# Using Lexical Semantic Techniques to Classify Free-Responses

Jill Burstein
Educational Testing Service - 11R
Princeton, New Jersey 08541
jburstein@ets.org

Randy Kaplan
Educational Testing Service - 17R
Princeton, New Jersey 08541
rkaplan@ets.org

Susanne Wolff
Educational Testing Service - 17R
Princeton, New Jersey 08541
swolff@ets.org

Chi Lu
Educational Testing Service - 17R
Princeton, New Jersey 08541
clu@ets.org

## Abstract

This paper discusses a case study in which lexical semantic techniques were used to implement a prototype scoring system for short-answer, free-responses to test questions. Scoring, as it is discussed in this paper, is a kind of classification problem. Responses are automatically scored by being assigned appropriate classifications. The ultimate goal is to develop a scoring system which can reliably analyze response content.

For this study, a domain-specific, concept-based lexicon, and a concept grammar were built to represent the response set, using 200 of 378 responses from the original data set. The lexicon is built, from individual words, and 2-word and 3-word terms from the training data. The lexicon is best characterized by Bergler's (1995) layered lexicon. Concept grammar rules are built by mapping concepts from the lexicon onto the concept-structure patterns present in a set of training responses. Previous attempts to score these responses using lexically-based statistical techniques and structure-independent content grammars were not reliable (Burstein and Kaplan (1995)). The results discussed in this paper illustrate the reliability of the lexical semantic methods used in the study.

## 1. Introduction

There is a movement in testing to augment the conventional multiple-choice items (i.e., test questions) with short-answer free-response items. Due to the large volume of tests administered yearly by Educational Testing Service (ETS), hand-scoring of these tests with these types of items is costly and time-consuming for practical testing programs. ETS is currently working on natural language understanding systems which could be used for computer-assisted scoring of short-answer free-responses (see Kaplan and Bennett (1994) and Burstein and Kaplan (1995)).[1]

The overall goal of our current research is to develop a scoring system that can handle short-answer free-response items. Such a scoring system has to be able to identify the relevant content of a response and assign it to an appropriate content category. Another consideration in the development of a scoring system is that the data sets that are available to us are relatively small. and the responses in these data sets lack lexico-syntactic patterning. The items which we work with are either experimental, or have been administered as paper-and-pencil exams. In the former case, there is a limited subject pool, and in the latter case, we rely on what has been put into electronic form. The response sets typically range from 300-700 responses which we have to use for training and testing. This is quite a different scenario from natural language understanding systems which can be designed using large corpora from full text sources, such as the AP News and the Wall Street Journal. This paper discusses a case study that examined how lexical semantic techniques could be used to build scoring systems, based on small data sets. Previous attempts to classify these responses using lexically-based statistical techniques and structure-independent content grammars were not reliable (Burstein and Kaplan (1995)). The results of this case study illustrate the reliability of lexical semantic methods.

For this study, a concept-based lexicon and a concept grammar were built to represent a response set. The lexicon can best be characterized by Bergler's (1995) layered lexicon in that the list of lexical entry words and terms can remain constant, while the features associated with each entry are modular, so that they can be replaced as necessary. Concepts in the concept grammars were linked to the lexicon. In this paper, concepts are superordinate terms which contain one or more subordinate, metonymic terms. A prototype was implemented to test our hypothesis that a lexical semantics approach to scoring would yield accurate results.

## 2. Test Item Types, Response Sets, and Lexical Semantics

### 2.1 Test Item Types and Response Sets

Our previous research with regard to language use in test items revealed that different test items use domain-specific language (Kaplan and Bennett (1994)). Lexicons restricted to dictionary knowledge of words are not sufficient for interpreting the meaning of responses for unique items. Concept knowledge bases built from an individual data set of examinee responses can be useful for representing domain-specific language. To illustrate the use of such knowledge bases in the

[1]In this paper, a response refers to an examinees 15 - 20 word answer to an item which can be either in the form of a complete sentence or sentence fragment.

development of scoring systems, linguistic information from the response set of an inferencing item will be discussed. For this item type, examinees are reliant on real-world knowledge with regard to item topic, and responses are based on an examinees own ability to draw inferences.

Responses do not appear show typical features of sublanguage in that there are no domain-specific structures, and the vocabulary is not as restricted. Therefore, sublanguage techniques such as Sager (1981) and Smadja (1993) do not work. In situations where lexico-syntactic patterning is deficient, a lexicon with specified metonymic relations can be developed to yield accurate scoring of response content. We define metonyms as words which can be used in place of one another when they have a domain-specific relation (Gerstl (1991))

### 2.2 Using Lexical Semantics for Response Representation

Our goal in building a scoring system for free-responses is to be able to classify individual responses by content, as well as to determine when responses have duplicate meaning (i.e., one response is the paraphrase of another response). In previous research, we used a concept-based approach similar to the one described in this study. The difference between the previous system and our current prototype is that in the previous system, concepts were not represented with regard to structure, and the lexicon was domain-independent. The underspecification of concept-structure relationships, and the lack of a domain-specific lexicon degraded the performance of that system (Kaplan and Bennett (1994). A second lexically-based, statistical approach performed poorly for the same reasons described above. The second approach looked at similarity measures between responses based on lexical overlap. Again, structure was not considered, and the lexicon was domain-independent which contributed to the system's poor performance (Burstein and Kaplan (1995)).

Any system we build must have the ability to analyze the concept-structure patterning in a response, so that response content can be recognized for scoring purposes. Given our small data set, our assumption was that a lexical semantic approach which employed domain-specific language and concept grammars with concept-structure patterns would facilitate reliable scoring. Our hypothesis is that this type of representation would denote the content of a response based on its lexical meanings and their relationship to the syntactic structure of the response.

It would appear that Jackendoff's (1983) Lexical Conceptual Structure (LCS) representation may be applicable to our problem. These structures are considered to be conceptual universals and have been successfully used by Dorr, et al (1995) and Holland (1994) in natural language understanding tasks. Holland points out, however, that LCSs cannot represent domain knowledge, nor can they handle the interpretation of negation and quantification, all of which are necessary in our scoring systems. Holland also states that LCSs could not represent a near-match between the two sentences, *The person bought a vehicle*, and *The man bought a car*. As is discussed later in the paper, our scoring systems must be able to deal with such near-match responses. Based on the above-mentioned limitations of LCSs, the use of such representation for scoring systems does not seem compatible with our response classification problem.

## 3. The Formulating-Hypotheses Item[2]

Responses from the *Formulating-Hypotheses* item (F-H) were used in this study. F-H is an experimental inferencing item in which an examinee is presented with a short passage (about 30 words) in which a hypothetical situation is described, and s/he composes up to 15 hypotheses that could explain why the situation exists. Examinee responses do not have to be in complete sentences, and can be up to 15 words in length. For example, an item referred to as the *police item* describes a situation in which *the number of police being killed has reduced over a 20-year period*. The examinee is the asked to give reasons as to why this might have occurred. Sample responses are illustrated in (1).

(1) Sample correct responses to the police item

a. Better cadet training programs
b. Police wear bullet-proof vests
c. Better economic circumstances mean less crime.
d. Advanced medical technology has made it possible to save more lives.
e. Crooks now have a decreased ability to purchase guns.

### 3.1 Required Scoring Tasks for F-H

Our task is to create a system which will score the data using the same criteria used in hand-scoring. In the hand-scoring process, test developers (i.e., the individuals who create and score exams) create a multiple-category rubric, that is, a scoring key, in which each category is associated with a set of correct or incorrect responses. A multiple-category rubric must be created to capture any possible response duplication that could occur in the examinees multiple response file. For instance, if an examinee had two responses, *Better trained police*, and *Cops are more highly trained*, the scoring system must identify these two responses as duplicates which should not both count toward the final score. Another reason for multiple-category assignment is to be able to provide content-relevant explanations as to why a response was scored a certain way. Our current prototype was designed to classify responses according to a set of training responses which had been hand-scored by test developers in a multiple-category rubric they had developed. For the police data set, there were 47 categories associated with a set of 200 training responses. Each rubric category had between 1 and 10 responses.

### 3.2. Characterization of police training data

The training set responses have insufficient lexico-syntactic overlap to rely on lexical co-occurrence and frequencies to yield content information. For instance, *police* and *better* occur frequently, but in varying structures, such as in the responses, *Police officers were better trained*, and *Police receiving better training to avoid getting killed in the line of duty*. These two

responses must be classified in separate categories: (a) *Better police training, general*, and (b) *Types of self-defense/safety techniques*, respectively.

Metonyms within content categories had to be manually classified, since such relations were often not derivable from real-world knowledge bases. For instance, in the training responses, *A recent push in safety training has paid off for modern day police*, and "*Officers now better combat trained...*, " the terms *safety training* with *combat trained*, needed to be related. Test developers had categorized both responses under the *Trained for self-defense/safety category*. *Safety training* and *combat train* were terms related to a type of training with regard to personal safety. The terms had to be identified as metonyms in order to classify the responses accurately.

## 4. Strategy for Representing Police Responses

As previously mentioned, there was insufficient lexico-syntactic patterning to use a contextual word use method, and domain-specific word use could not be derived from real-world knowledge sources. Therefore, we developed a domain-specific concept lexicon based on a set of 200 training responses over all categories. Each single, relevant word or 2-3 word term was linked to a concept entry. Small concept grammars were developed for individual rubric categories. These grammars were based on the conceptual-structural representations identified in the training response set.

As much as possible, it was important that the rules represented the relationship between multiple concepts within a phrasal constituent. The phrasal constituent itself, that is, whether it was an NP or a VP did not seem relevant. It was only meaningful that a constituent relationship occurred. Without this structural information, the concepts could occur in any position in a response, and automatic category assignment would not be reliable (Burstein and Kaplan (1995)). The procedure used to identify conceptual and syntactic information, retrieves concepts within specific phrasal and clausal categories. Once a response was processed, and concept tags were assigned, all phrasal and clausal categories were collapsed into a general phrasal category, XP, for the scoring process, as illustrated in (4), below. There were some cases, however, where we had no choice but to include some single concepts, due to the limited lexico-syntactic patterning in the data.

### 4.1. The Scoring Lexicon for the Police Item

What we term the *scoring lexicon* can best be illustrated by Bergler's (1995) *layered lexicon*. The underlying idea in Bergler's approach is that the lexicon has several layers which are modular, and new layers can be plugged in for different texts. In this way, lexical entries can be linked appropriately to text-specific information. In the layered lexicon approach, words are linked to definitions within some hierarchy. Bergler's approach also has a meta-lexical layer which maps from syntactic patterns to semantic interpretation that does not affect the lexicon itself. By comparison, our scoring lexicon, contains a list of base word forms (i.e., concepts).[3] The definitions associated with these concepts were typically metonyms that were specific to the domain of the item. These metonym definitions were subordinate to the words they defined. In

[3]Suffixation was removed so that part of speech did not interfere with conceptual generalizability.

the spirit of the layered lexicon, the definitions associated with the superordinate concepts are modular, and can be changed given new domains.

For this study, metonyms for each concept were chosen from the entire set of single words over the whole training set, and specialized 2-word and 3-word terms (i.e., domain-specific and domain-independent idioms) which were found in the training data. The lexicon developed for this study was based on the training data from all rubric categories. In (2), below, a sample from the lexicon is given. Our concept grammars, described in Section 4.2, are in the spirit of Bergler's notion of a meta-lexical layer that provides a mapping between the syntax and semantics of individual responses.

In our lexicon, concepts are preceded by #. Metonyms follow the concepts in a list. Lexical entries not preceded by # are relevant words from the set of training responses, which are metonyms of concepts. These entries will contain a pointer to a concept, indicated by '% <*concept*>'. A sample of the lexicon is illustrated below.

(2) Sample from the Police Item Lexicon

#BETTER [ better good advance improve increase ...
efficient modern well increase ]
ADVANCE [ %better ]

### 4.2 Concept Grammar Rules for the Police Item

The concept grammar rule templates for mapping and classifying responses were built from the 172 training set responses in 32 categories.[4] The training data was parsed using the parser in Microsoft's Natural Language Processing Tool (see MS-NLP(1996) for a description of this tool). For this study, suffixes were removed by hand from the parsed data. Based on the syntactic parses of these responses and the lexicon, a small concept grammar was manually built for each category which characterized responses by concepts and relevant structural information. The phrasal constituents were unspecified. Sample concept grammar rules are illustrated in (3).

(3) Sample Concept Grammar Rules for *Types of self-defense/safety*

a. XP:[POLICE],XP:[BETTER,TRAIN],XP:[SAFETY]
b. XP:[TRAIN ],XP:[POLICE,SAFETY],
XP:[BETTER,SAFETY]
c. XP:[POLICE,BETTER,TRAIN],
XP:[SAFETY,DANGER,SITUATION]
d. XP:[SPECIALIST],XP:[TRAIN SAFETY]

### 4.3 Processing Responses for Category Assignment

[4]Some categories were not considered in this study due to insufficient data.

Responses were parsed, and then input into the phrasal node extraction program. The program extracted words and terms in Noun Phrases (NP), Verb Phrases (VP), Prepositional Phrases (PP), Infinitive Clauses (INFCL), Subordinate Clauses (SUBCL), Adjective Phrases (ADJP) and Adverb Phrases (ADVP). All phrasal and clausal constituent nodes were then collapsed into a generalized representation, XP. All single XPs and combinations of XPs were matched against the concept grammars for each content category to locate rule matches. This procedure is illustrated below.

(4)
a. Input:
Cops are better trained in self-defense

b. Tag Phrasal Nodes of Parsed Response:
[Cops=POLICE]NP
[better=BETTER,trained=TRAIN]VP
[self-defense=SAFETY]PP

c. Collapse Phrasal Nodes:
XP:[Cops=POLICE]
XP:[better=BETTER,trained=TRAIN]
XP:[self-defense=SAFETY]

d. Match Tagged Nodes to Concept Grammar Rules:
XP: [POLICE], XP:[BETTER,TRAIN],XP:[SAFETY]

### 4.4 Does Manual Preprocessing of the Data Outweigh the Benefits of Automated Scoring?

Since the preprocessing of this response data is done by hand, the total person-time must be considered in relation to how long it would take test developers to hand score a data set in a real-world application. We must address the issue of whether or not a computer-based method would be efficient with regard to time and cost of scoring.

In this study, the manual creation of the lexicon and the concept grammar rules for this data set took two people approximately one week, or 40 hours. Currently, we are developing a program to automate the generation of the concept grammars. We expect that once this program is in place, our preprocessing time will be cut in half. So, we estimate that it would take one person approximately 8 -10 hours to create the lexicon, and another 8 - 10 hours to do the preprocessing and post-processing required in conjunction with the automatic rule generation process currently being developed.

The F-H item is currently only a pilot item for the Graduate Record Examination (GRE), which administers approximately 28,000 examinees, yearly. For the F-H item, each examinee can give up to 15 responses. So, the maximum number of responses for this item over the year would be approximately 420,000. Each examinee's response set would then typically be scored by two human graders. It is difficult to estimate how long the manual scoring process would take in hours, but, presumably, it would take longer than the approximately 40 hours it took to build the lexicon and

concept grammars. Certainly, it would take longer than the 20 hours estimated, once the automatic rule generator is implemented. Therefore, assuming that the accuracy of this method could be improved satisfactorily, automated scoring would appear to be a viable cost-saving and time-saving option.

### 5.1 Initial Results

One hundred and seventy-two responses were used for training. These responses were used to build the lexicon and the concept grammar rules. An additional, independent set of 206 test responses from 32 content categories was run through our prototype. The following were the results.

**Table 1: Results of Automatic Scoring of Responses**

| Response Set | Coverage | Accuracy |
|---|---|---|
| **Total Set of Responses (Training Set + Test Set)** | **92% (347/378)** | **90% (313/347)** |
| **Test Set Only** | **87% (180/206)** | **81% (146/180)** |

### 5.2 Error Accountability

Most of the errors made in classifying the data can be accounted for by four error types: (a) *lexical gap*, (b) *human grader misclassification*, (c) *concept-structure problem*, (d) *cross-classification*. The lexical gap error characterizes cases in which a response could not be classified because it was missing a concept tag, and, therefore, did not match a rule in the grammar. In reviewing the lexical gap errors, we found that the words not recognized by the system were metonyms that did not exist in the training, and were not identified as synonyms in any of our available thesaurus or on-line dictionary sources. For instance, in the response, "*Police are better skilled...,"* the phrase *better skilled*, should be equated to *better trained*, but this could not be done based on the training responses, or dictionary sources. Forty percent of the errors were lexical gap errors. The second problem was human grader misclassification which accounted for 1 percent of the errors. In these cases, it was clear that responses had been inadvertently misclassified, so the system either misclassified the response, also. For example, the response, *Officers are better trained and more experienced so they can avoid dangerous situations*, was misclassified in *Better trained police, general*. It is almost identical to most of the responses in the category *Better intervention/crook counseling*. Our system, therefore, classified the response in *Better intervention/crook counseling*. Concept-structure problems made up 30 percent of the errors. These were cases in which a response could not be classified because its concept-structural patterning was different from all the concept grammar rules for all content categories. The fourth error type accounted for 17 percent of the cases in which there was significant conceptual similarity between two categories, such that categorial cross-classification occurred.

### 5.3 Additional Results Using an Augmented Lexicon

As discussed above, 40 percent of the errors could be accounted for by lexical gaps. We hypothesized that our results would improve if more metonyms of existing concepts were added to the lexicon. Therefore, we augmented the lexicon with metonyms that could be accessed from the test data. We reran the scoring program, using the augmented lexicon on the same set of data. The results of this run were the following.

**Table 2: Results from Automatic Scoring Using an Augmented Lexicon**

| Response Set | Coverage | Accuracy |
|---|---|---|
| **Total Set of Responses (Training Set + Test Set)** | **96% (364/378)** | **96% (341/364)** |
| **Test Set Only** | **93% (193/206)** | **93% (178/193)** |

The improvement which occurred by augmenting the lexicon further supports our procedure for classifying responses. Based on these results, we plan to explore ways to augment the lexicon without consulting the test set. Furthermore, we will use the augmented lexicon from this second experiment to score a set of 1200 new test data.[5]

### 6. Conclusion

Our results are encouraging and support the hypothesis that a lexical semantic approach can be usefully integrated into a system for scoring the free-response item described in this paper. Essentially, the results show that given a small set of data which is partitioned into several meaning classifications, core meaning can be identified by concept-structure patterns. It is crucial that a domain-specific lexicon is created to represent the concepts in the response set. Therefore, the concepts in the lexicon must denote metonyms which can be derived from the training set. Relevant synonyms of the metonyms can be added to expand the lexicon using dictionary and thesaurus sources. Using a layered lexicon approach (Bergler (1995)) allows the words in the lexicon to be maintained, while the part of the entry denoting domain-specific meaning is modular and can be replaced. The results of this case study illustrate that it is necessary to analyze content of responses based on the mapping between domain-specific concepts and the syntactic structure of a response. As mentioned earlier in the paper, previous systems did not score responses accurately due to an inability to reliably capture response paraphrases. These systems did not use structure or domain-specific lexicons in trying to analyze response content. The results show that the largest number of erroneous classifications occurred due to lexical gaps. Our second set of results shows that developing new methods to augment the lexicon would

[5]We did not use these 1200 test data in the initial study, since the set of 1200 has not been scored by test developers, so we could not measure agreement with regard to human scoring decisions. However, we believe that by using the augmented lexicon, and our concept grammars to automatically score the 1200 independent data, we can get a reasonable idea of how well our method will generalize, based on our assessment of the scoring decisions made by the program.

improve performance significantly. In future experiments, we plan to score an independent set of response data from the same item, using the augmented lexicon, to test the generalizability of our prototype. We realize that the results presented in this case study represent a relatively small data set. These results are encouraging, however, with regard to using a lexical semantics approach for automatic content identification on small data sets.

**References**


Bergler, Sabine. (1995). From Lexical Semantics to Text Analysis. In Patrick Saint-Dizier and Evelyne Viegas (eds.), *Computational Lexical Semantics*, Cambridge University Press,New York, NY.

Burstein, Jill C. and Randy M. Kaplan. (1995). On the Application of Context to Natural Language Processing Applied to the Analysis of Test Responses. *Proceedings from the Workshop on Context in Natural Language Processing*, IJCAI, Montreal, Canada.

Dorr, Bonnie, James Hendler, Scott Blanksteen and Bonnie Migdoloff. (1995*). On Beyond Syntax: Use of Lexical Conceptual Structure for Intelligent Tutoring*. In V. Melissa Holland, Jonathan Kaplan and Michelle Sams (Eds), Intelligent Language Tutors, Lawrence Erlbaum Publishers, Mahwah, NJ.

Gerstl, P. (1991). A Model for the Interaction of Lexical and Non-Lexical Knowledge in the Determination of Word Meaning. In J. Pustejovsky and S. Bergler (Eds), *Lexical Semantics and Knowledge Representation*, Springer-Verlag, New York, NY.

Holland, V. Melissa. (1994). Intelligent Tutors for Foreign Languages: How Parsers and Lexical Semantics Can Help Learners and Assess Learning. In Randy M. Kaplan and Jill Burstein (Eds), *Proceedings of the Educational Testing Service Conference on Natural Language Processing and Technology in Assessment and Education*, Educational Testing Service, Princeton, NJ.

Jackendoff, R.S. (1993). *Semantics and Cognition*. MIT Press, Cambridge, MA.

Kaplan, Randy M. and Randy E. Bennett. (1994). Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypothesis Item. (RR-94-08). Princeton, NJ:Educational Testing Service.

MS-NLP. (1996). http://research.microsoft.com/research/nlp. Microsoft Corporation. Redmond, WA.

Sager, N. (1981). *Natural Language Information Processing: A computer grammar of English and its applications*, Addison-Wesley, Reading, MA.

Smadja, Frank. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*. 19(1), 143-177.