Is there a Greater Analytic Potential for Open-ended Survey Questions? A Comparison of Analytic Strategies

Casey Langer Tesfaye
American Institute of Physics
Georgetown University

Abstract:

Because the time and expense involved in traditional qualitative coding methods are increasingly prohibitive, text analysis techniques are increasingly seen as as an attractive alternative for qualitative analysis. This paper provides an exploratory comparison of traditional qualitative coding with a series of text analysis and natural language processing strategies. I compare the depth and usability of the findings from different strategies, and some advantages and disadvantages of each. By integrating methods and knowledge from the fields of linguistics and natural language processing, survey methodologists can gain a better understanding of the patterns inherent in textual data. By taking advantage of these patterns, we gain the potential to analyze answers from open ended survey questions more efficiently, the ability to unpack more meaning from textual data, and the potential to use naturalistic data, such as social media sources or conversational transcripts, to complement other research findings. This analysis is based on the open ended responses from the 2008 Nationwide Survey of High School Physics Teachers.

Keywords: content analysis, corpus linguistics, natural language processing, open-ended questions, sociolinguistics, survey methodology, word clouds, word frequencies

Introduction:

One of the essential assumptions behind survey research is that one response cannot represent a whole population. Closed questions work well under this assumption, because they aggregate easily. They are relatively easy and inexpensive to analyze, and we are usually able to generalize their results to the sample frame and subsequently the wider population with some degree of certainty. In contrast, open ended questions are significantly more difficult to analyze, more expensive and time consuming and markedly more difficult to generalize. Despite these complications, open ended questions are an important aspect of survey research.

In the absence of the cues that response domains provide, respondents orient toward open ended questions in a variety of ways, both in terms of the response domains they choose and their orientation within those domains. To add to this variation, it is very unlikely that the word choice of any substantive response will duplicate another, meaning that it is quite unlikely that any two responses will be the same. As a result of this variation, open ended questions are

more exploratory in nature. They are more of a window into what respondents may be thinking about a given topic than a quantifiable reflection of their views. They are ideal for occasions when it's not clear how responses to a question will naturally group. It is this exploratory quality that both makes these questions essential and makes them so difficult to analyze and report. Some quantitative analysts respond to the variation inherent in open ended responses by characterizing all open ended analysis as deeply subject to judgment. Some would prefer to replace open ended questions entirely by conducting more research, like focus groups or cognitive testing, earlier in the survey development process and then using those findings to construct better closed questions with valid response domains±. Traditional qualitative analytic methodology accounts for the variation in responses through inter-rater reliability. Responses are coded by more than one coder in tandem, using a set of codes that are developed by both coders emergently as they proceed through the dataset. Instances where the coders don't initially agree on the codes are discussed, or adjudicated, until consensus is reached. The amount of staff time needed to code open ended responses is often prohibitive. Although open ended questions are commonly used, the majority of open ended responses are never analyzed in any detail.

Open ended responses are not just difficult to code; they are also particularly difficult to report. There are two main reasons for this. One: we can't make any assumptions about the responses that we don't observe. So if one respondent doesn't answer the question at all, we can make no assumptions about their underlying views. Moreover, if one respondent responds about aspect A of the question and another responds about aspect B of the same question, we can't make any assumptions about the underlying views of the first respondent toward aspect B or the second respondent toward aspect A. This makes discussions about findings tricky. Although the responses are aggregated into codes, the number of respondents falling into a given code is not necessarily meaningful in any contrastive way. Should we discuss the frequency of a code in the context of all of the respondents who answered a question, or just the respondents who gave substantive responses (isolating these can be difficult), or just the respondents who gave substantive responses about aspect A of the question?

The second reason why reporting responses to open ended questions is so difficult is because of the inherent relational power of narrative. When we choose individual responses to report, we choose them for their eloquence. These eloquent responses often have a small narrative, or illustrative storytelling component that makes them particularly convincing. There is a natural tendency for people to treat a single eloquent response as representative of all responses. I've seen readers of our reports put a greater weight on a single comment than on the numeric findings that accompany it. Given a situation where 95% of respondents oppose something, it is not uncommon for a single quote from one of the 5% who support it to be better remembered and recalled more often than the 95:5 split. Because of this, open ended responses can feel like a minefield to report. This can only add to our reluctance to analyze these results.

However, the technical environment surrounding text analysis is already beginning to change the dynamics surrounding open ended questions and inspire a new hope of finding useful, cost efficient ways of using open ended responses. Word clouds are fast gaining popularity in survey reports. Survey analysis software packages are beginning to include, develop, and heavily market text analysis extensions, and qualitative coding software packages are developing and integrating more text analytic abilities. The results of these approaches look dramatically different from traditional results, and they introduce a new set of advantages and disadvantages. These programs are responding to the demands of the field to gain a window into the content of the responses without requiring as much staff time. They take advantage of a massive and growing field of research called Natural Language Processing (NLP), which is the intersection between linguistics and computer science.

Some of the difficulties in extracting semantic information from text that programmers in the field of Natural Language Processing face mirror the challenges that we face in traditional qualitative analysis. This is no surprise to linguists, whose work focuses on the ways in which language is and is not patterned. When we, as people who use our language skills every day, think about language, we naturally think of the topicality or aboutness of our communication, or 'what are we talking about?' But let's take a step back and think about conversation. Imagine that you and I are seated together in a hotel lobby during a conference, and one of us asks the other, "So what do you think of the conference so far?" Experience tells us that we can't predict the response to this question, and we would be even less successful at predicting the conversation that would develop after that prompt. And yet we, as survey researchers, often ask similar questions of our respondents and expect the answers to fall into patterns.

Students of linguistics quickly learn that language is not patterned on or organized by its aboutness. It is, however, thoroughly patterned. It is patterned in the way that words are structured, or formed and reformed using common roots (e.g. "dog" "dogged" "doggedly," "fast" "faster" "fastest," and "is" "was" "will be"). It is structured by the parts of speech of the words we use, and the order in which words are arranged into phrases and phrases are arranged into sentences. We know that "Man bites dog" doesn't equate to "Dog bites man" because of the structural order of subjects, verbs and objects. We know that the verb "run" functions differently than bite (e.g. "Dog runs man") because run, which doesn't work with an object is a different type of verb than bite, which usually does have an object. We know that "teacher" is related to "instructor," but that relationship is different than the relationship between "teacher" and "school" or "instructor" and "student." Language is patterned on grammatical rules that we rarely consider, and language is also patterned in more conscious ways. When we communicate we make many strategic decisions with varying levels of consciousness. These decisions include the tense of the verbs we choose to use, whether or how we choose to represent our opinions or emotions, whether or not we wish to clarify

our degree of certainty about a topic, whether to repeat or reframe the original words used to introduce a topic, and more. All of these patterns are particularly useful in the unpacking of words, and each of these areas is a thriving field of study. The following are a few of the many domains that hold potential for the analysis of open ended survey questions.

Fact vs. Opinion

There has been a considerable amount of research devoted to the disaggregation of factual vs opinionated language in text analysis. For a simple example, opinions are often marked by expressions about their level of certainty (e.g. "I think" or "I'm pretty sure") or by their source (e.g. "People say that…" or "I suspect that…") (Chafe and Nichols, 1986). These types of patterns, and others, can be isolated by programmers in order to isolate topics from commentary about topics. Pang and Lee (2008) provide an excellent summary of some of this research, which has been greatly successful to date.

Sentiments or Emotions

There is also a great deal of research devoted to detecting the presence of an affective or emotional orientation in text. Within the field of Natural Language Processing, this area is known as Sentiment Analysis. Sentiment Analysis has already begun to permeate the field of Public Opinion Research. Some automated programs can separate out topics and collective sentiment toward those topics from banks of open ended questions (e.g. Open Amplify, http://www.openamplify.com).

Specialized Vocabulary

The use or lack of use of specialized vocabulary in discourse can reflect the conscious rejection of the speaker toward specific terms, it could reflect a lack of familiarity with that terminology, or it could reflect subtle differences in the topic being referenced. Any of these possibilities could prove analytically useful. An analysis of the use of specialized vocabulary could provide some insight into the dispersion or resonance of a key concept or method of communication.

Repetition

Repetition is a frequently used device in conversation (Tannen, 1989), and it can be used to establish a link between a response and its particular genesis. This could be useful, for example, when comparing the influence of various news services on a group of people. If people are consistently more likely to repeat phrases from twitter than from their nightly news program, there is a higher likelihood that they are orienting toward twitter over the nightly news as their preferred news source.

Ngram

Looking at common and rare combinations of words has proved a very effective strategy and a powerful tool for the analysis of textual data. Gary King and Will

Lowe (2003) were able to successfully analyze public opinion data using an event frequency strategy to look at word combinations.

Word Sense

Traditionally we think of meanings as a property of words rather than words as a property of meanings. In the field of NLP, words are organized into senses, or individual meanings, and indexed by these senses, unlike traditional dictionaries, which are organized by the words themselves. These senses are further indexed by their relationships to other words (e.g. "mother" to "grandma"|"grandmother.") Whereas an adherence to individual words can limit our ability to see conceptual patterning in our textual data because of the complicated relationships between words and their meanings, linking a dataset to a word sense dictionary like wordnet (http://wordnet.princeton.edu/) can give us the ability to better isolate conceptual patterning independent of specific word choice (Deerwester et al, 1990). Word sense is an important tool for addressing polysemy (words that have multiple meanings) and synonymy (meanings that have multiple words).

Agency vs Passivity

There is also quite a bit of research on the relationship between the placement of nouns in sentences and the underlying attitudes of the speaker. For example, in the previous sentence, I'm more focused on the existence of the body of research than the people behind it, and I expressed this bias by choosing passive voice. This can be a particularly interesting area of study when examining differences in the ways groups are discussed. An individual noun placement may not be independently meaningful, but this type of analysis strongly lends itself to patterning and can make a very strong case in the comparison of attitudes toward groups or individual entities. For example, series of textual examples including "dog bites man," "dog begs man for food," and "man feeds dog" can work to construct the different roles of men and dogs. This type of analysis is based on directly on syntax.

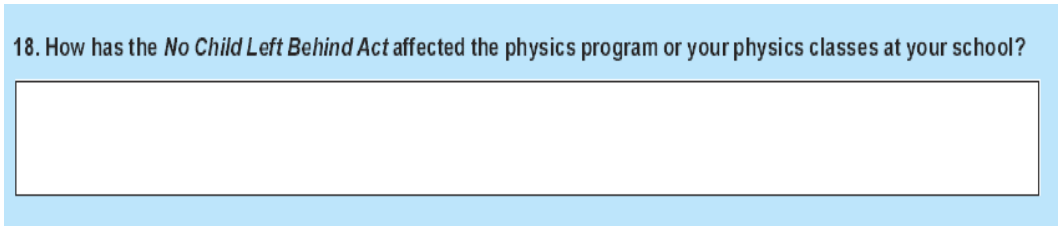
Temporality

Temporality is also a rich area of study in Natural Language Processing. When we speak we choose verbs according to the past, present, future, or ongoing nature of the actions they entail. "I already brushed my teeth" has a different temporal meaning than "I am brushing my teeth right now" or "tonight I will brush my teeth." Organizing text by the tense or by aspect of its verbs can show either the changing nature of something over time or changing attitudes toward something over time.

Methods:

In the following analysis, I will discuss a selection of strategies for the analysis of textual data from open ended survey responses, some of the specific advantages and disadvantages of those strategies, and the relative analytic promise of each. I will compare traditional qualitative methodology, frequency counts, word clouds, alternative frequency counts, program-based content

analysis, topic/content coding, and temporal coding using the responses to the following question from the 2008 National Survey of High School Physics Teachers about the No Child Left Behind Act (NCLB):



Traditional Qualitative Coding

The responses to this question were analyzed using a quasi-traditional qualitative methodology. Instead of multiple coders working in tandem, a single coder developed a coding schema in conjunction with other analysts who were familiar with the response set and then the coder proceeded to independently code the data over the course of a couple of months, consulting the analysts for advice when particularly difficult cases arose. The coder's results are not exactly representative of traditional methodology, but they are close. Below are the resulting codes and the frequencies of each. For this paper, I decided not to include any of the categories that garnered less than 29 responses, of which there were quite a few. It is important to note that the numbers of responses in these categories are misleading, because the numbers include responses from private school teachers who were not subject to the NCLB regulations. Unfortunately, because of the way the data was handled, we aren't able to separate those responses out. The total number of open ended responses from public school teachers is actually 623.

Code	**Number**
None/Little	722
Comment	196
NA / Don't know	163
Con	142
Other	136
Time - Instructional	97
Students - Unprepared	89
Time - Lost	84
Standards	70
Lower Level	59
Test - Scores	52
Paperwork	45
Money - Lack of	43
Test - General	40
Class Size - Small	38
Future	37
Students - Apathetic	36

Background	34
Teaching - Not	34
Money - Redirected	30
Students - Unprepared Math	29
Curriculum - Guided	29

At this point, these codes are probably the most detailed representation available of the content of the responses to this NCLB question. They show what a wide variety of dimensions respondents used to orient toward the question. Although the response frequencies start out robust, they quickly deteriorate. The responses include some potentially conflicting information, because about half of the responses said there was no effect and the other half described an effect of the policy. Although there are facts about the effects of the policy embedded in the answer set, they couldn't necessarily be used to fully explain the factual effects of the policy. And although there are some opinions in the response set, it's not clear that a summary of these opinions would be in any way representative of the group as a whole. Despite the large amount of time devoted to this coding, the usefulness of the analysis is debatable and reporting on it will be particularly difficult.

Word Frequencies

In contrast to traditional qualitative coding, word frequency counts hold a great deal of allure. There are over ten times as many words as responses within this response set (623: 6447), and word count software is plentiful, quick and remarkably easy to use. Currently no other tool can generate a faster analysis. However, a quick count of the word frequencies in the dataset demonstrates that word frequency is not very meaningful on its own.

Word	**Frequency**
the	240
to	200
not	198
physics	153
students	137
has	135
no	123
it	123
of	113
a	106

tool/source: wordcounts.com

Function words (like "the" "a" "and") are the most common words, but they don't independently carry much content. In a large enough body of words, the frequency distribution for all of the words, including the function words, will tend to follow a Zipfian distribution, in which the frequency of a given word is

inversely proportional to its frequency rank. Given this distribution, the actual frequency count of words diminishes rapidly enough that the words that we most want to target in word frequency counts are significantly less frequent than we would hope to see.

Fortunately, the majority of word frequency counters, including the one I used (wordcounts.com) will remove the function words from the frequency counts in order to more quickly access the words that carry more meaning. This is a subtle, but very important adjustment. Here is the same distribution, but without the function words. It is also adjusted by word stem (e.g. "Physics" and "Physical" are combined into one word with a common stem and listed as "physic*"). As a testament to the rapid turnaround of word counting websites, wordcounts.com was up for sale when I returned to repeat these frequencies without the word stem adjustments. There are still quite a few other sites that count word frequencies, with or without a variety of possible adjustments.

Word	**Frequency**
student	160
physic*	153
test	89
affect	62
take	61
science	54
effect	52
school	46
class	40
time	38

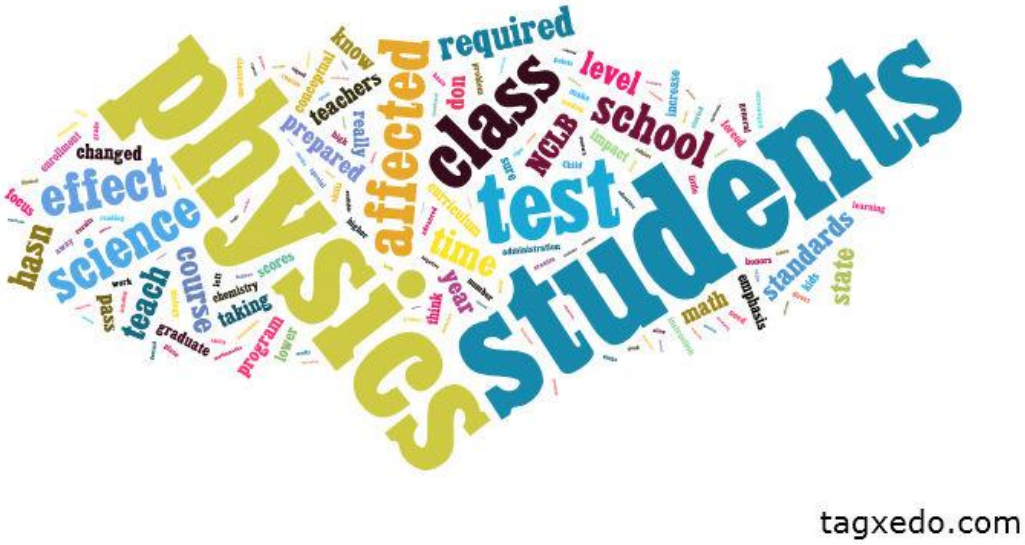
tool/source: wordcounts.com

* note: this apparent misspelling is due to an adjustment by word root

This frequency list unarguably carries more information than the one with the function words included, but it doesn't provide enough information to answer the original question. Word frequency lists can't independently function as a useful or informative analytic strategy.

Word Clouds

Another tactic for text analysis that is gaining popularity in survey research is the word cloud. Word clouds are as easy to generate as word frequency lists, and programs that generate them are abundant. Word clouds are particularly compelling because of both their cost benefit advantages and visual appeal. They are nearly as quick and easy to use as the word counters, but the visual component makes it easier to make quick frequency comparisons across words. Below is a word cloud using the responses to the NCLB question.



Although both the cloud and the word frequencies demonstrate that the respondents were indeed answering the question that was posed to them, the answer to the question cannot be summarized informatively by looking at the word cloud. Seeing, for example, that the respondents mentioned students or tests, doesn't give any information about how they oriented toward the topics. We don't know about the magnitude of the changes that resulted from the NCLB regulations or the teachers' feelings about the law. We don't know what modifiers originally surrounded the words (e.g. "not" "weren't," or "other"). For example, the word "affected" was prominent in the NCLB word cloud. On its own in the context of the word cloud, one could easily assume that the prominence of "affected" had an entirely different implication than a closer look at the word in the context of the data suggests. But "affected" has more than one potential meaning.

"I do not believe that NCLB has affected the physics program."

"I do not particularly know how it has affected me if at all."

Although these examples refer to different meanings of a single verb, word clouds are especially vulnerable to words that have multiple parts of speech, or specific metaphorical meanings.

Syntax-Independent Word Frequencies and Clouds

Pew Charitable Trusts has demonstrated a particularly interesting strategy for handling the grammatical complications that mitigate the effectiveness of the word clouds. They collect and aggregate single word descriptors of a topic. Here is an example from the 2008 campaign cycle (Pew, 2008):

"What ONE WORD best describes your impression of John McCain?"

Source: http://pewresearch.org/pubs/968/candidates-in-a-word

This question construction maps especially well onto word clouds. The result is a particularly fast and useful picture. This strategy not only works for word clouds, but also for word frequency counts, which, as can be seen below, can be used comparatively, because in this condition the words have no context to lose and can be taken at face value.

The One Word that Best Describes...

McCain		Obama	
#		#	
58	Old	55	Inexperienced
34	Patriot	36	Change
28	Bush/Bush-like	20	Intelligent
25	Experienced	20	Young/Youthful
21	Honest	15	Charismatic
18	Conservative	14	New
17	Hero	12	Energetic
16	Leader/Leadership	12	Hope/Hopeful
14	Strong	12	Liberal
11	Good	10	Honest
10	Integrity	9	Fresh
9	Maverick	9	Scary
9	Same	8	Different
9	Trust/Trustworthy	7	Enthusiastic
8	Honor/Honorable	7	Unqualified
8	Qualified	6	Committed
8	Republican	6	Good
7	Courageous	6	Innovative
7	Lies/Liar	6	Inspiring
6	Dedicated	6	Liar
		6	Socialist
N=629		N=629	

Based on registered voters. Figure shows number of respondents who offered each response; these numbers are not percentages.

Source: http://pewresearch.org/pubs/968/candidates-in-a-word

We could have accomplished results that were similarly usable by asking physics teachers for one word descriptions of NCLB, and that is indeed an attractive option for future versions of the Nationwide Survey of High School Physics Teachers. A word cloud based on this type of question would better reflect the respondents' answers. It would be a significantly easier analysis to conduct, and the results would be informative and meaningful. The Pew strategy is a solid one.

However, neither word clouds nor word frequency counts can account for lexical variation (e.g. "tests" or "standards") or variation in word meanings. This strategy can't account for synonymy or polysemy. Concepts with greater lexical variation are significantly less likely to be reported, even though they may be equally prevalent in the response set. It would certainly be possible to group words into categories and report the words, categories or both. But coding words adds to the analytic burden inherent in these questions. Alternatively, the words could be linked to wordnet (http://wordnet.princeton.edu/) and aggregated by word sense instead of word meaning. Word sense disambiguation would be both easier, because most of these words share the same part of speech (adjectives), and more difficult, because of the lack of context. However, in contrast to the ease and speed of the word cloud and word counting programs, word senses bring about a set of conceptual decisions that must be addressed in the process of the analysis.

This syntax-independent methodology provides useful and meaningful analysis at a quick turnaround for a minimal expense, which makes it an attractive and sensible option for firms with limited time or resources. However it can't provide any nuanced sense of respondent opinion. Single words can provide some useful and meaningful insight, but they can't provide the depth of information that a whole phrase, or even a series of sentences, could provide.

Automated or Semi-automated Content Analysis

Although I mentioned the large and changing field of software packages, the open source resources are far more powerful and flexible, and they take better advantage of the fields of knowledge in NLP and linguistics that are constantly evolving and growing. As a result, I will mostly discuss it in the context of programming methodology.

In a traditional qualitative analysis, coders devote a great deal of time and attention to their coding, developing coding guidelines as they work. All of this coding could be done by computer program, whether inefficiently on a case by case basis or more efficiently by writing the emergent coding guidelines into decision based programming modules. The prospect of writing coding decisions into a programming language often generates a lot of skepticism from survey researchers. This skepticism is probably based on both the assumption that the results of the programs would not be checked by human eye and an assumption that a human would make many corrections that a programmer would not make. But these assumptions are simply not true. A programmer could indeed make any adjustment that a coder could make. A program can also function as a detailed record of the decision making processes in the analysis and serve as a starting point for future analysis (especially if a question is repeated in another survey). Additionally, programming based strategies have the advantage of being able to take advantage of the work of the field of natural language processing, including wordnet (the word sense dictionary discussed earlier), open source part of speech taggers, word stemmers (like wordcounts.com used to tag different forms of a single wordstem as morphological combinations of a single root), and more. Programming based strategies can easily build on the past work of many programmers, increasing the benefits and decreasing the cost over time by saving future staff time and resources

However, even a programming based approach to content coding is complicated by the inherent variation in open ended responses. Responses to the NCLB question fell into many domains, including the following:

- Yes/no
- Standardized testing
- Enrollments
- Preparation of students

- Conceptual Physics
- Political Orientation
- Funding

And then within each of these domains, there is still a great deal of variation in the type of response. Here is an example of a series of responses that are quite similar to each other, and would traditionally be grouped together into a single code:

> "It has had little impact."
>
> "It has had no apparent impact."
>
> "It has had no effect."
>
> "It doen [sic] not affect in a positive or negative manner"
>
> "It has not affected our physics program."
>
> "It has not affected physics"
>
> "It has not affected the physics program. Enrollment is constantly low."

Although these responses are all quite similar, they are saying different things, commenting about different aspects of the policy. The traditional qualitative coder collapsed these responses and summarized them by saying that the majority of respondents reported that NCLB had little or no effect. A stored program is a nice place to keep a record of the set of responses within a certain code and a particularly fast and easy way of tweaking the coding structure as the project progresses. These responses likely need to be collapsed together, because in the context of the sheer volume and variation of responses to the NCLB question this variation is comparatively small. But later analytic decisions may call some of these choices into question, and a program is significantly easier to refer back to and to alter than a set of coded text.

Linguistics Based Strategies

Survey researchers are not trained to think deeply about units of analysis. Each respondent represents a single unit of analysis. But in linguistics, we are trained to carefully consider our units of analysis. Language is patterned in many ways, and each potential unit of analysis has a different set of patterns associated with it. We saw from the word frequency distributions that words are generally not independently informative units of analysis, and we see from the responses above that individual responses can hold complicated or even conflicting information. It is difficult to deal with this level of complication when processing

the responses as individual units of analysis when the meaning encoded in the responses is contained in the interaction between grammar and content. Responses can still be tagged, and responses can be mapped back to the respondent level in order to look at co-occurrences or take advantage of other information in the dataset, but we can also explore, or toggle between, other types of units of analysis, including:

Units of analysis:
Respondents
Sentence
Phrase
Grammatical pairs (e.g. noun, verb)
Common Combinations
Word sense
Nouns
Verbs
Word

Working with these smaller units of analysis provides a few important analytic advantages. First, it allows us to take advantage of the inherent systematicity of grammar and use the set of patterns around each type of unit to extract relevant information in a structured and comparative way. Second, it allows us to take advantage of a variety of open source NLP resources that were developed in order to take advantage of those patterns. For example, the Natural Language Toolkit, or NLTK (http://www.nltk.com), offers a wide variety of open source language tools through Python, a flexible, powerful open source programming language. In order to bypass morphological differences in words (e.g. “fast” “faster” “fastest,” or “is” “was” “will be”), NLTK offers a variety of stemmers. NLTK also offers part of speech tags that are detailed enough to isolate individual verb tenses, singular or plural nouns, or syntactic positions. If more or less detail is needed, more tags can be generated from the open source ones or custom tags can be developed. NLTK also offers tagsets that isolate common combinations of words, or ngrams, as well as wordnet, the word sense dictionary that includes relational hierarchies and accounts for synonymy and polysemy. There are other options for tags, including syntactic trees, specific or custom tagsets, and more.

To put these to use, let's consider a very different type of response:

> “More kids in my room are unprepared mathematically than I had expected, which may or may not be due to NO CHILD LEFT BEHIND. This is the main problem I face as a teacher, trying to teach unprepared kids a difficult subject that they really don't understand or see great value in.”

This response is quite similar to the others shown above. The teacher is unsure whether or not to attribute his or her challenges to the regulation, but took the opportunity to discuss his or her challenges nonetheless. The teacher may not be sure whether or not NCLB had affected his or her class directly, but the use of capital letters certainly seems to attribute some degree of agency. The teacher's response is not restricted to the immediate question, but it is informative. It is also significantly more memorable than the responses we just saw, because it adds a narrative dimension. This is the kind of quote that survey researchers love and hate: love because it's such a high quality depiction of the circumstance of the respondent, lending a palpable dimension that a numeric data point never could, and hate because of its memorability. Juxtaposed next to 3000 respondents, this response will likely be more memorable and likely carry more weight than any quantitative data point.

Although the responses that we've discussed so far would probably lead us to believe that NCLB had had a minimal effect on high school physics teachers, not all of the data agrees. In fact, a significant proportion of the data does detail significant effects of the regulation. These responses tend to be significantly longer and contain significantly more detail. Here are a few more:

> "Creates a serious problem. Cannot adequately split time to help all children as adequately as I would like."

> "Students are attempting to take a college preparatory class they are not adequetly [sic] prepared to take. Rigor is NOT something they are used to nor do they possess the motivation necessary for success. Ther [sic] is more emphasis on students scoring well on a state test rather than learning physics for the sake of learning."

> "Standardized state test caused change in class alignment. This school WAS going to a physics first direction until the standardized test came out with Biology-oriented questions. Push to give summative assessments with selected response to prepare students for standardized testing."

> "More students dumped into Physics that are unprepared so the content is much lower - almost a physical science class rather than a Physics class."

Coding responses like these is significantly more challenging than coding the other responses. At this point, we have to take a good look at our data and carefully consider our coding structure and analytic goals before deciding how to treat these responses.

Topic/Comment

These responses have quite a bit of nuanced information in them, but one possibility for looking closer at them is to focus on the nouns and present them with their modifiers or contexts, using a topic-comment format. Here is a quick look at these few responses using that strategy:

Topic	**Comment**
Students	not prepared
	not sufficiently motivated
	not used to rigor
	dumped into physics
Tests	changed class alignment
	garnered more emphasis
	pushed
	need to be prepared for

The results of this kind of strategy can appear similar to the results of traditional qualitative coding, but their basis is quite different. Instead of a coder choosing important elements from individual, nuanced responses, a set of nouns is isolated, and their context is displayed. In this way, the topic/comment format is more exhaustive. It is also potentially much less labor intensive than traditional coding, and the code developed during this process could easily be applied to other projects.

Temporality

Once we have isolated the parts of speech within the responses, we can separate out the responses by verb tense or aspect in order to develop a rough picture of the responses that were oriented toward the future implications of NLCB regulations, which were oriented toward current hardships, which were discussing their past experiences with NLCB, and which experiences were ongoing. We may find, for example, that hardships are more concentrated in the future, and ambiguity of more concentrated in the past. This would give us an idea of the temporal effects of NCLB on the respondents, and this would better reflect the staged implementation of the policy.

Here is an example of a temporal orientation in the NCLB responses:

Past:	“We have not seen much change"
Present:	“Not many students take physics because it's not required.”
Future:	“not at all yet--in two years all students will be required to take

	phys”

This is an especially useful tool for this set of responses, because it renders the roughly half of responses that previously appeared to be minimally informative significantly more useful. In fact, looking at this set of responses through a temporal lens shows that the ambiguity of the teachers is largely clustered in the past, the challenges that teachers face in the classroom are largely ongoing, and the focus about new regulations and mandated testing is largely clustered in the future.

The temporal analysis itself is not a simple one, because verb tense is not as simple as one might expect. Phrases often contain multiple verbs that comprise a single action. Additionally, there are quite a few different types of verbs to consider, and temporal orientation is not just a function of verb tense, but verb aspect. Temporal coding is a very active area of NLP research, and many research groups are working to develop tools and modules to handle these complications. This is one area of analysis where the potential benefits are large, and analytic efforts will be cumulative and benefit from a greater field of research.

Grammar based resources are not just convenient programming tools; they significantly change our ability to analyze text. In the process of looking for grammatical patterns, we can better understand the patterns underlying the respondents' orientation toward the question. By parsing at the level of a clause, we can examine clauses both as an element in a body of teacher commentary and as an element in a single teachers' response. Once the responses have been tagged by part of speech, we are able look for patterns in subject verb pairs and object verb pairs. In doing this, we can see which nouns were most often coupled with affective verbs, or we can see which nouns were most likely to be used agentively and which were more likely to be used passively. Once we have tagged parts of speech and isolated common combinations of words, we can isolate which aspects of the question were most frequently keyed on by the respondent. The patterning inherent in syntax and language use can be accessed directly through programming and through these open source tagsets, and these are the tools that can be used to unlock new levels and dimensions of meaningful patterns in textual data.

There is a significant advantage to parsing the responses into units that are smaller than the level of individual responses. Content coding quickly yields very small numbers of working frequencies that don't collapse well together. But looking at grammatical patterning can unlock a greater usability and hidden patterns in textual data. Although we are significantly limited by our small set of responses, many of these responses tend to be comprised of multiple points or topics. Not only can grammatical coding increase the number of observations in a dataset, but it allows us to answer a set of questions that our data is better suited to answer in more patterned and reliable ways. Using our growing understanding of the systematic grammatical patterning within a response set and the corresponding

body of research, we can begin to ask questions about patterns in the methods of the responses, not just the content.

Some may wonder about the validity of these techniques±. Indeed, validity is an important aspect to consider in any analysis. Some datasets are particularly well suited for a quantitative evaluation of the validity of their qualitative analysis. Pang and Lee (2008) did a validation study of sentiment analysis using congressional floor speeches and subsequent floor votes. But most open ended survey questions do not readily have validation measures like these. In the case of our NCLB question, the question in this analysis was the only question related to NCLB in the survey form. However, the 2004 National Survey of High School Physics Teachers did include both a quantitative and qualitative measure. It's unclear to me whether these would be a perfect conceptual match, and it's unclear to me which question would produce a more valid set of responses. There is also a possibility of finding a larger dataset to work with. Maybe a transcript from a discussion among high school physics teachers about NCLB or a more detailed exchange on the internet. A study along these lines could utilize many of the same strategies as our present study. But the participant base of the discussion would probably not be as neatly representative of physics teachers as a whole as our present one sixth working sample of the whole population. Nonetheless, grammar based strategies lend themselves better to answering questions of validity. Inter-rater reliability is often mistakenly projected to be a measure of the validity of traditional strategies.

Conclusion:

There are large bodies of research in the fields of Natural Language Processing and Linguistics that are devoted to unpacking textual data. Traditionally survey methodologists have been trained to think about textual data exclusively in terms of its content, but thinking about different kinds of patterns unlocks a strong potential for analyses with greater usability and reliability. Because language is patterned grammatically, not topically, changing the unit of analysis to appropriate grammatical units and following the rules associated with those units enables us to analyze qualitative data in new ways.

Although there has been a historical hesitancy to mix programming and text analysis, there are significant analytic advantages to programming based approaches. The decision making strategies of the coder are preserved in the code and easy to change and build upon. Time spent coding can be used to benefit other projects and contribute to a growing bank of working knowledge instead of being stuck in project specific archives. Code can be repeated when questions are repeated, in order to gauge change over time. The costs and benefits associated with programming textual analysis change the equation such that there is a wider potential for the use of textual data that often goes unanalyzed.

Statistical software packages and qualitative coding software packages recognize the current unbalance in costs and benefits associated with the analysis

of open ended responses. The field of providers is constantly evolving and growing, and we are often subjected to marketing materials from expensive qualitative analytic tools, but these tools are responding more to the demand to describe content than a desire to find patterns in textual data. There is indeed tremendous potential for software, but software needs to better anticipate our needs than adapt to our traditions. In the meantime, open source tools are also constantly growing and evolving, and programming based open source resources offer the most power and flexibility at this point.

I have mentioned a few areas where methodology from linguistics and natural language processing could be applied to the analysis of open ended survey responses, but these are by no means exhaustive. They only begin to scratch the surface of what could potentially be a very fruitful methodology for survey researchers to pursue. These strategies have a wider analytic potential as well. The ability to handle textual data in these research validated ways will allow us to better take advantage of natural patterns in our the data, and the ability to find useful and informative patterns in textual data opens us up to analyze a wider array of data sources, including social media, blogs or journal entries, and interview or Focus Group transcripts. If data is indeed the new oil, natural language processing could be the new rig.

Author's Note:

I would like to sincerely thank Dr Graham Katz of Georgetown University for his insight, Dr Susan White of the American Institute of Physics for her support and feedback, Mark McFarling of the American Institute of Physics for his technical support and insight, Garrett Anderson of the American Institute of Physics for his editorial help, Dr Anna Trester of Georgetown University for her support, and Dr Mansour Fahimi of Marketing Systems Group for his truly excellent feedback as the remarkably well prepared session discussant. The specific areas of this paper that are written in response to Dr Fahimi are indicated with a ±

To follow up with me about any of the ideas in this paper, please send email to clanger@aip.org

Bibliography:

Bennett, Winfield S; Herlick, Tanya; Hoyt, Katherine; Liro, Joseph; Santisteban, Ana. (1989). Toward a Computational Model of Aspect and Verb Semantics. *Machine Translation*, Vol. 4(4), pp. 247-280.

Bird, Steven, Ewan Klein, and Edward Loper. (2007). *Natural Language Processing in Python*. http://nltk.org/

Bird, Steven; Klein, Ewan; Loper, Edward. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. Accessed online at: http://www.nltk.org/book

Button, Graham; Casey, Neil. (1985). Topic Nomination and Topic Pursuit. *Human Studies*, Vol. 8(1), pp. 3-55.

Chafe,Wallace; Nichols, Johanna. (1986). *Evidentiality: The coding of epistemology in language*. Norwood, NJ: Ablex.

Deerwester, Scott; Dumais, Susan T; Furnas, George W; Landauer, Thomas K; Harshman, Richard. (1990), Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. Vol. 41 (6), pp. 391-408.

Erickson, Frederick. (1982). Money tree, lasagna bush, salt and pepper: Social construction of topical cohesion in conversation among Italian-Americans. In Deborah Tannen (ed.), *Text and Talk*, pp. 43-70. Washington, DC: Georgetown University Press.

Goffman, Erving. (1981). Replies and responses. In: E. Goffman, *Forms of Talk*, pp. 5-77. Philadelphia: University of Pennsylvania.

Greene, Stephan; Resnick, Philip. (2009). More than Words: Syntactic Packaging and Implicit Sentiment. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association Computational Linguistics*, pp. 503–511, Boulder, Colorado.

King, Gary, and Will Lowe. (2003). An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization*. Vol. 57, pp. 617-642. copy at http://j.mp/lxhNuB

McEnery, Tony; Xiao, Richard; Tono, Yukio. (2006). *Corpus-Based Language Studies*. London: Routledge.

Pang, Bo, and Lee, Lillian (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. Vol. 2(1-2), pp. 1-135

Pew Research Center for the People & the Press. (2008). The Candidates: In a Word. http://pewresearch.org/pubs/968/candidates-in-a-word. Accessed November, 2011.

Princeton University (2011). *Wordnet. A lexical database for English*. Retrieved November, 2011, from http://wordnet.princeton.edu/

Pustejovsky, James; Kippen, Robert; Littmman, Jessica; Sauri, Roser. (2005). Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*. Vol. 39, pp. 123-164.

Sacks, Harvey. (1992). Topic (April 17, 1968). In E. Schegloff (ed.), *Lectures on Conversation*. Blackwell.

Schaeffer, Nora Cate; Presser, Stanley. (2003). The Science of Asking Questions. *Annual Review of Sociology*, Vol. 29, pp. 65-88.

Schuman, Howard; Presser, Stanley. (1979). The Open and Closed Question. *American Sociological Review*, Vol. 44 (5), pp. 692-712.

Tagxedo.com. Accessed November, 2011, from http://www.tagxedo.com

Tannen, Deborah. (1989). Repetition in conversation: toward a poetics of talk. In *Talking Voices* (pp. 36-97). Cambridge: Cambridge University Press.

Wordcounter.com. Accessed November, 2011, from http://www.wordcounter.com