Towards Improving Synonym Options in a Text Modification Application

Jill Burstein
Educational Testing Service
Princeton, NJ 08541
jburstein@ets.org

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812
tpederse@d.umn.edu

November 2, 2010

Abstract

The Language Muse^{TM}system (LM)[1] is an application that supports classroom teachers in text modification for middle- and high-school learners who are non-native English speakers (NNES). The application performs linguistic analysis on classroom texts - highlighting lexical, syntactic, and pragmatic features, indicating potential areas of linguistic complexity. One of the lexical feature options that LM offers are synonym candidates for words in a text. This study investigates how to improve the current synonym detection feature, using a distributional method that identifies similar words [Lin98, LC03] and WordNet. For single words and multi-word expressions, human judges annotated "acceptability" of synonym candidates from both resources. Humans attained high levels of agreement. Outcomes indicated that the distributional method and WordNet provide complementary options that together could provide improved resources.

1 Introduction

Research in synonym detection and lexical substitution has application in a number of areas of NLP research and applications, including word sense disambiguation [MKW04], textual entailment [CHT$^+$07, DW09], paraphrase research [BM01, Boo04], question answering [CFH08, KMB$^+$05], and machine translation [CBKO06]. Synonym detection research also has a place in

[1]The Language Muse system was previously known as Text Adaptor.

educational applications. Synonym detection is used to identify paraphrases in automated scoring of constructed responses [LC03]. Automated essay scoring systems that use Latent Semantic Analysis examine semantic similarity between gold-standard student or test-taker responses [FKL98, LLF00]. Recent work successfully uses synonyms of words in an essay test question (topic) to improve off-topic essay detection for an automated essay scoring application [LH10]. WordNet [Mil95, Fel98] is used in commercial applications, such as Visual Thesaurus (www.visualthesaurus.com) in the context of vocabulary teaching. Freely available synonym systems, such as WordNet, and Lin's (1998) distributional method are being used increasingly to develop educational applications. This motivates the need to explore the real-life use of these resources. Specifically, how do distributional methods and WordNet complement each other, and how might we take advantage of this complementary behavior to improve these current resources?

In previous related work, Lin and Pantel [LP02] investigate using WordNet to support a clustering algorithm. In the context of preposition attachment, Calvo, Gelbukh, and Kilgarriff [CGK05] investigate the performance of distributional methods as compared to WordNet. Agirre, Alfonseca, Hall, Kravalova, Pasca & Soroa [AAH$^+$09] investigate the use of WordNet and distributional methods for cross-lingual similarity. In educational applications, Sukkarieh & Stoyanovich [SS09] have compared Lin (1998) to WordNet in the context of building gold-standard responses for a short-answer scoring system.

1.1 General Approach

WordNet [Mil95, Fel98] is a widely-used lexical database containing word sense and word relation information for about 150,000 English nouns, verbs, adjectives, and adverbs. Lin (1998) is a distributional method used to detect similar words. LM uses a statistical method to model words in context in large corpora from which a matrix of words is produced. The matrix contains the target words, and similar words with similarity values that indicate the likelihood that the similar words are related. Values are between 0-1, and values closer to 1 indicate a higher likelihood of similarity. In Lin (1998) work, newswire texts are used to build the matrices.

The goal of this study is to determine if WordNet and Lin synonyms provide complementary information, and therefore, might be used to supplement one another. More specifically, the outcomes of the evaluation will be used to inform development of LM, an educational application that supports teachers' understanding of linguistic complexity in text, and helps them

to develop language-appropriate materials for non-native English speaker (NNES) learners (see section 2.1).

A straightforward annotation task was completed. Annotators labeled the "acceptability" of synonym candidates from WordNet and the Lin matrices for single words and multi-word expressions in middle- and high-school social studies, science and language arts texts. Results indicated that humans attained high levels of agreement in evaluating candidate synonyms. Also, both WordNet and the Lin matrices provide complementary synonyms that together are richer than either source individually.

2 Synonyms for Vocabulary Explanation

Throughout the United States, non-native English speakers (NNES) often comprise a large proportion of classroom students. Middle- and high-school classroom teachers who teach subject areas (e.g., social studies, science) often find themselves having to modify or develop additional explanation around vocabulary in subject-area texts to ensure that the language is appropriate for NNES learners [CMR07].

While there are many kinds of linguistic complexity in text that may interfere with a NNES learner's comprehension (lexical, syntactic, or pragmatic), difficult or unfamiliar vocabulary can be a big contributor. That said, teachers often use synonyms to support *basic comprehension* (using easier words and additional explanation), or *vocabulary development* (offering more difficult words).

2.1 The Language Muse^{TM}system

Subject-area teachers are often not specifically trained to do text modification - that is, to make adjustments to language in a text to accommodate NNES learners. The working assumption is that if teachers are trained to recognize linguistic complexity in text, they will be better prepared to create instructional materials that can be used by all students in their classrooms.

In light of this, LM is an application that was developed to support teachers' linguistic awareness of language and structure of text that could impede a NNES learner's comprehension of core content (Anonymous). Currently, LM allows teacher users to import a text, and offers analysis of a number of lexical and syntactic features (including, key concepts, synonyms, cognates, morphologically complex words, contractions, confusable words, complex noun phrases, lengthy prepositional phrases, and complex sentences).

Teachers can use the linguistic analysis to modify texts, either by adding explanation (e.g., synonyms), or by simplifying complex structures (e.g. partitioning complex sentences). The system is also being developed so that teacher users can develop activities and assessments around a text. LM has been piloted in teacher training programs to develop teachers' sensitivity to linguistic aspects in subject-area text that could interfere with NNES' accessibility to core content. As part of the pilot studies, post-study surveys have been administered to collect feedback from teachers about what they like about the system, and where they would like to see improvement. While teachers appreciate the synonym support, they also consistently comment that they are not always happy with the synonym candidates which are sometimes too difficult or inappropriate. The teachers' feedback has forced us to consider how we might improve the synonym candidates offered, and has motivated the following research questions:

1. Can a distributional method and WordNet resources be combined to provide better synonym candidates in LM?, and

2. Can a distributional method and/or WordNet provide a reasonable source of candidate synonyms for multi-word expressions, currently not available in LM?

In the context of educational applications, this work also has implications for corpus-building of thesauri for NNES learners. More broadly, this work furthers our understanding of straightforward differences between WordNet and a distributional method that could support other kinds of NLP applications that require synonym detection.

3 Experiment 1 Data : Single Words

In the single word synonym annotation task, five typical classroom texts were used: two social studies texts, two science texts, and one language arts text. Texts spanned grades 5, 7, 8, 9 and 12. The number of words per text is as follows, by grade order, 902, 287, 374, 300, and 855. While there were only five files in this task, the annotators had to evaluate thousands of synonyms.

3.1 Data Preparation

For each text, Lin and WordNet synonym candidates were provided for target words using the same criteria that LM uses to select words for which

synonyms will be provided. It is important to point out that LM uses a modified set of Lin matrices (referred to henceforth as modLin). These modLin matrices were built for educational software using 300 million words of text from a corpus of fiction, non-fiction, and textbooks, and newswire from the San Jose Mercury News (see Leacock and Chodorow (2003) for details).

Words (in the 5 texts) for which synonyms are provided are referred to as *target words* in this paper, and selection happens as follows. In LM, synonyms are provided for words based on a word frequency setting provided by the teacher user. Users can request synonyms for higher (more common) or lower (more rare) frequency words. The system will then provide synonyms for all words equal to or less than the user selected frequency. Frequencies are determined using a standard frequency index based on Breland, Jones, and Jenkins (1994)[2].

The default frequency (used in this study) indicates that the word appears once in approximately 10,000 words. For the target words that meet the frequency criterion, LM will provide synonyms if the modLin similarity value is greater than or equal to 0.172. This default setting was determined based on discussions with users about their satisfaction with synonym candidate. The 0.172 value corresponds very closely to the mean value (0.174) of all similarity values in the modLin matrices. To prevent over-generation of synonym candidates, only the first 4 synonyms that adhere to the 0.172 threshold are offered.

For this experiment, for the five texts, we used the existing criteria in LM (described above) to generate up to 4 modLin synonym candidates, which were augmented with candidate synonyms from WordNet. To prevent over-generation of the WordNet synonym's candidates, we select only the first 3 senses associated with each possible part of speech (according to WordNet) of the target word. Note that if a sense for a given part of speech only provides the target word itself as synonym, then it is skipped and the next sense is used.

3.2 Annotation Task

Two annotators completed the task. Both work in education, and have linguistics background. The annotators were given a labeling protocol with the task description and examples. They communicated with the authors by

[2]The formula to determine a word.s standard frequency index value is as follows. SFI = 10(Log10(1,000,000*F/N) + 4), where F is the word frequency and N is the total number of words.

e-mail if there were questions about the task. Both annotators were privy to all questions and answers about the protocol. The core task was to go through the texts, and where candidate synonyms were provided for a word, place an in front of a candidate that could be a synonym for that word. Synonyms appeared in brackets after the target word.

The central task was expressed as follows in the annotation protocol: *The purpose of this task is to identify, from a list of possible candidates, synonym candidates that could be appropriate substitutions for the original word given its use in a sentence.*

4 Experiment 1 Results : Single Words

The annotations described above were analyzed with the following measures. **Coarse-grained measures** treat all the candidates generated for a given target word as a set, and evaluate if both annotators selected any one of the candidates from the set, without regard to which one. **Coarse grained agreement (CGA)** is the sum of the probability that both annotators agreed that there was one valid candidate for a given target word, and the probability that they both agreed that there were no valid candidates for the target word. **Coarse-grained precision (CGP)** is the probability that the annotator found at least one valid candidate for a target word.

Fine-grained measures examine each candidate individually, and compare when annotators agreed on whether or not a specific candidate was a viable synonym for the target word. **Fine-grained agreement (FGA)** is the sum of the probability that the annotators agreed that a candidate was viable and the probability that they agreed a candidate was not viable. **Fine-grained precision (FGP)** indicates how many of the candidates each agreed was viable. We also compute Cohen's kappa (unweighted) to assess fine and coarse grained agreement [Coh60], and interpret those scores according to the Landis and Koch (1977) scale.

4.1 Overall Results

The single word annotation task was carried out two times: in May 2010, and in June 2010 following a period of discussion where the annotators resolved differences in their interpretation of the task. It is important to note that in the context of LM, teachers are looking at synonym candidates not only as substitutes for words, but also as a means of explanation. For instance, for the word *sports*, the candidates, *basketball*, *baseball*, *football*, exemplifying types of sports might offer helpful explanation to a learner who may

Table 1: 5-files (May) : Kappa = 49.03

Coarse	A1 Yes	A1 No	total
A2 Yes	(55%) 405	17	(57%) 422
A2 No	159	(22%) 162	321
total	(76%) 564	179	743

be familiar with a particular sport. In light of this, the definition of strict substitute was not necessarily appropriate for this task. The need to redefine "acceptable synonym" required a second annotation round (June 2010). In June, the annotators re-calibrated and settled on an altered definition in which synonyms were selected based on their appropriateness for likely teaching goals related to basic vocabulary comprehension and vocabulary development.

In this task for the May and June rounds, the annotators evaluated synonym candidates generated for 743 single target words found in the five texts. Together, modLin and WordNet provided a total of 7,171 candidate synonyms for these target words. Results from both annotations are presented below. Coarse-grained agreement and precision is shown in Tables 1 and 2, and fine-grained figures appear in Tables 3 and 4.

Note that Tables 1 – 6 and 10 – 13 are cross-classification (i.e., contingency) tables. Rows represent one of the annotators and the columns represent the other. The number of cases where the annotators agree is shown on the diagonal, and their disagreement is shown in the off-diagonal. For example, Table 1 shows that there were 405 target words for which both annotators agreed that at least one candidate synonym was viable. They also agreed that for 162 target words none of the candidates were acceptable. Thus, their total coarse-grained agreement (CGA) was 77% (55% + 22%). The marginal totals in Table 1 show that one of the annotators selected many more target words with at least one synonym (A1 found 564, while A2 found 422). These marginal totals are used to compute the coarse grained precision (CGP) for each annotator. This is simply the ratio of their individual counts of viable synonyms with the total number of candidates. In Table 1 A1 has coarse-grained precision of 76% (564/743) while in the case of A2 it was 57% (422/763). This shows that in general A1 was more satisfied with the candidates than was A2. Finally, for all tables with agreement data we also show Cohen's kappa in the table title.

The coarse- and fine- grained results from the May annotation pointed

Table 2: 5-files (June) : Kappa = 75.2

Coarse	A1 Yes	A1 No	total
A2 Yes	(65%) 484	9	(67%) 495
A2 No	68	(24%) 180	248
total	(74%) 552	191	743

Table 3: 5-files (May) : Kappa = 56.49

Fine	A1 Yes	A1 No	total
A2 Yes	(13%) 919	199	(16%) 1,118
A2 No	807	(73%) 5,250	6,053
total	(24%) 1,726	5,445	7,171

to systematic differences in how the annotators had approached the task (we observed that A2's synonyms were nearly a subset of A1's). We asked the annotators to discuss their differences, and repeat the annotation approximately one month later. After both rounds of annotation, A2 reported that she had viewed the task as a "strict lexical substitution task" during the May annotation, and that she relaxed her criteria to also include synonyms that could be used for explanatory purposes.

As a result of this change, Tables 1 and 2 show that the overall coarse-grained agreement rose to 89% (65% + 24%), and kappa increased to 75.2 (substantial agreement). In addition, Tables 3 and 4 show that there was a very significant increase in the overall fine-grained agreement (going from kappa 56.49 (moderate agreement) to 78.87 (substantial agreement).

The coarse- and fine- grained results from the May and June annotations indicate that modLin and WordNet, together, are generating a number of acceptable (viable) candidates. There was at least one valid candidate for 65% of the target words (coarse grained) and approximately 18% of the generated candidates were valid (fine grained).

However, we want to understand, further, whether annotators prefer synonyms from WordNet or modLin. To answer this, we conducted a fine-grained analysis that broke down the candidate synonyms based on their origin (modLin or WordNet) as well as by their underlying part of speech and sense distinctions.

Table 4: 5-files (June) : Kappa = 78.87

Fine	A1 Yes	A1 No	total
A2 Yes	(18%) 1,308	112	(20%) 1,420
A2 No	407	(75%) 5,344	5,751
total	(24%) 1,715	5,456	7,171

4.2 WordNet versus modLin

In the following analyses we break down the 7,171 candidates into two groups, those that came from WordNet (5,036 candidates, 70%) and those that came from modLin (2,135 candidates, 30%). Only 275 candidates are offered by both WordNet and modLin for the same given target word. This indicates that the two methods are potentially highly complementary, and that it is relatively rare for a target word to have overlapping candidate synonyms in WordNet and modLin. For example, for the target word *rascal*, modLin offers *scoundrel*, while WordNet offers *rogue, knave, rapscallion, scalawag, scallywag, varlet, imp,* and *scamp* For the target word *hated*, modLin offers *despise*, while WordNet offers *detest*.

Tables 5 and 6 show the agreement between annotators when evaluating synonym candidates that come from WordNet and modLin, respectively. These tables show that approximately the same number of candidates were selected as viable from each source (702 for WordNet versus 606 for modLin). However, since WordNet offers more candidates than modLin (5,036 versus 2,135) the fine-grained precision of the annotators is much higher in modLin than in WordNet. With the modLin synonyms, A1 selected 673 (32%) as valid, while A2 selected 643 (30%). Fine-grained precision for WordNet is much lower (21% for A1 and 15% for A2). Overall kappa is significantly higher with modLin (88.58, near perfect agreement) as opposed to WordNet (72.59, substantial agreement). This suggests that the options provided by modLin are generally easier to evaluate (since the annotators are more likely to agree). WordNet may have some difficulties in this regard since it includes archaic and obscure senses, whereas the modified thesaurus generated by Lin's method focuses on relatively common words. In addition, neither WordNet nor modLin incorporate word sense disambiguation (at this time) and so candidates that are not used in a sense appropriate to the context could be generated.

Table 5: 5-files (June) WordNet : Kappa = 72.29

Fine	A1 Yes	A1 No	total
A2 Yes	(14%) 702	75	(15%) 777
A2 No	340	(78%) 3,919	4,259
total	(21%) 1,042	3,994	5,036

Table 6: 5-files (June) modLin : Kappa = 88.58

Fine	A1 Yes	A1 No	total
A2 Yes	(28%) 606	37	(30%) 643
A2 No	67	(67%) 1,425	1,492
total	(32%) 673	1,462	2,135

4.3 Part of Speech Distinctions

Tables 7 and 8 examine agreement, based on the underlying part of speech of the candidates. WordNet candidates carry with them part of speech tags, however, the modLin candidates are not associated with a part of speech. If a WordNet candidate was "selected" for a given target word, then we assumed that all the modLin candidates would also use that same part of speech. While this is not foolproof, it is not possible to use a part of speech tagger or other tool since the modLin candidates only appear in a list, and the modLin matrices do not provide information on part of speech. This assumption allowed us to assign parts of speech to 2,063 of the 2,135 (97%) modLin candidates.

Note that Tables 7 – 9 and 14 – 16 are not cross-classification tables, but rather summary tables that break down the results of a previous table in more detail. For example, Table 7 provides a part of speech level breakdown of Table 5. Tables of this form indicate the count of candidates which the annotators agreed were viable (syn) and the number that they agreed were not viable (no syn). Fine-grained agreement (FGA) is shown, which is the sum of the probabilities associated with the syn. and no syn. counts above. Finally, we show the Kappa value for the agreement between the annotators for each category.

Table 7 shows the 5,036 WordNet candidates divided into nouns, verbs, adjectives and adverbs, and Table 9 shows the 2,036 modLin candidates by part of speech. It should be noted that the number of noun and verb candidates from WordNet is nearly the same (2,010 nouns and 2,191 verbs,

Table 7: 5-files (June) : WordNet by PoS

Fine (5,036)	Noun (2,010)	Verb (2,191)	Adj (583)	Adv (252)
syn.	281	207	130	84
no syn.	1,470	1,877	425	147
FGA	87%	95%	95%	92%
Kappa	61.07	76.7	87.1	82.25

Table 8: 5-files (June) : modLin by PoS

Fine (2,063)	Noun (1,554)	Verb (188)	Adj (206)	Adv (115)
syn.	411	63	83	46
no syn.	1,065	118	113	65
FGA	95%	96%	95%	97%
Kappa	87.51	91.86	90.09	92.85

which together account for 83% of the candidates). By contrast, 1,554 of the 2,063 (73%) modLin candidates are nouns, while just 40% of the WordNet candidates are nouns.

In general the annotators had much higher levels of agreement for modLin nouns and verbs (almost perfect agreement) as compared to WordNet nouns and verbs (substantial agreement). This might partially be due to the fact we do not filter WordNet for archaic or extremely obscure synonyms, which might lead to some confusing candidates. We do, however, have a filter for obscene language for WordNet.

Tables 7 and 8 illustrate that the annotators preferred a larger number of the modLin nouns to WordNet nouns (411 versus 281). Given the smaller number of modLin candidates the overall fine-grained precision associated with modLin was much higher (26% for modLin nouns versus 14% for WordNet nouns). Interestingly, while modLin generated a much smaller number of verbs (188 versus 2,191), the percentage of the modLin candidate verbs judged as viable (fine-grained precision) was much higher (34% versus 9%). The same pattern was also observed with the adjectives and adverbs (where modLin generates a smaller number of candidates, but with higher fine-grained precision).

Table 9: 5-files (June) : WordNet by sense

Fine (5,036)	# 1 (2,435)	# 2 (1,474)	# 3 (1,105)	other (22)
syn.	395	159	137	11
no syn.	1,890	1,474	868	11
FGA	94%	89%	91%	100%
Kappa	80.25	59.71	68.08	100

4.4 Sense Distinctions

Table 9 shows the distribution of WordNet candidates across the different senses. In general the synonyms came from the first three senses of a candidate word, although in a few cases (22) we needed to look beyond these senses since the only synonym provided by a synset was the target word itself. The cases where the synonyms come from a sense higher than 3 are shown in Table 9 in the column *other*.

Table 9 shows that there was greater agreement among the annotators when considering candidates from WordNet sense 1. This is not surprising since WordNet sense 1 is generally a word's most common sense. Some senses beyond sense 1 can become obscure, and so annotator disagreement is more likely.

This suggests that accurate word sense disambiguation could certainly decrease the number of candidates and improve the precision of the candidates. However, incorrect word sense disambiguation could result in the failure to generate reasonable candidates, so at this point we elected to over-generate so as not to miss possible candidates.

5 Experiment 2 Data: MWEs

The second task was to annotate synonym candidates that were generated for multi-word expressions. We elected to keep this separate from the single word task since multi-word expressions generally have much less semantic ambiguity associated with them (than single words).

5.1 Data Set

A total of 11 texts (the 5 used by the two annotators in the single word task plus an additional 6) were used for this task. The text represented the

same genre (social studies, science and language arts texts from 5th - 12th grade). More texts could be used since there were fewer target multi-word expressions in each. MWEs come from a manually abridged version of a list of all the MWEs found in version 3.0 of WordNet (including phrasal verbs, collocations, and compound words). This abridged list is used in LM to highlight these kinds of terms to the user. Synonyms are currently not provided - only highlighting. One of the reasons for undertaking this study is to determine if synonym candidates can be added to LM for target MWEs by WordNet and/or modLin. For this task, the target MWEs were identified in the 11 texts, and candidate synonyms from WordNet and modLin were offered in the texts for the annotation task. Note that for these experiments we did not limit the modLin candidates to any threshold or number of occurrences, and we took all of the synonyms associated with all of the senses in WordNet as candidates. We opted to run the risk of over-generation since it seemed likely that MWEs would have fewer candidates in general due to their overall lower ambiguity.

5.2 Annotation Task

For this task, the authors completed the annotation. One of the authors has experience working with teachers in the context of LM, so has somewhat similar common experience to the first two annotators. As annotators, the authors labeled acceptable (viable) synonym candidates with , using the same annotation protocol as was used for the single word task.

6 Experiment 2 Results: MWEs

We carried out an analysis similar to that performed for single word synonyms, which includes an overview of coarse- and fine- grained results, a comparison of WordNet candidates versus modLin candidate, and breakdowns of results based on part of speech and senses.

6.1 Overall Results

Table 10 shows the coarse-grained agreement between the two annotators. Similar to the May annotation of the single word data, for this task, one of the annotators tended to select a subset of the other annotator's synonyms. The more selective annotator was holding the candidates to the standard of being substitutable for the target MWE, whereas the more liberal annotator interpreted the task in terms of whether or not the synonym might be used

Table 10: MWE : Kappa = 50.04

Coarse	B1 Yes	B1 No	total
B2 Yes	(35%) 175	26	(40%) 201
B2 No	100	(40%) 197	297
total	(55%) 275	223	498

Table 11: MWE : Kappa = 52.65

Fine	B1 Yes	B1 No	total
B2 Yes	(3%) 314	129	(4%) 443
B2 No	382	(92%) 9,483	9,865
total	(7%) 696	9,612	10,308

to offer additional explanation to a learner for the target MWE. This is most likely related to the fact that the annotator has experience working with teachers.

As a result of this difference in interpretation, the overall accuracy is moderate (75%) as is the kappa value (50.4). Given this level of agreement, we believe a resolution and discussion stage similar to that carried out by the single word synonym annotators would be helpful, and would likely raise agreement both at the coarse- and fine-grained level.

Table 11 shows the fine-grained analysis of the MWE synonym candidates. What is striking is the relatively small percentage of candidates that were judged as acceptable. While their fine-grained agreement is quite high (95%), this is mostly due to their agreement on candidates that are not suitable synonyms. In general, MWEs tend to be more specific and more limited in the range of contexts in which they can be used (compared to single words) so this may account for the very high percentage (92%) of candidate synonyms that are judged to be unacceptable by both annotators.

6.2 WordNet versus modLin

Given the high number of candidates that are agreed to be unacceptable, and the very low fine-grained precision for each annotator (7% and 4%), we wanted to again compare WordNet and modLin to see if either has a significant advantage in generating synonym candidates. Of the 10,308 candidates, 1,776 come from WordNet and 8,532 come from modLin. There are only 92 candidates generated by both modLin and WordNet for any given

Table 12: MWE WordNet: Kappa = 48.71

Fine	B1 Yes	B1 No	total
B2 Yes	(11%) 200	74	(15%) 274
B2 No	213	(73%) 1,289	1,502
total	(23%) 413	1,363	1,776

Table 13: MWE modLin : Kappa = 49.18

Fine	B1 Yes	B1 No	total
B2 Yes	(1%) 114	55	(3%) 169
B2 No	169	(96%) 8,194	8,363
total	(2%) 283	8,249	8,532

target word, so again we can see that these two methods are highly complementary. Tables 12 and 13 show the agreement between the annotators for the WordNet and modLin candidates. The quantity and percentage of candidate synonyms that are judged acceptable by both annotators is somewhat higher when they come from WordNet (200 and 11%) as opposed to modLin (114 and 1%). While fine-grained agreement is somewhat higher for modLin, this is again due to high agreement that a large number of the candidates are inappropriate. The high level of agreement occurs in part because certain target MWEs can generate a very large number of modLin candidates because they occur in similar contexts (and may often not be synonyms, resulting in a large number of synonyms that aren't viable).

6.3 Part of Speech Distinctions

As was the case in the single word task, we are interested in whether or not candidates associated with a certain part of speech tend to be more viable than others. Table 14 shows that WordNet tended to generate both noun (26%) and verbs (54%) candidates for synonyms, while Table 15 makes it clear that modLin tends to generate candidate synonyms that are nouns (75%) more often than other parts of speech. However, we observed that a few kinds of MWE target words resulted in very large number of candidates - for example any city name resulted in a list of approximately 50 other cities, most of which weren't viable synonyms. This occurred more with nouns, and resulted in a large number of noun candidates relative to the other parts of speech.

Table 14: MWE : WordNet by PoS

Fine (1,776)	Noun (464)	Verb (959)	Adj (82)	Adv (271)
syn.	56	53	6	85
no syn.	295	824	69	101
FGA	76%	91%	91%	69%
Kappa	35.86	51.65	58.71	39.38

Table 15: MWE : modLin by PoS

Fine (7,792)	Noun (5,874)	Verb (832)	Adj (100)	Adv (986)
syn.	52	39	2	16
no syn.	5,708	741	95	930
FGA	98%	94%	97%	96%
Kappa	46.83	56.67	55.88	42.45

6.4 Sense Distinctions

Table 16 shows that of the WordNet candidates, nearly all of them were associated with sense 1. This makes intuitive sense since in general the polysemy of MWEs is fairly modest, and there are relatively few possible senses and in many cases only one. Given this it is interesting and indeed somewhat surprising that relatively few of the MWE candidates from WordNet were judged to be viable.

7 Conclusions

NLP is a growing contributor to educational applications. In LM, a text modification application designed for teachers, several NLP methods are used, including automated summarization, machine translation, and syntactic parsers and synonym detection for the purpose of highlighting linguistic structure, and specific vocabulary, as part of the larger goal of providing linguistic analysis of classroom texts for teachers. This study was motivated by an interest to improve the current synonym detection capability (modified Lin matrices) that are currently use in LM. This interest was motivated by user feedback. In light of this, there were two application-specific research

Table 16: MWE : WordNet by sense

Fine (1,776)	# 1 (1,039)	# 2 (240)	# 3 (121)	other (376)
syn.	170	18	3	9
no syn.	644	187	111	347
FGA	78%	85%	94%	95%
Kappa	46.03	44.89	43.11	44.60

questions that sparked this study. The first asks if the manually-created lexical database, WordNet, could serve as a source of synonym candidates for single words that would complement a modified thesaurus created automatically using Lin's (1998) distributional approach. The second asks if WordNet and/or Lin provide a reasonable source of synonym candidates for MWEs. The answer to the first question seems to be clearly "yes". WordNet provides additional synonyms not discovered by the distributional method that are appropriate to use in context or to explain the underlying meaning of a word. The answer to the second question is not as clear. In general, there was a high level of agreement that most of the synonym candidates for the MWE target words were not viable, and the number of acceptable candidate was quite small. In terms of broader implications for the educational applications, moving forward with this work, we envision completing more educationally-guided annotation tasks that might require annotators to select synonym candidates that are acceptable only for certain proficiency levels of NNES learners, and with regard to the level and the goal of basic vocabulary comprehension or vocabulary development. Such annotation could support corpus-building or enhancements for NNES learners that can be used in educational applications. More broadly, knowledge of this complementary relationship between WordNet and a distributional method could inform any applications that require synonym detection.

Acknowledgements

This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute.

References

[AAH+09] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[BM01] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France, July 2001. Association for Computational Linguistics.

[Boo04] C. Boonthum. istart: Paraphrase recognition. In Daniel Midgley Leonoor van der Beek, Dmitriy Genzel, editor, *ACL 2004: Student Research Workshop*, pages 31–36, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[CBKO06] C. Callison-Burch, P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June 2006. Association for Computational Linguistics.

[CFH08] P. Clark, C. Fellbaum, and J. Hobbs. Using and extending WordNet to support question-answering. In *Proceedings of the Fourth Global WordNet Conference*, pages 111–119, University of Szeged, Hungary, 2008.

[CGK05] H. Calvo, A. Gelbukh, and A. Kilgarriff. Distributional thesaurus vs. WordNet: A comparison of backoff techniques for unsupervised pp attachment. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 177–188, Mexico City, February 2005.

[CHT+07] P. Clark, P. Harrison, J. Thompson, W. Murray, J. Hobbs, and C. Fellbaum. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 54–59, Prague, June 2007. Association for Computational Linguistics.

[CMR07] M Calderón and L. Minaya-Rowe. *Teaching reading, oral language and content to English language learners – How ELLs keep pace with mainstream students.* Corwin Press, Thousand Oaks, CA, 2007.

[Coh60] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[DW09] G. Dinu and R. Wang. Inference rules and their application to recognizing textual entailment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 211–219, Athens, Greece, March 2009. Association for Computational Linguistics.

[Fel98] C. Fellbaum, editor. *WordNet: An electronic lexical database.* MIT Press, 1998.

[FKL98] P. Foltz, W. Kinsch, and T Landauer. Analysis of text coherence using Latent Semantic Analysis. *Discourse Processes*, 25:285–307, 1998.

[KMB+05] B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, Ö. Uzuner, and A. Wilcox. External knowledge sources for question answering. In *Proceedings of the 14th Annual Text Retrieval Conference (TREC)*, Gaithersburg, MD, November 2005.

[LC03] C. Leacock and M. Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405, 2003.

[LH10] A. Louis and D. Higgins. Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95, Los Angeles, California, June 2010. Association for Computational Linguistics.

[Lin98] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 768–774, Montreal, 1998.

[LLF00] T. Landauer, D. Laham, and P. Foltz. The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 15(5):285–307, 2000.

[LP02] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 577–583, Taipei, Taiwan, 2002.

[Mil95] G.A. Miller. WordNet: A lexical database. *Communications of the ACM*, 38(11):39–41, November 1995.

[MKW04] D. McCarthy, R. Koeling, and J. Weeds. Ranking WordNet senses automatically. Technical Report CSRP 569, University of Sussex, January 2004.

[SS09] J. Sukkarieh and S. Stoyanchev. Automating model building in c-rater. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 61–69, Suntec, Singapore, August 2009. Association for Computational Linguistics.