that the workbench did not consider ambiguous and thus the subjects did not annotate these words. These words were *two*, *seventeen*, *million*, and *seven* and are underlined in the text found in Appendix A. I removed these words from the dataset and they did not count against any accuracy measurements. Similarly, the workbench asked subjects to annotate five words that the gold standard did not contain. These words were *books*, *not*, *such*, *something*, and *several* and are in boldface in the text found in Appendix A. These words were left in the dataset and as a result counted against the accuracy measures.

The order in which subjects used each method turned out to have a large effect on the speed and accuracy with which the subjects performed the annotations. The subjects performed the third task, no matter the method, faster and more accurately than the first two tasks. The results also indicate a significant time difference between the first and second task, although no significant accuracy difference exists between these two tasks.

5.2 Speed Measurements

Quickly collecting semantic date is one goal of the workbench; thus, I analyzed the speed of annotation. The workbench times the annotation of each word during the task. I used this data to calculate the total time each subject spent on each task. I compared the time each task took based on the order in which the subjects performed the tasks and the method used. The results are presented in Table 5.1.

Task/Method	Time (minutes)
First Task	19.33
Second Task	16.59
Third Task	10.72
Random Method	15.73
Most Common First Method	15.87
Lesk Method	15.17

Table 5.1 *Speed Results*

Using a t-test I found that the differences in time were significant based on ordering, with p values of 0.01. In other words, subjects performed the second task faster than the first and the third task faster than the second no matter what method they used. Figure 5.1 shows the time for each task and Figure 5.2 shows the time for each task divided by group. I did not find a significant time difference based on the method used.

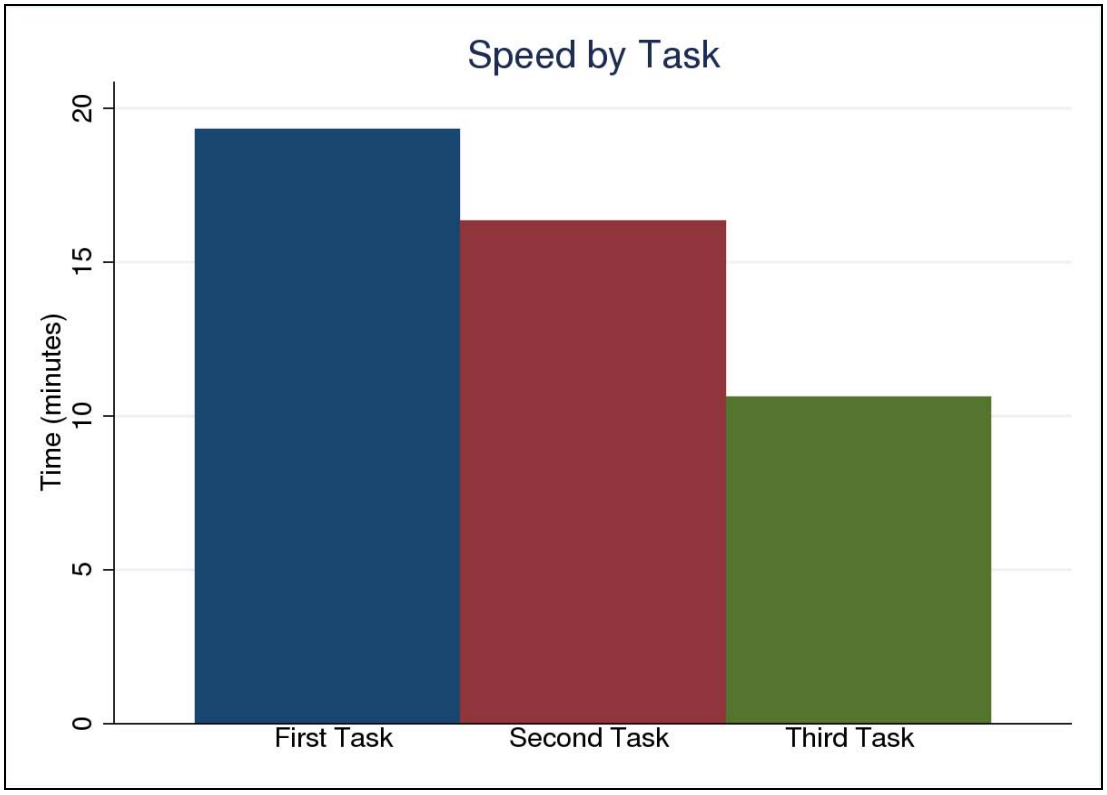


Figure 5.1 *Time, in minutes, for each task.*

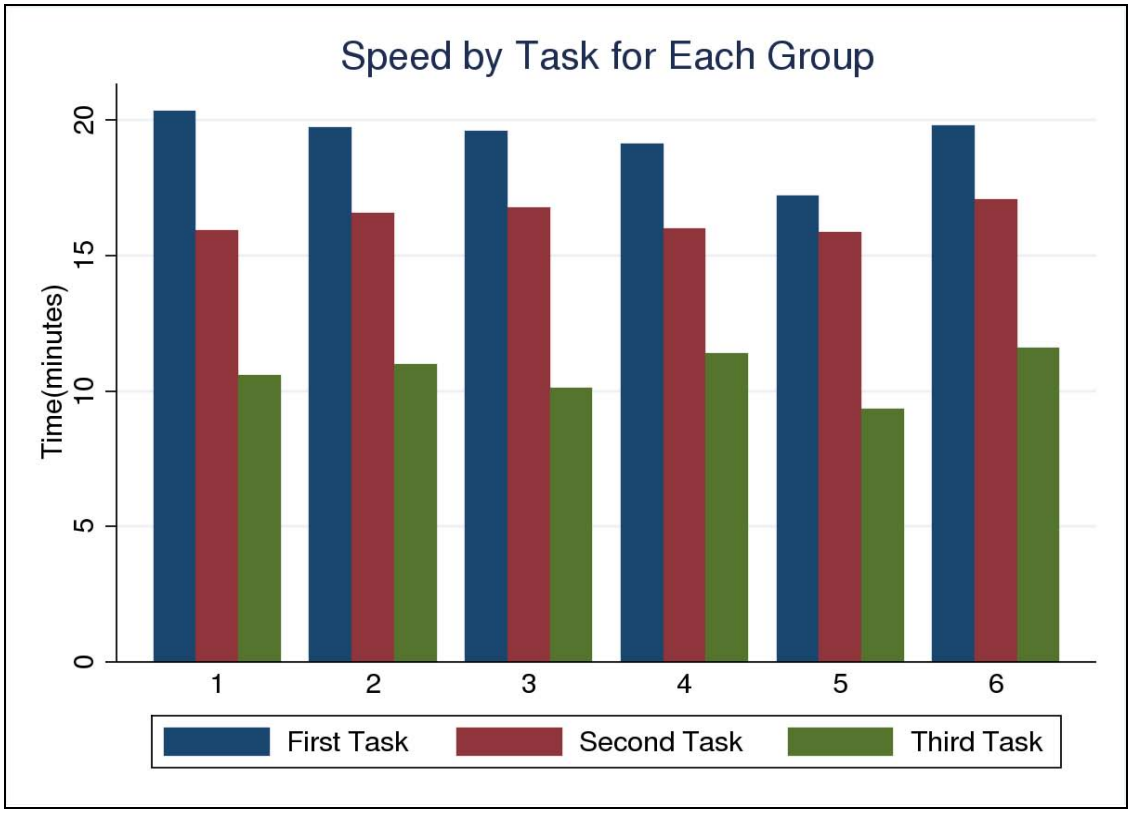


Figure 5.2 *Time, in minutes, for each task, broken down by group*

5.3 Accuracy Measurements

The second goal of the workbench is to accurately collect the semantic data; thus, I also analyzed the accuracy of the annotations. The workbench records the sense the subject chooses for each word annotated during the task. I compared the sense chosen by the subject to the correct sense, based on the gold standard annotation. I used the kappa statistic to get a measurement of accuracy between each subject and the gold standard. To calculate kappa, I measured the percent agreement between the subject and the gold standard and then controlled for the percent agreement expected by chance, using the following equation.

K = (agreement - agreement expected by chance) / (1-agreement expected by chance),

where the agreement expected by chance is 1/(the number of possible definitions for a word) averaged over all words. For more details about calculating kappa, see Cohen. The results are presented in Table 5.2.

Task/Method	Kappa
First Task	0.4339
Second Task	0.4024
Third Task	0.5121
Random Method	0.4097
Most Common First Method	0.4766
Lesk Method	0.4621

Table 5.2 *Accuracy Results*

Again, the ordering of the tasks produced significant differences in the kappa statistic of the annotations between the tasks, with *p* values of 0.01 between the first and third task and between the second and third task. Figure 5.4 displays the kappa statistic for each task and Figure 5.5 shows the kappa statistic for each task broken down by group. The difference between the first and second task was not significant.

The results also suggest that the Most Common First and *Lesk Method*s produce more accurate results than the *Random Method*, with a *p* value of 0.01 between the *Random Method* and the *Most Common First Method* and a *p* value of 0.025 between the *Random Method* and the *Lesk Method*. I found no significant difference between the kappa statistic of the *Most Common First Method* and the *Lesk Method*. Figure 5.6 shows the kappa statistic for each method and Figure 5.7 show the kappa statistics for each method broken down by group.

The kappa statistics, although they show a significant difference from each other, are still relatively low. As noted earlier, automatic methods can produce accuracy levels of 60%, which is greater than most of the accuracy numbers presented in these results. However, the accuracy measure obtained in this study may be low for several reasons. First, the measure for expected agreement by chance is an upper bound, at 0.32, that only calculates the number of possible definitions presented to the subject and does not take into account the options to collocate or change the stem of a word. Figure 5.3 shows accuracy of the annotation, when chance agreement is set to zero, compared with the kappa statistic values and demonstrates that finding a tighter bound on chance agree should

increase the accuracy values. Second, for those subjects who had missing data, the missing data produces a mismatch thus lowering the accuracy number. Third, the accuracy increased the more time the subject spent using the tool, therefore, if subjects train for a longer period of time the accuracy may increase.

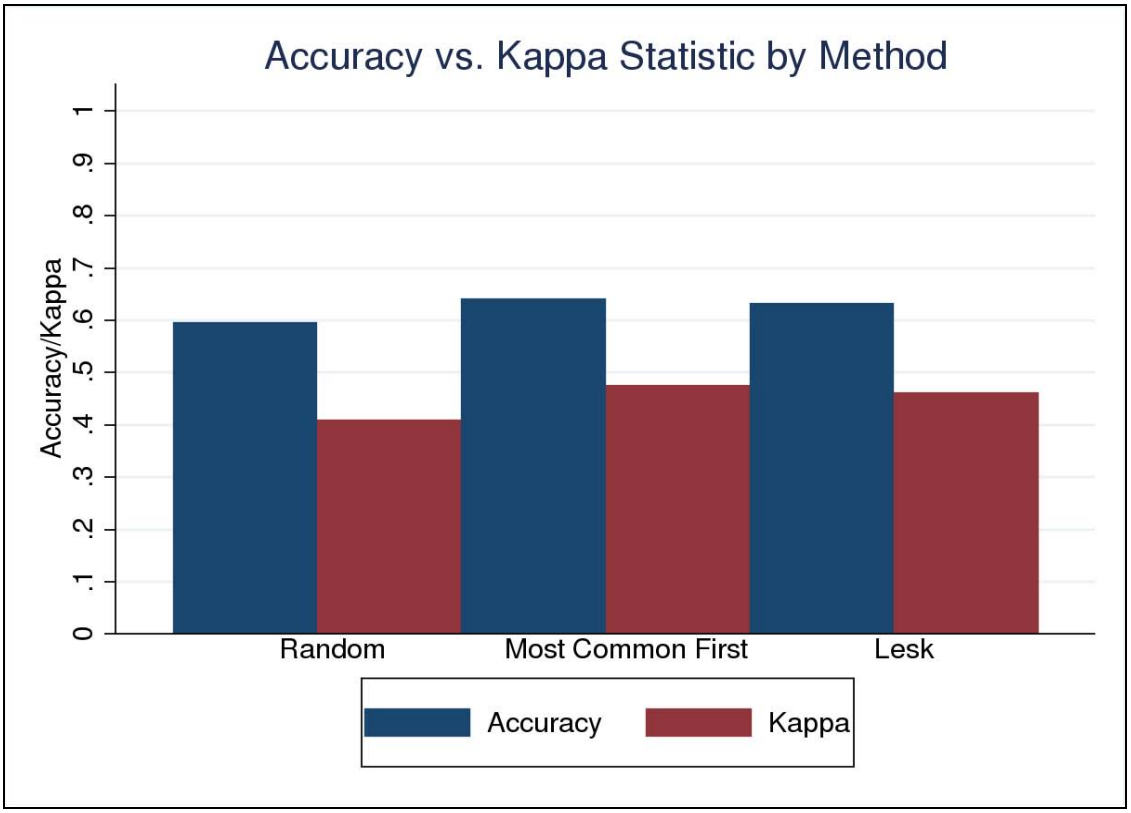


Figure 5.3 *The accuracy with chance agreement set to zero compared to the kappa statistic.*

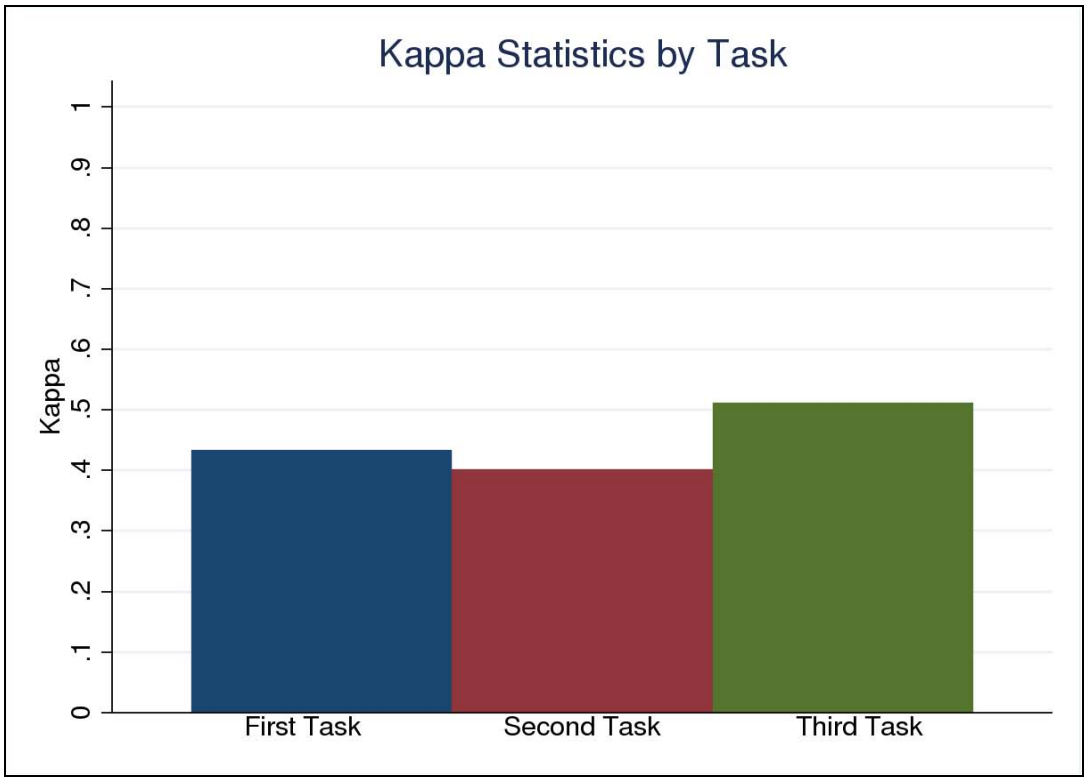


Figure 5.4 *Kappa Statistic by Task*

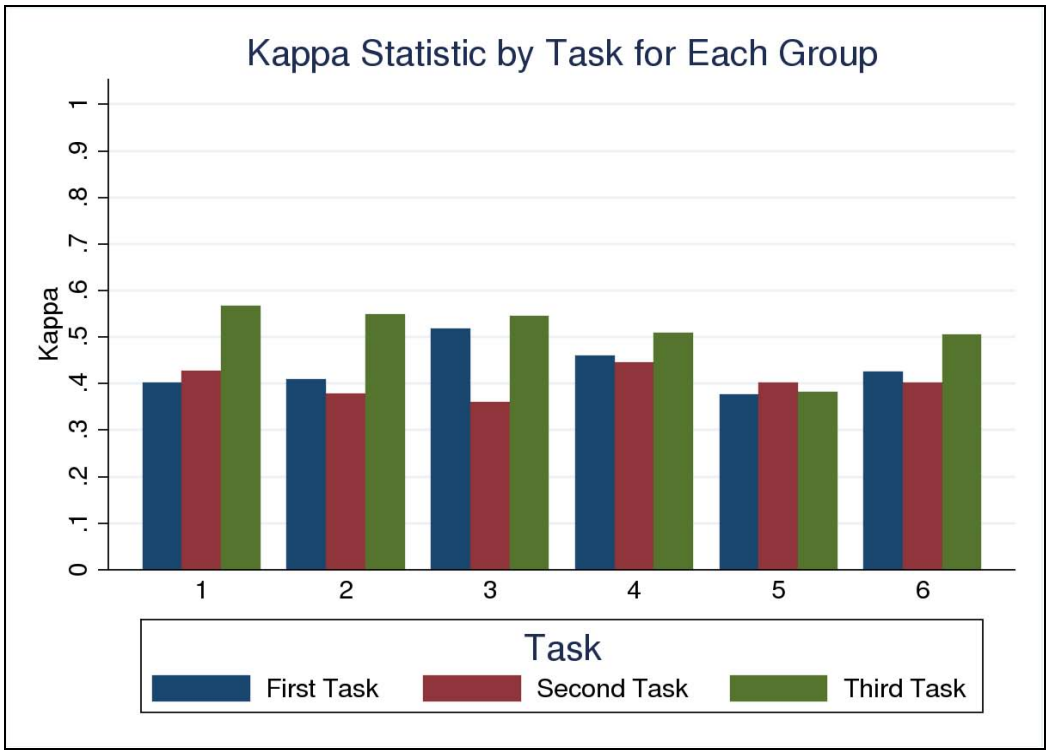


Figure 5.5 *Kappa statistic by task, broken down by group*

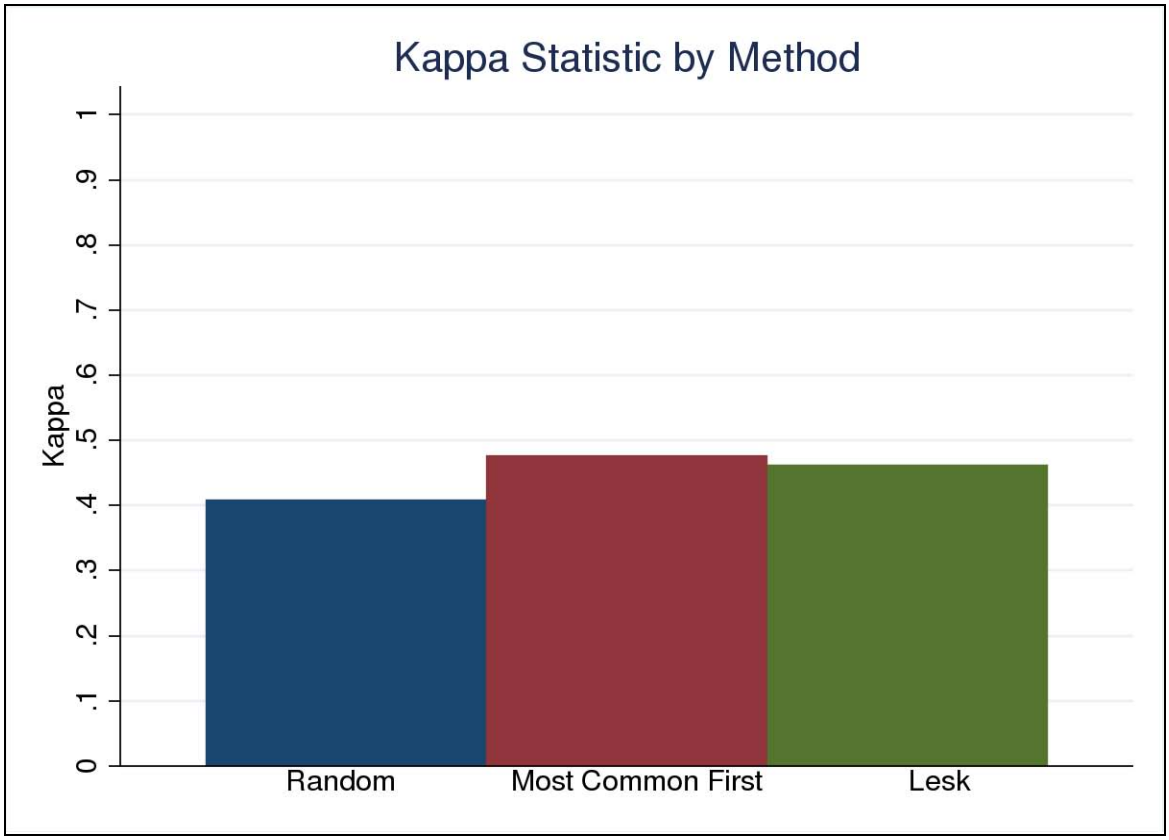


Figure 5.6 *Kappa statistic by method*

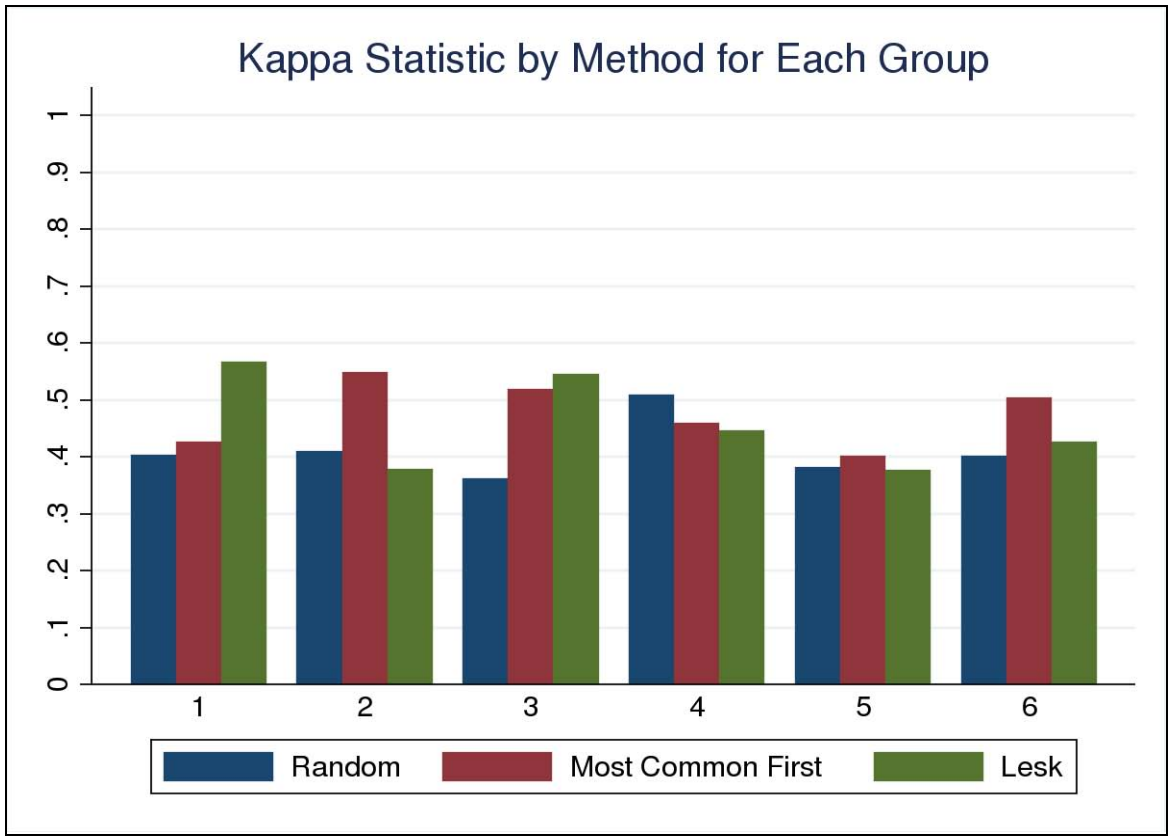


Figure 5.7 *Kappa statistic by method, broken down by group*

	Time	Std Dev	# obs	t	p	Kappa	Std Dev	# Obs	t	p
Task 1 -> Task 2	19.33	5.81	41	5.00	0.01	0.4339	0.0889	38	1.69	0.10
	16.59	5.04	41			0.4023	0.1094	38		
Task 1 -> Task 3	19.33	5.81	40	13.7	0.01	0.4339	0.0889	38	4.67	0.01
	10.72	3.25	40			0.5121	0.0940	38		
Task 2 -> Task 3	16.59	5.04	40	12.9	0.01	0.4023	0.1094	38	5.60	0.01
	10.72	3.25	40			0.5121	0.9040	38		
Random -> Common	15.73	5.90	40	0.12	0.91	0.4097	0.1084	38	3.13	0.01
	15.87	5.92	40			0.4766	0.0958	38		
Random -> Lesk	15.73	5.90	41	0.52	0.61	0.4097	0.1084	38	2.62	0.01
	15.17	6.23	41			0.4621	0.1086	38		
Common -> Lesk	15.87	5.92	40	0.76	0.45	0.4766	0.0958	38	0.66	0.51
	15.17	6.23	40			0.4621	0.1086	38		

Table 5.3 *Summary of Results*

6. Future Studies

The results of this study suggest two major extensions. First, the training task should be increased to counter the effects of ordering. A subject should train until they reach a threshold where more training does not significantly improve their speed and accuracy. By controlling for the effects of training, hopefully an effect between methods will emerge. Second, the task should be performed over several different tasks. The improved speed and accuracy for the third section could have been due to the words given in the third section of the text. If multiple texts are used one can control for the effects of a particular text.

Other potential extensions include varying the order in which words are presented to the user. For example, annotations may be faster if the wizard asks subjects to annotate every instance of a particular word in the text before annotating a different word because after annotating the first instance user become familiar with the possible definition of the word. Also, other methods should be included in the study. For example, an improved Lesk method, that makes a guess for every word, may increase the speed and/or accuracy of the Lesk task. Similarly implementing a one sense per discourse heuristic may also improve the annotations.

7. Contributions

During the course of this project, I have developed a set of experimental methods that I used to test the Story Workbench. I found that the *Most Common First Method* and the *Lesk Method* lead to more accurate annotations than the *Random Method*. I have also identified the need for future tests that control for the ordering effect found from this study.

References

Cohen, Jacob. (1960). A Coefficient of Agreement for Nominal Scales. Education and Psychological Measurement. Vol. XX, No. 1.

Fellbaum, C., Ed. (1998). Wordnet: An Electronic Lexical Database. Cambridge, Ma, MIT Press.

Palmer, M., H. T. Ng, et al. (2006). Evaluation of WSD Systems. Word Sense Disambiguation: Algorithms and Applications E. Agirre and P. Edmonds. Dordrecht, The Netherlands, Springer. 33: 75-106.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, ACM SIGDC Proceedings of the 5th annual international conference on Systems documentation, p. 24 - 26.

Semcor. Princeton University. http://www.cs.unt.edu/~rada/downloads.html

Appendix A: Texts

Training Text

Principals of the 13 schools in the Denton Independent School District have been reelected for the 1961-62 session upon the recommendation of Supt. Chester O. Strickland.

Text

Committee approval of Gov. Price Daniel's "abandoned property" act seemed certain Thursday despite the adamant protests of Texas bankers.

Daniel personally led the fight for the measure, which he had watered down considerably since its rejection by two previous Legislatures, in a public hearing before the House Committee on Revenue and Taxation.

Under committee rules, it went automatically to a subcommittee for one week. But questions with which committee members taunted bankers appearing as witnesses left little doubt that they will recommend passage of it.

Daniel termed "extremely conservative" his estimate that it would produce 17 million dollars to help erase an anticipated deficit of 63 million dollars at the end of the current fiscal year next Aug. 31.

He told the committee the measure would merely provide means of enforcing the escheat law which has been on the **books** "since Texas was a republic". It permits the state to take over bank accounts, stocks and other personal property of persons missing for seven years or more.

The bill, which Daniel said he drafted personally, would force banks, insurance firms, pipeline companies and other corporations to report **such** property to the state treasurer. The escheat law cannot be enforced now because it is almost impossible to locate **such** property, Daniel declared.

Dewey Lawrence, a Tyler lawyer representing the Texas Bankers Association, sounded the opposition keynote when he said it would force banks to violate their contractual obligations with depositors and undermine the confidence of bank customers.

"If you destroy confidence in banks, you do **something** to the economy", he said. "You take out of circulation many millions of dollars". Rep. Charles E. Hughes of Sherman, sponsor of the bill, said a failure to enact it would "amount to making a gift out of the taxpayers' pockets to banks, insurance and pipeline companies".

His contention was denied by **several** bankers, including Scott Hudson of Sherman, Gaynor B. Jones of Houston, J. B. Brady of Harlingen and Howard Cox of Austin.

Cox argued that the bill is "probably unconstitutional" since, he said, it would impair contracts. He also complained that **not** enough notice was given on the hearing, since the bill was introduced only last Monday.

Appendix B: Experimental Procedure

I. Introduction and Instruction

A. Background

Word sense disambiguation is the task of associating a given word in a text with a meaning (sense) that is different from other meanings associated with the word. Some of these distinctions between meanings are coarse while others are subtle. For example, the word *bank*, in its noun form, has two coarse meanings, one as a sloping piece of land and the other as a holder of money. These coarse meanings are usually not too difficult to distinguish when assigning word senses. However, under the meaning of a holder of money, there are more subtle divisions. For instance, the word *bank* could refer to the financial institution of a bank, the physical bank building, or a personal money container such as a piggy bank. The need to distinguish between these subtle meanings is what makes the task of word sense disambiguation difficult.

In this study, we are testing a method of word sense disambiguation to determine how useful it is for the task of assigning the appropriate meaning to a word in context.

B. Instructions

During the course of this study, you will perform three different tasks.

1. Training

First, you will be given a practice task that will familiarize you with the tool you will use to perform word sense disambiguation.

2. Annotation Task

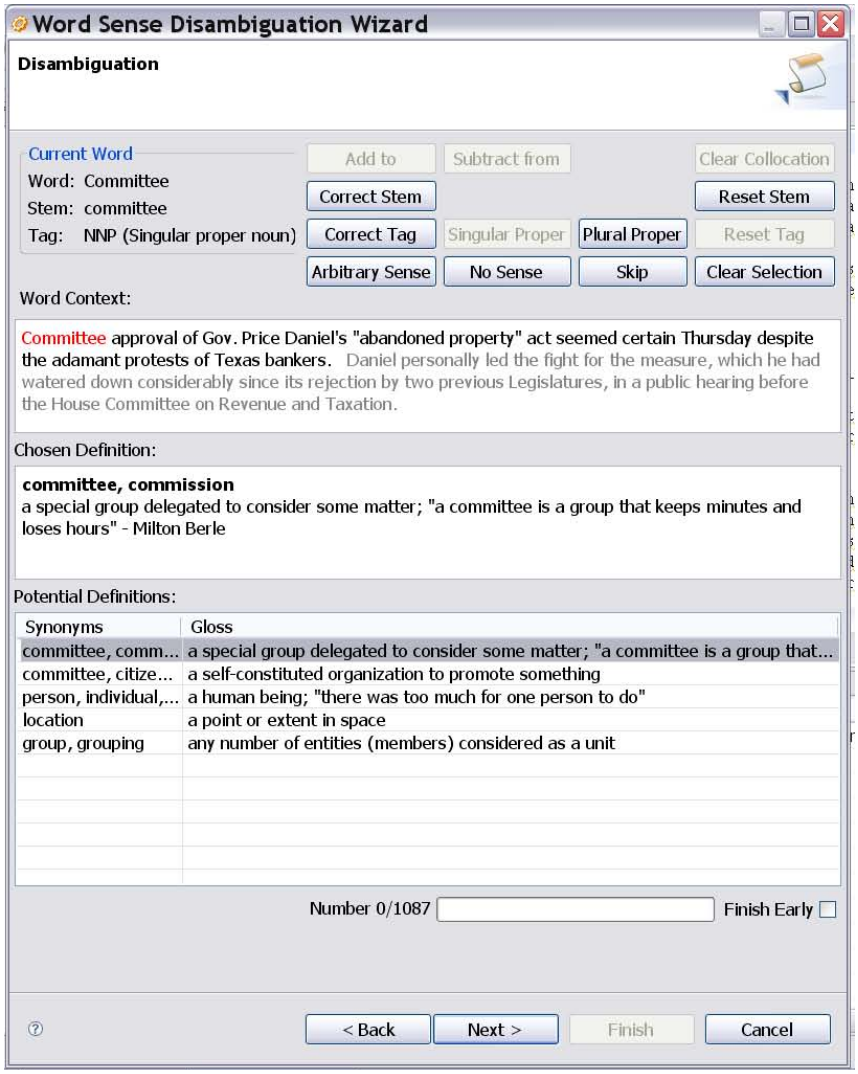
After you have completed the practice task for the method, you will move on to the actual test. You will be given a text and ask to tag the words with the appropriate sense using the method that you just learned in the training session.

3. Survey

After completing these tests, you will be asked to fill out a survey about your cultural and linguistic background, which we will use to analyze differences in the word sense assignments between subjects. The survey will also ask you questions about your experience using the provided annotation tool.

C. Directions for tagging

The program will guide you through this task in the following manner. First, you will be asked to read a text of approximately 250 to 2000 words. Once you have read the text, you will be asked to answer a few reading comprehension questions about the text. Next, the program will show you three sentences from the text with one word highlighted, as shown in the Word Context box in the figure below. The highlighted word is the word that you will annotate. In a box below the sentences, you will be given a list of potential definitions for the word you are annotating. You will use the up and down arrow keys or the mouse to navigate to the correct sense. You should use the context provided in the surrounding sentences to choose the correct sense of the word. If the appropriate sense is not provided you may choose the No Sense option. After choosing the correct sense for the highlighted word, click next and the program will provide you with another word from the text to annotate. You will repeat the annotation task for each word in the text.



D. Tagging Hints

Make Collocation

If two or more words next to each other in the text should be considered together you should make a collocation. For example, names of people, groups, or locations, such as Professor John Smith or New York Times should be collocated. Other instances might include two word verbs such as take off or take over. To make a collocation highlight the words that should be together and click Add To. You made undo the collocation by selecting Clear Collocation or you may remove words from the collocation by highlighting the words you wish to remove and then selection Subtract From.

Correct Stem

A stem is the part of the word that remains unchanged by tense and agreement. For example the stem of the words *running*, *runs*, or *ran* would be *run*. In some cases, you will need to correct the stem of a given word. If there are no definitions for a word or the definitions given do not fit the word (this will often happen when you make a collocation) you probably need to correct the stem. For example, if you just made a collocation for take over you will need to change the stem from take to take over to get the proper definitions. If you know what stem to use click on correct stem and type the stem into the box. If you are unsure of the correct stem click on choose arbitrary sense and start typing what you think the stem might be. As you type you will be given a list of senses that match the word you are typing. Choose the correct definition from the list. You may reset the stem to its original form at any time by clicking Reset Stem.

Correct Tag

In some cases, you may need to correct the part of speech tag associated with a given word. You should do this if you feel the part of speech is incorrect for the word given or the workbench gives you an error message that the chosen definition does not match the given part of speech tag. To correct the tag, click on correct tag and choose the appropriate part of speech. You may reset the tag to its original form at any time by clicking Reset Tag.

Appendix C: Reading Comprehension Questions

Training Text Questions

1. What is this story about?
   a. the Denton school district
   b. principals
   c. re-election
   d. all of the above

2. Who recommended the principals for reelection?
   a. the governor
   b. the superintendent
   c. Senator Grover
   d. no one

3. What district were the principals from?
   a. Garland
   b. Rowlett
   c. Denton
   d. Richardson

4. What is the length of time they have been elected for?
   a. ten years
   b. five years
   c. two years
   d. one year

5. How many schools were principals elected for?
   a. 13
   b. 1
   c. 20
   d. 10

Text Questions

1. What had Governor Daniel done since the act had previous been rejected?
   a. threatened those who opposed the act
   b. made the act less stringent
   c. changed his opinion
   d. added more provisions to the act

2. What does the abandon property act do?
   a. allow banks to be in charge of any property that was abandon while in the bank's care
   b. require banks to report abandoned property

c. allow the state to take over bank accounts, stocks and other personal property of persons missing for seven years or more
d. b and c

3. Who is protesting the act?
   a. bankers
   b. farmers
   c. citizens
   d. Governor Daniel

4. Why do they argue the bill is unconstitutional?
   a. it is discriminatory
   b. it will impair contracts
   c. it violates individual rights
   d. none of the above

5. Why does Dewey Lawrence argue that the act will hurt the economy?
   a. the government will take money that would have otherwise been invested
   b. banks will lose the money they use to make loans
   c. people will lose confidence in banks and stop investing money in them
   d. a and b

Appendix D: Culture Survey

Semi-Automatic Semantic Annotation for Cross-Cultural Analysis

Whitman Richards, Mark Finlayson, and Diana Moore

Participant Survey

Basic Information

Subject ID Number _____

Gender [ ] Male [ ] Female

Highest Degree Earned

[ ]Highschool or GED [ ]Some Highschool [ ]Some College
[ ]College Degree [ ]Masters or Professional Degree [ ]Doctoral Degree

Cultural Origin

Country of Birth_____

Countries where you were raised_____

Ethnicity

[ ]White/Caucasian [ ]Black [ ] AsianIEast-Asain [ ] Middle-Eastern
[ ]Native American [ ] Hispanic/Latino [ ]Other

What is the name by which you call your ethnic group,if any? (e.g. African-American, Mexican-American,Northern Irish, Pakistani, Asian-American, etc.)_____

Language Proficiency

Please rate your fluency on a scale from **1** to ***5,*** where:

1 = Elementary Proficiency
2 = Limited working proficiency
3 = Professional Working Proficiency
4 = Full Professional Proficiency
5 = Native or Bilingual Proficiency

First Language _____Fluency_____
Second Language_____Fluency_____
Third Language_____Fluency_____
Other Languages_____

Background Knowledge

Please describe any experience you've had with programming, linguistics, artificial intelligence, or cognitive science. Please describe relevant classwork, if any

Appendix E: User Experience Survey

Semi-Automatic Semantic Annotation for Cross-Cultural Analysis

Whitman Richards, Mark Finlayson, and Diana Moore

User Experience Survey

Subject ID Number _____

Please answer the following questions about your experience using the text annotation tools.

1. Which method did you prefer using to annotate text?

a. First Method

b. Second Method

c. Third Method

Why?

2. Which method was faster?

a. First Method

b. Second Method

c. Third Method

Why?

3. Which method was easier to use?

a. First Method

b. Second Method

c. Third Method

Why?

4. Did you feel that one method helped you to produce better annotations than the other? If so, which one?

a. Yes, First Method

b. Yes, Third Method

c. Yes, Second Method

d. No

Why?

5. Was there anything unclear about how to perform the annotation?

a. Yes

b. No

If so. please explain

6. What improvements could be made to make the tools easier to use?