Computational studies of commonsense science:
An exploration in the automated analysis of clinical interview data

Bruce Sherin
School of Education and Social Policy
Northwestern University
2120 Campus Drive
Evanston, IL 60208
847-467-2405
bsherin@northwestern.edu

An earlier version of this paper was presented at the 2009 Meeting of the AERA, in the session Modeling Micro-Processes of Learning and Conceptual Change, Lauren Barth-Cohen and Mariana Levin, co-chairs, San Diego, CA, April, 2009
Abstract

Although researchers studying commonsense science knowledge have employed a variety of methods, one-on-one clinical interviews have played a unique and central role. The data that result from these interviews take the form of video recordings, which in turn are often compiled into written transcripts, and coded by human analysts. This coding of interview data can be seen as consisting of two steps: (1) the creation of a coding scheme and (2) the application of that coding scheme to the data corpus. In this manuscript, I explore several techniques for automating this type of analysis, techniques that are drawn from the field of statistical natural language processing. For this exploration, I make use of a data corpus consisting of 21 clinical interviews in which middle school students were asked to explain the seasons. In prior work with the same corpus, Dam and Kaufmann (2008) showed that it is possible to use techniques based on Latent Semantic Analysis (LSA) to automate the application of a coding scheme developed by human analysts. In this manuscript, I build on and extend their work, and I explore the possibility of using LSA, combined with cluster analysis, to automate both the creation and application of a coding scheme. Although these computational techniques are not yet sufficiently developed to replace human analysts, the results are nonetheless somewhat encouraging. Paradoxically, I will argue that the ability of the computational techniques to induce a coding scheme has more immediate promise than the ability to apply a given scheme.
1 Introduction

It is now widely accepted that many of the key issues in science instruction revolve around the prior conceptions of students. The central observation is that students come to science instruction already possessing a great deal of knowledge about the natural world. Some of these prior understandings are a result of earlier school-based instruction; some are learned from the popular media and informal discourse; and some are acquired simply by living in and interacting with the natural world. Here I will refer to these prior understandings as commonsense science knowledge, while acknowledging that some of this knowledge may have been acquired in earlier schooling.

Although researchers studying commonsense science knowledge have employed a variety of methods, one-on-one interviews have played a central role. In most cases, students are asked open-ended questions that allow them to respond at length. And, in many cases, the interviewer is given the freedom to spontaneously construct follow-up questions, in order to further explore the meaning of a response. This more open type of dialogue, often referred to as a clinical interview (Clement, 2000; diSessa, 2007; Ginsburg, 1997) is the focus of this manuscript.

The data that result from open-response interviews – extended utterances, consisting of natural language, gestures, and drawings – pose a number of difficulties for analysis. Broadly speaking, there are two ways that this type of data is used to make scientific claims about prior conceptions. One approach is to employ a strategy of competitive argumentation, in which selected examples are drawn from a larger corpus and are used, in written arguments, to support an author’s position and refute alternatives. This approach suffers from some relatively obvious difficulties; for example, it can be difficult to establish that selected examples are in some manner representative of the larger corpus, and are the result of some type of systematic analysis.

The second broad category of approach involves the development of a coding scheme, which is applied systematically to the corpus. However, even if the coding scheme is applied systematically, it may still be difficult to support general claims, particularly if the interviewer had the freedom to vary the prompts employed across interviews.

The purpose of this paper is to describe a new suite of analytic techniques that can be employed to analyze the data produced by clinical interviews about prior science conceptions. These new techniques are automated in the sense that a portion of the analysis that is usually carried out by hand is instead performed by a computer. Ultimately, what I am striving toward is the development of software algorithms that take raw transcripts of interviews as input, and produce coded transcripts as output.

There are many reasons to believe that this type of automation should be extremely difficult. At first glance, it would seem to require that we solve the full problem of natural language processing; that is, it seems to require that we construct a system that has many of the same capabilities for understanding natural language possessed by humans. Furthermore, the speech that occurs in commonsense science interviews can pose particular difficulties for comprehension. Student utterances are often halting and ambiguous, gestures can be very important, and external artifacts such as drawings are frequently referenced. Even human analysts can have significant difficulty understanding what students are saying in commonsense science interviews.
As a solution, I explore here the use of techniques from statistical natural language processing. In using these statistical techniques, I stop far short of attempting to solve the full problem of natural language comprehension. Instead, these techniques simply “count words;” they look at which words occur, how frequently they occur, and in what contexts. Thus, the algorithms that underlie techniques from Statistical NLP can be much easier to implement than true NLP. The question, of course, is whether they are able to capture useful attributes of our data.

Statistical NLP is a broad field, and one that is currently in rapid expansion. In the long run, I believe that it will be worthwhile to explore the use of a wide range of statistical techniques for the analysis of transcript data. Here, I make a small start: I explore the use of a technique called Latent Semantic Analysis (LSA), augmented with cluster analysis. These choices make sense for a number of reasons. LSA has already been employed, with some success, in applications relevant to educational research (e.g., Foltz, Kintsch, & Landauer, 1998; Graesser, Lu, Jackson, & Mitchell, 2004; Landauer, Foltz, & Laham, 1998; Shapiro & McNamara, 2000; Wade-Stein & Kintsch, 2004). In addition, initial attempts to apply LSA to my research team’s data proved promising, and thus justified further exploration (Dam & Kaufmann, 2008). Finally, as I will explain, given our team’s longer-term research goals and theoretical perspective, cluster analysis provided a natural extension of the work with LSA.

Thus, this paper is concerned with exploring the viability of computational techniques from Statistical NLP for the analysis of interview data about student prior conceptions. I must say a bit more about what exactly is being promised. The coding of interview data can be seen as consisting of two big steps: (1) the creation of a coding scheme and (2) the application of that coding scheme to the data corpus. (Of course these steps are generally carried out iteratively.) I will be describing computational techniques that automate both of these steps. In the most developed forms of these techniques, raw transcripts of clinical interviews are input to computational algorithms, which induce the coding scheme and apply that coding scheme to code the raw transcripts. It might seem to be more surprising that it is possible to automate the induction of a coding scheme, than to automatically apply a coding scheme developed by a researcher. However, paradoxically, I will argue that the ability of these computational techniques to induce a coding scheme has more immediate promise than the ability to apply a given scheme.

For this paper, I make use of a data corpus consisting of 21 clinical interviews in which middle school students were asked to explain the seasons (Sherin, Krakowski, & Lee, in revision). This work takes off from the foundational work done by Gregory Dam and Stefan Kaufmann (2008), which employed techniques based on Latent Semantic Analysis to apply a given coding scheme to this same corpus. The work described in this paper elaborates on these LSA-based techniques and also adds techniques based on cluster analysis. The use of cluster analysis is what makes it possible to automate the induction of the coding scheme.

1.1 Overview of this manuscript

The remainder of this manuscript will proceed as follows. In Section 2, I situate the present work within the existing literature, including research on commonsense science knowledge, and prior work on applications of LSA in education research. Then, in Section 3, I introduce the data corpus employed in this work and I overview the results that were obtained when this corpus was coded by human analysts.
Next, Sections 4 through 7 describe the computational techniques and the results obtained when they are applied. The techniques fall into two broad categories. First, Sections 4 and 5 describe computational analyses that only apply a coding scheme developed by human analysts. Sections 6 and 7 then introduce the use of cluster analysis, and show how it is possible to computationally induce a coding scheme. Finally, Section 8 concludes the manuscript, and reflects on the prospects of the techniques described here both for research on commonsense science, and the field in general.

Before proceeding, one feature of my presentation deserves some emphasis. Throughout Sections 4 through 7, I attempt, as much as possible, to describe the algorithms employed so that readers could, in principle, program these algorithms on their own. I adopted this approach in part because the goals of this manuscript are methodological; because my aim is to introduce new methods to this field, I believe it is appropriate to describe those methods in some detail. This is necessary, of course, because it can allow readers to begin using these methods themselves. But, more centrally, it is necessary because I would like readers to be able to assess the viability of these new methods; I would like readers to be able to judge, for themselves, whether the techniques presented in this manuscript are likely to be useful beyond the narrow circumstances described in this work. That requires, I believe, that readers have at least a basic working understanding of each of the main algorithms I employed. Furthermore, I believe that the fact that these methods are successful in itself poses some interesting theoretical questions for the field.

2 Literature Review and Theoretical Background

2.1 Commonsense science knowledge

As stated above, it is now widely accepted that many of the key issues in science instruction revolve around the prior conceptions of students. This focus on commonsense science leads to a perspective in which the central task of science instruction is understood as building on, adapting, and, when necessary, replacing students’ prior knowledge. One outcome of this focus has been the growth of a veritable industry of research on students’ prior conceptions. The bibliography compiled by Pfundt and Duit (Duit, 2009), which lists literature on the science conceptions of teachers and students, provides one measure of the scale of this effort. As of early 2009, the bibliography had over 8300 entries, spanning a wide range of scientific disciplines, including such disparate domains as quantum physics and health.

The great number and diversity of research on commonsense science is not accidental. Rather, it is a result of some core features of the research endeavor. When we set out to study students’ prior conceptions, we are not typically interested in general features of commonsense science that span domains, ages, and populations. Instead, the goal is to map out the specific prior conceptions, held by specific populations, in relation to specific domains. We want to know, for example, what students believe about the shape of the earth (Vosniadou & Brewer, 1992), evolution (Samarapungavan & Wiers, 1997), and nutrition (Wellman & Johnson, 1982). Furthermore, it is an assumption of this research endeavor that, even within a given domain and population, student prior conceptions may be diverse and idiosyncratic. In sum, research on prior conceptions is about mapping the “flora and fauna” of student conceptions, attending to as much of the richness and diversity as possible.
It is for this reason that *interviews* – and clinical interviews, in particular – are frequently the method of choice for researchers studying prior science conceptions. Certainly, there is a role for a wide variety of methods. But, if we truly wish to map the flora and fauna of student prior conceptions across a range of domains, then we need methods that make it possible for us to uncover a wide range of conceptions.

In discussing the literature on commonsense science, it has become commonplace to distinguish two theoretical poles. At one extreme, is the *theory-theory* perspective. According to this perspective, commonsense science knowledge consists of relatively well-elaborated theories; it is assumed that students enter instruction possessing theories that are, in some respects akin to the theories possessed by scientists, and that instruction must replace those theories (e.g., McCloskey, 1983). At the other extreme, is the *knowledge-in-pieces* (KiP) perspective. In this perspective, it is assumed that: (a) commonsense science knowledge consists of a moderately large number of elements – a system – of knowledge (Smith, diSessa, & Roschelle, 1993) and (b) the elements of the knowledge do not align in any simple way with formal science domains (Sherin, 2001).

If we accept the KiP perspective, we are faced with a number of challenges for the interpretation of clinical interview data. If commonsense science consists of a system of knowledge elements, we must assume that, in general, answers given will be generated, in-the-moment, out of the elements that comprise the system. If our goal is to use a clinical interview to identify knowledge, then we must somehow see through the answers given to the underlying knowledge elements that generated them.

There are many possible theoretical perspectives that lie somewhere between the two extreme poles I associated with the theory-theory and KiP perspectives. I believe that the computational methods described in this paper should be of interest to a broad range of researchers who study commonsense science, and adopt a range of theoretical perspectives. However, the exploration of computational methods presented in this paper was biased by my own theoretical perspective, which lies closer to the KiP pole. As I hope will become evident, my exploration of computational methods has been driven by a desire to get at the more basic knowledge – the pieces – that I believe comprise commonsense science knowledge.

How might the new computational methods contribute to the field of commonsense science research? There are at least two sorts of reasons that one might imagine that automation of this sort would be beneficial. First, coding data by hand, or extracting examples from a corpus, is a significant amount of work. If we could automate any of this analysis, it seems possible we might save researchers time and effort. However, the techniques presented in this paper have not reached the point where they can substitute for the work done by humans, and I am not sure whether that will be possible in the near term. There is, however, a second way in which the automated techniques described here can be of immediate use. I will argue that they are useful not because they can replace human-based analyses, but because they can provide independent *support* for that analysis. This type of support is particularly important given the inherent difficulty of producing scientifically convincing analyses of data from clinical interviews.

### 2.2 Natural Language Processing and Latent Semantic Analysis

The field of Natural Language Processing (NLP) is concerned with giving computers the ability to both understand and generate natural language. Researchers in this field are typically linguists
and computer scientists. Statistical NLP is a subfield of NLP that uses a variety of statistical and probabilistic methods to solve some of the thorny problems in NLP that are not amenable to other sorts of solutions. Analyses from traditional NLP are frequently concerned with the parsing of individual sentences in order to extract their meaning. This means understanding the structure of sentences so as to identify components such as noun phrases, prepositional phrases, and the parts of speech of the individual words that comprise them. In contrast, analyses in Statistical NLP are usually concerned with analyzing statistical properties of a large corpus of text. These analyses might look simply at how often certain words appear in the corpus. But there are, of course, more sophisticated analyses.

At the heart of the new methods I will describe in this paper is a technique called Latent Semantic Analysis (LSA). It is reasonable to think of LSA as a relatively sophisticated technique from Statistical NLP. However, LSA had its origins in the field of Information Retrieval, which was historically separate from the field of NLP, and from linguistics more broadly. Information Retrieval is primarily concerned with the retrieval of information from large electronic databases. In the prototypical situation, there is a corpus of documents, numbering in the hundreds or even thousands. The goal is for users to be able to type queries into a computer-based system, which will respond by returning the documents in the corpus most closely related to the query. Although the field of Information Retrieval was initially distinct from NLP, the fields have become increasingly merged. (For a lucid introduction to Statistical NLP, see Manning & Schütze, 1999.)

Early information retrieval systems were based on straightforward word-matching algorithms. However, these systems suffered from the limitation that they would only return documents in which there was a literal match to the words employed in the query; they could not return documents that were on the desired topic, but employed different terms. LSA was one technique that was developed to fill this need (Berry, Dumais, & OBrien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). It allows the document retrieval system to recognize words or passages that are similar in meaning to query terms, even if the literal terms that appeared in the query do not appear in the passage.

To accomplish this, LSA makes use of a vector space model. In a vector space model, the meaning of a block of text – a word, paragraph, essay, etc. – is associated with a vector, usually in a high dimensional space. So, two blocks of text have the same meaning to the extent that their vectors are the same. In this way, LSA makes it possible to compute the similarity in meaning between any pair of words or blocks of text. In Section 4, I will describe, in some detail, the algorithms employed in the particular variety of LSA used in this work.

2.3 Applications of LSA in education

Outside of information retrieval, some of the earliest and most persistent uses of LSA have been in applications related to education (Haley, Thomas, De Roeck, & Petre, 2005). These applications have been of two broad types. First, LSA has been used as a research tool by educational researchers – that is, as a means of analyzing data, in order to study thinking and learning. Second, LSA has been used as a component of intelligent instructional systems. The majority of these educational applications, across both types, have been focused on the teaching of reading and writing.
For example, LSA-based systems have been employed to automatically score essays written by students (Landauer, et al., 1998; Landauer, Laham, & Foltz, 2003; Pérez, Gliozzo, Strapparava, & Alfonseca, 2005). Typically, these systems work by first computing a meaning vector for a student essay, and then comparing this meaning vector to the vectors associated with one or more reference essays. For example, a student’s essay might be compared to an “ideal” essay constructed by a researcher.

Another early use of LSA-based systems in education was to judge the difficulty of texts, and their appropriateness for children with differing reading abilities (Foltz, et al., 1998). These systems work by judging the coherence of texts. This is typically done by computing the meaning vectors for portions of a text, and then comparing these vectors. Texts that exhibit sharp changes in meaning are judged to be less coherent and thus more difficult to read.

In a number of applications, students are asked to summarize a passage or document that they have just read, and an LSA-based system is used to evaluate these summaries. In evaluating them, the system can do more than simply judge the quality of the student’s summary as an essay; it can also be used as a means of determining the extent to which the student understood the material contained in the passage. For example, Shapiro and McNamara (2000) had students read and summarize portions of psychology textbooks. Using LSA, these summaries were then compared both to the text the students read, and to model essays composed by experts.

In most of these uses of LSA, the data consisted of text that was written by participants in the research. However, in a few instances, LSA has been applied to transcriptions of verbal data. For example, in the study just mentioned, Shapiro and McNamara (2000) found that LSA could be applied successfully both to written summaries of the textbook and to transcriptions of verbal summaries given by students. Similarly, Magliano and Millis (2003) applied LSA to think-aloud protocols that students produced as they read passages of text.

As mentioned above, LSA has been used as a component of intelligent instructional systems. For example, intelligent systems have been constructed that provide feedback to students on summaries that they write of a given text passage (Wade-Stein & Kintsch, 2004; Wiemer-Hastings & Graesser, 2000). One LSA-based system, AutoTutor, is of particular interest here because it has been applied to teach science-related subject matter (Graesser, et al., 2004). AutoTutor teaches physics by first posing a problem or question. The student responds by typing a response into the system. The system then evaluates that response by using LSA to compare the student’s text to a set of predefined expectations and misconceptions; the expectations are pre-specified components of a correct response and the misconceptions are possible erroneous ideas that might be expressed by the student. Based on this analysis, the system responds by posing further questions to the student, either to help correct the misconceptions, or to draw out more components of a complete answer to the original problem.

I want to say a bit about where the work described in the present paper fits within the space of uses of LSA in education. First, in this work, LSA is used as an analytic tool for researchers; it is a tool that I use in order to see into the knowledge possessed by students. I will not be describing an LSA-based system that is used by students.

Second, I apply LSA to verbal data. The great majority of applications of LSA in education use text that is typed by a student, either in the form of an essay or short responses. Furthermore, prior research that has worked with verbal data has employed data that is very different than that employed in the present work. For example, the work by Shapiro and McNamara (2000) and
Magliano and Millis (2003), which I mentioned above, employed a more constrained type of think-aloud protocol, focused on passages of text that were just read. In contrast, the verbal data employed in this work consists of relatively free-flowing discussions involving back-and-forth between an interviewer and interviewee. Furthermore, gestures played a central role in these interviews, and students also drew and referred to diagrams as they spoke.

There have been prior applications of LSA to more open dialog, but that prior research has had concerns that differ greatly from that of the present work. For example, Foltz and colleagues used LSA-based techniques to measure the coherence of speech of patients with schizophrenia (Foltz, 2007). And Feinerer and Wild (2007) applied LSA to interviews conducted as part of marketing research.

I will conclude by mentioning what I believe is the most important way in which the present work diverges from prior uses of LSA in education. In all prior applications, answers given by students, whether in written or verbal form, were evaluated by comparison to a predefined model. This model might be, for example, some portion of the text just read, or an ideal answer constructed by the researcher. For example, AutoTutor analyzes a student’s response by comparing it to a set of misconceptions and expectations, which are embodied in a set of documents constructed by researchers. Some of the analyses I describe will employ idealized comparison documents of this sort. However, as mentioned above, I will also describe techniques for automatically inducing a set of conceptions from the data itself.

### 3 The data corpus and coding by human analysts

#### 3.1 The data corpus

The data used in this work was drawn from a larger corpus collected by the NSF-funded *Conceptual Dynamics Project* (CDP).¹ As part of that project, interviews were conducted with middle school students on a wide range of science-related topics, usually in the context of curricular interventions on related subject matter. For the present work, I draw on a set of interviews in which students were asked about issues pertaining to the earth’s climate and seasons. Our group has 21 interviews on this subject that were conducted prior to any related intervention.

My analysis focuses on the portion of these interviews in which students were asked to explain why the earth experiences seasons. This portion of the interview always began with the interviewer asking “Why is it warmer in the summer and colder in the winter?” After the student responded, the interviewer would, if necessary, ask for elaboration or clarification. The interviewer had the freedom, during this part of the interview, to craft questions on-the-spot in order to clarify what the student was saying.

Next, the student was asked to draw a picture to illustrate their explanation. Then, once again, the interviewer could ask follow-up questions for clarification. Our interviewers were also prepared with a number of specific follow-up questions to be asked, as appropriate, during this part of the interview. Some of these questions were designed as challenges to specific explanations that

---
¹ NSF grant #REC-0092648. *Conceptual dynamics in complex science interventions* (B. Sherin, PI).
students might give. On average, the portion of our interview in which students were asked about the seasons lasted 4.25 minutes, but it ranged from less 1.5 minutes to over 9 minutes.

The seasons have long been a popular subject of study in research on commonsense science, and a significant number of studies have set out to study student and adult understanding in this area (Atwood & Atwood, 1996; Newman & Morrison, 1993; Sadler, 1987; Trumper, 2001). Looking across these studies, it is clear that it is difficult for individuals of all ages to give a fully correct explanation of the seasons. However, beyond that simple generalization, there does not appear to be any clear consensus about the set of explanations that are given, or the frequency with which any explanations appear. Part of the problem seems to be that slightly different prompts can elicit a great diversity of explanations. For example, in one study, Atwood and Atwood (1996) gave a written test to 39 preservice teachers. In their analysis, they identified 16 different explanations of the seasons given by the teachers. Furthermore, only four of these explanations were given by more than one of the student teachers.

The difficulty of explaining the seasons, coupled with the diversity of responses, point to the features of this subject matter that have made it a fertile area in which to study commonsense science knowledge. Explaining the seasons is challenging enough that even educated adults find it difficult. At the same time, the question is accessible enough that individuals with a wide variety of backgrounds and ages can make progress in constructing sensible explanations. The same features of the seasons that have made it a popular area of focus for other researchers, also make it productive for the purposes of the present work. However, ultimately, we will have to consider whether the results reported here are likely to be replicable in other areas of subject matter that might be less optimal for this kind of research.

3.2 Coding of the data by human analysts

In prior work with our seasons data, Conceptual Dynamics researchers have adopted a strongly KiP perspective (Sherin, et al., in revision; Sherin, Lee, & Krakowski, 2007). We assume that students possess a system consisting of many knowledge elements – the “pieces” – that may potentially be drawn upon as they endeavor to explain the seasons. When a student is asked a question during an interview, some subset of these elements are activated. The student then reasons based on this set of elements, and works to construct an assemblage of ideas in the service of explaining the seasons. We refer to this assemblage of ideas as the dynamic mental construct or DMC, for short.

Furthermore, in prior work, Conceptual Dynamics researchers have attempted to understand the dynamics of interviews at the level of the knowledge elements; we have attempted to identify specific knowledge elements, and to describe in some detail how DMCs are constructed in specific interviews (Sherin, et al., in revision). In contrast, in the present paper, the analysis is restricted to the level of DMCs. What that means is that I will look only at explanations that are constructed by students, not at the process by which these explanations are constructed out of existing knowledge. However, I do hope to give some sense for how feasible it will ultimately be to employ computational techniques to capture dynamics at the level of knowledge elements.

Although there is substantial diversity in the DMCs constructed by students in our data corpus, it is possible to place these DMCs into three broad categories. The first category, closer-farther, is illustrated by the diagram in Figure 1a. In closer-farther explanations, the earth is seen as orbiting the sun in such a way that it is sometimes closer to the sun and sometimes farther. When
the earth is closer to the sun then it experiences summer; when it’s farther away it experiences winter. Note that, in closer-farther explanations the entire earth experiences the same season at a given time.

The second category of DMC, *side-based*, is illustrated in Figure 1b. Side-based explanations are focused on the rotational motion of the earth, rather than its orbital motion. In side-based explanations, the earth rotates so that first one side, then the other, faces the sun. The side facing the earth at a given time experiences summer, while the other side experiences winter.

The third and final category of DMC, *tilt-based*, is depicted in Figure 1c. Tilt-based DMCs depend critically on the fact that the earth’s axis of rotation is tilted relative to a line connecting it to the sun. In a tilt-based explanation, the hemisphere that is tilted toward the sun experiences summer and the hemisphere that is tilted away experiences winter. This category includes the normative scientific explanation, as well as some non-normative explanations.²

As discussed above, it is an assumption of the KiP perspective that interviewees will, in general, be constructing answers out of more basic knowledge resources during an interview. In prior work, we have argued that this happens in a very visible and dramatic way in our seasons data corpus (Sherin, et al., in revision). In a majority of cases, the students interviewed did not have a fully-formed explanation of the seasons ready-to-hand, and they were therefore in the position of needing to invent one. Thus, students’ explanations were highly dynamic and not always amenable to easy coding.

² In the normative explanation, the earth maintains a fixed tilt as it orbits the sun. For this reason, first one hemisphere, then the other, it tilted toward the sun. The hemisphere that is tilted toward the sun is struck more directly by sunlight, and is thus warmer.
Nonetheless, if we insist that our analysts code each student interview in terms of one of the above three categories of DMCs, we can do so in the majority of cases, and with high reliability. I attempted to code all 21 interviews as belonging to one of the three main categories – closer-farther, side-based, and tilt-based. In addition, I coded some interviews as *shift* if a student clearly expressed more than one prominent DMC in the course of the interview. The results are shown in Table 1. A second researcher, not involved in earlier stages of the research, coded the 21 interviews with the same four categories. We agreed in every case but one, resulting in a Cohen’s Kappa of .94. (The one disagreement was for a student, Richard, that I coded as side-based, but the second researcher coded as *shift*.)

<table>
<thead>
<tr>
<th></th>
<th>Closer-Farther</th>
<th>Side-Based</th>
<th>Tit-based</th>
<th>Shift</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td><strong>Second Coder</strong></td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>21</td>
</tr>
</tbody>
</table>

*Table 1. Coding of the data corpus*

Figure 2. Complete text of the seasons portion of the interview with Jill.

---

Int. Can you tell me why it’s warmer in the summer and colder in the winter?

Jill It’s because the sun um, we rotate around the sun [gestures in a circle with pointed index finger] like in an axis, but its not a perfect circle, and when, and then like, or not an axis like we orbit [gestures in a circle], it’s like not a perfect circle it’s like egg-shaped almost [using both hands, makes an oval shape] but not very noticeable and the sun, the earth is on an axis [points index finger up] on that orbit [spins finger that’s held up] that when it goes around like there’s one part that’s closer and one part that’s farther so that’s kind of [using other hand, points to different sides of finger that’s held up], that explains why it’s warmer near the equator cause that’s where the sun hits it the most, but it depends on how the orbit is, like in the summer the sun is closer to the earth because we’re, like the, the, our orbit’s circle [gestures in a circle] is not like a perfect circle around the sun [makes oval with both hands], it’s not like, this is the sun [indicates a point in space], it’s not like there [makes oval shape with both hands] it’s like that kind of [shifts oval shape over].

Int. Okay.

Jill So this is where the [points to one side of the orbit] like that’s the summer part and that’s the winter part [points to other side of the orbit].

Int. Okay, okay, ya know maybe [hands Jill a sheet of paper].

Jill Here’s the sun, and that one [begins to draw], so this would be the summer [labels summer on diagram] this is winter [labels winter].

Int. And its uh, cause

Jill Cause when the earth is closer the sun rays hit it’s like closer, so it’s hot.

Int. Okay um, I don’t know, I’ve heard that when its winter in the US its summer in Australia, have you heard that?

Jill No.

Int. No? Yeah, I was just wondering. Yeah, that’s what I hear, so I was just wondering how that would work given what we’ve talked about a description?

Jill Um, um, I th- maybe um, where Australia is placed on the earth the sun hits it less but I don’t think it, so I think it’s just colder there, but I don’t think its actually like ss- I think the whole earth is at summer at one time, like it goes together but where Australia is placed it gets weird seasons because of that, like the temperature of the seasons are different.
To give a flavor for the content of our interviews, the text from one interview, with a student Jill, is given in Figure 2. The primary DMC constructed by Jill is a closer-farther explanation. Jill emphasizes throughout the interview that the orbit of the earth is not a perfect circle, and states that it will be warmer when the earth is closer to the sun. For example, in her first utterance she says: “it depends on how the orbit is, like if, in the summer the sun is closer to the earth because … our orbit’s circle is not like a perfect circle around the sun.” Furthermore, when asked to draw a picture, she produced the diagram shown in Figure 3, which is consistent with her closer-farther DMC.

![Figure 3. Jill's drawing.](image)

It is worthwhile to look at the transcript of Jill’s interview in Figure 2 and reflect on some of the challenges that will be faced as we attempt to reproduce the human analyses computationally. Clearly, Jill’s language is halting and her thoughts are often left incomplete. Furthermore, the computational techniques will only have access to the spoken words; thus, some of the information that helps a human analyst disambiguate Jill’s utterances will not be available to the computational techniques. Our human analysts had access to the drawing shown in Figure 3. Furthermore, the human analysts usually worked directly from videos of the interviews, and thus had access to Jill’s gestures, intonation, facial expressions, and so on. In the transcript shown in Figure 2, I have attempted to describe some of Jill’s gestures, but it is hardly possible to do them justice. Jill gestured constantly as she spoke, and she often referred to her gestures, as when she said “our orbit’s circle [gestures in a circle] is not like a perfect circle around the sun [makes oval with both hands], it’s not like, this is the sun [indicates a point in space], it’s not like there [makes oval shape with both hands] it’s like that kind of [shifts oval shape over].”

3.3 The work of Dam and Kaufmann (2008)

As mentioned earlier, the work described in this paper builds very directly on prior work by Dam and Kaufmann (2008). Gregory Dam was a researcher on the Conceptual Dynamics Project who forged an initial relationship with Stefan Kaufmann, a linguist. Their work made use of data collected by the Conceptual Dynamics Project, the same data corpus that was used for the work discussed in this article. As I describe my own techniques in the sections that follow, I will point to similarities and differences of my techniques to the techniques employed by Dam and Kaufmann. Here, I will just briefly summarize the analyses they performed.

Recall that LSA is a vector space model, and that the meaning of a unit of text is associated with a vector in a high dimensional space. In Dam and Kaufmann’s primary analysis, they began by computing a vector for the transcript from each student interview. Then they wrote a set of idealized answer documents, one corresponding to each of the three categories of DMCs, and

---

3 All names are pseudonyms.
they computed a vector for each of the idealized answers. Finally, the vector for each student transcript was compared to these three idealized vectors. The most similar idealized answer vector was taken to be the code for that student.

Using this procedure, along with some refinements to be described below, Dam and Kaufmann report an agreement with human coders that is as high as 90%. For a number of reasons, however, the results reported here will not be directly comparable. Most importantly, our research group’s human analysis has changed in the years since Dam and Kaufmann completed their work.

Dam and Kaufmann also conducted one additional analysis. They selected two students who were thought to dramatically shift their explanation over the course of their interview, and then segmented the transcripts for these students into 50-word segments. They then coded each of the segments using the same procedure used for coding whole transcripts. Dam and Kaufmann concluded, tentatively, that there was some evidence that their techniques could capture meaning in 50-word segments, and that it could thus possibly be used to trace shifts in student explanations during an interview.

My analysis in Section 4, which follows, adheres fairly closely to the procedure employed by Dam and Kaufmann in their primary analysis. In that analysis, I compare student transcripts to idealized response documents. In Section 5, I extend Dam and Kaufmann’s analysis of segmented transcripts. Then, as mentioned earlier, the analyses in Sections 6 and 7 depart more dramatically from Dam and Kaufmann’s techniques. In those analyses, I make no use of comparison documents. Instead, I use cluster analyses to discover a coding scheme.

4 Analysis using comparisons to idealized responses

I am now ready to present the first type of computational analysis, the type that, following Dam and Kaufmann (2008), makes use of the three coding categories developed by human analysts. Here I begin to describe my analysis procedures in more detail. Throughout my description of the analysis process the reader is encouraged to refer to Figure 4, which describes the overall plan for analysis in schematic form. In brief, what I am aiming for is a system that takes a transcript document as input, and outputs a code for that transcript. In reality, as shown in Figure 4, three inputs are required: in addition to the transcript documents, there are a training corpus and idealized response documents. In the sub-sections that follow, I will go step-by-step through the procedure in Figure 4.

4.1 Processing the training corpus with Latent Semantic Analysis

The analysis process begins at the top-left of Figure 4, with the training corpus. All applications of LSA need a way to associate a vector with each word that will appear in the data of interest. To do this, a large corpus of text, consisting of hundreds or thousands of documents, is used to teach the LSA system the “meaning” of individual words (in the form of the vectors in a high-dimensional space).

The training corpus employed for this work consists of 196 documents, totaling about 300,000 words, assembled by hand from publicly available web pages.4 (By the standards of LSA, this is

---

4 The training corpus is available from the author upon request.
a relatively small training corpus.) Following Dam and Kaufmann (2008), as well as the guidance provided by Shapiro and McNamara (2000) and Wolfe and Goldman (2003), the corpus assembled is domain-specific. All of the documents included in the corpus pertain to topics relating to earth’s seasons and climate. They are somewhat varied in character, however. Some are highly technical and clearly geared to a more expert audience; others are much more informal and, in some cases, were clearly written with children in mind. I will say more about the choices underlying the composition of the training corpus in Section 8.

![Diagram](Image)

**Figure 4. Overview of analysis using idealized responses.**

Prior to analysis, the documents comprising the corpus were assembled into one large computer file, with boundaries between documents indicated by markup codes. The entire file was then fed into a Latent Semantic Analysis engine. For the present work, I made use of a particular LSA implementation known as InfoMap (Takayama, Flourney, Kaufmann, & Peters, 1999).[^5]

![Table](Image)

**Figure 5. Building the co-occurrence matrix.**

The processing of the training corpus begins with the building of the co-occurrence matrix. This matrix is depicted schematically in Figure 5. Each row in this matrix is labeled, down the left side, with one of the words that appears in the training corpus. With the exception of some words

to be discussed below, every word that appears in the training corpus is included as a row label in the matrix. For the corpus employed in this work, there were 12,493 rows.

Across the top of the co-occurrence matrix are the 51st through 1050th most common words in the training corpus. These are selected because it is believed that these moderately common words are likely to be “content-bearing;” they are the words that are most likely to help distinguish the meaning of one passage from another (Takayama, et al., 1999).

The cells in the body of the co-occurrence matrix are filled in as follows. The analysis engine walks through the training corpus, one word at a time, until it encounters a word that appears as a label on the left-hand side of the matrix. It then marks out a fixed window around this focal word in the training corpus, which begins 15 words before the focal word, and ends 15 words after the focal word. It then looks to see if any of the 1000 words at the top of the matrix columns appear in this fixed window. If one of these words does appear, then the value in the corresponding cell is incremented.

Thus, suppose, for example, that the analysis engine is walking through the training corpus and encounters the word “earth.” It would then set up a window stretching backward 15 words and forward 15 words, and look to see if any of the column labels appear within this window. If, for example, the word “radiate” appears within 15 words of the word “earth,” the cell at the intersection of the “earth” row and “radiate” column would have its value increased by one.

My description of the algorithm to this point has glossed over several details. First, not every word that appears in the training corpus is allowed to appear as a row label in the co-occurrence matrix. Rather, a list of stop words are specified in advance and these words are simply ignored by the InfoMap engine. In the analyses described in this paper, I employed a standard stop list of 765 words that is included with the InfoMap package. The list includes extremely common words such as “the” and “then,” and also not-quite-as-common words such as “regular” and “zero.” I also employed a go list of 21 words, created by Dam and Kaufmann (2008) (refer to Figure 6). Words that appear in the go list are included as row labels whether or not they appear in the stop list.

<table>
<thead>
<tr>
<th>facing toward</th>
<th>towards</th>
<th>away</th>
<th>slanted</th>
<th>slant</th>
<th>leaning</th>
<th>farther</th>
<th>further</th>
<th>closer</th>
<th>nearer</th>
<th>near</th>
<th>far</th>
<th>direct</th>
<th>sun</th>
<th>tilt</th>
<th>axis</th>
<th>earth</th>
<th>sun</th>
<th>light</th>
<th>angle</th>
</tr>
</thead>
</table>

Figure 6. The list of go words.

After the cells of the co-occurrence matrix are filled in as described above, the raw counts are transformed in two ways. First, each count is multiplied by a weighting factor:

\[
w = t_f (\log(D + 1) - \log(d_f))
\]

where \(t_f\) is the total number of times that the column label appears in the corpus, \(d_f\) is the number of documents in the corpus in which the column label appears, and \(D\) is the total number of documents. What this means is that terms that appear a large number of times in the training corpus, while being localized to a small number of documents, are given a higher weight.

Second, the square root of each entry is taken in order to smooth out the effect of extreme values.

At this point, we have a matrix with 12,493 rows, each of which has 1000 entries, and is labeled by a word on the left. Alternatively, we can say that each of the row labels is now associated with a 1000-dimensional vector. This set of vectors can be interpreted as a vector space model of
The sort discussed in Section 2.2. Thus, we can associate the “meaning” of each row word with the direction in which its 1000-dimensional vector points.

However, LSA introduces an additional, critical step to reduce the size of co-occurrence matrix and the vectors that comprise it. Throughout my presentation here, I have been attempting to describe the algorithms employed in such a manner that readers could, in principle, program these same algorithms on their own. However, at this point I am going to have to skip some steps. Unlike all of the other algorithms employed in this work, the technique used to reduce the size of the co-occurrence matrix, singular value decomposition (SVD), is complex enough that it is not amenable to a brief description. For the purposes of the present exposition, it is sufficient for the reader to know that, in this work, the dimensionality of the matrix is reduced by SVD so that it has 200 columns rather than 1000.6

The reduced co-occurrence matrix is shown schematically in Figure 7. This reduced matrix has the same labels down the left side as the full matrix. However, the column headings no longer have a simple interpretation (they are no longer associated with words in the training corpus). Each row in the reduced matrix is normalized, which means that each of the 200-dimensional vectors specified by a row has a length of 1.

Like the co-occurrence matrix in Figure 5, the reduced matrix can be interpreted as providing a dictionary of meanings for each of the 12,493 words that appear as row labels. This matrix, which is referred to as the Word Space, is the final output of the analysis of the training corpus.

The reduction of the co-occurrence matrix has two benefits. The first of these benefits is relatively simple to understand: the reduction makes computations using the matrix more manageable. The second benefit is of more import, but is also more subtle: the reduction uncovers “latent” associations among words. For illustration, I will focus on two rows from the hypothetical co-occurrence matrix presented in Figure 5, the sun row and the earth row. (For convenience, these two rows are reproduced in Figure 8.) Note that there are no columns in which both sun and earth have non-zero entries; sun only has an entry in the third column, and earth only has an entry in the first two columns. If this were true for all of the column entries in these two rows, then, in a vector space model, this would mean that there isn’t any overlap in meaning between the words sun and earth.

Figure 7. The reduced co-occurrence matrix.

<table>
<thead>
<tr>
<th></th>
<th>200 dimensions ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>.2685 .3050 .1381</td>
</tr>
<tr>
<td>earth</td>
<td>.2963 .2091 .1320</td>
</tr>
<tr>
<td>solar</td>
<td>.2312 .0343 -.0201</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. Two rows from the row co-occurrence matrix.

<table>
<thead>
<tr>
<th></th>
<th>32</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>7</td>
<td>82</td>
</tr>
<tr>
<td>earth</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 See Section 8 for a discussion of the choice of 200 as the number of columns in the reduced matrix.
However, after the reduction in dimensionality through the SVD, there might be columns in which both *sun* and *earth* have non-zero entries. If so, this implies that there is overlap in meaning between these words. (This is the case in my hypothetical reduced matrix in Figure 7.) This is a sign that SVD is uncovering latent relationships among words (Deerwester, et al., 1990; Landauer, et al., 1998).

Although it will not play a role in the later steps of my analysis, it is instructive to see how we can use the Word Space to compute the similarity in meaning of pairs of words that appear as row labels. In line with the vector space model, two words are similar in meaning to the extent that their vectors point in the same direction. Mathematically speaking, we can say that the similarity of two words $u$ and $v$ is given by the dot product of the corresponding rows from the Word Space matrix:

$$similarity = \sum_{i=1}^{200} u_i v_i$$

What this means is that, to compute the similarity of two words, we multiply the entries that appear in corresponding columns, and then sum the resulting products.

<table>
<thead>
<tr>
<th></th>
<th>angle</th>
<th>axis</th>
<th>orbit</th>
<th>rotate</th>
<th>slant</th>
<th>spin</th>
<th>tilt</th>
</tr>
</thead>
<tbody>
<tr>
<td>angle</td>
<td>1.000</td>
<td>0.097</td>
<td>0.015</td>
<td>-0.043</td>
<td>-0.057</td>
<td>-0.022</td>
<td>0.275</td>
</tr>
<tr>
<td>axis</td>
<td>1.000</td>
<td>0.263</td>
<td>-0.031</td>
<td>-0.072</td>
<td>0.134</td>
<td>0.412</td>
<td></td>
</tr>
<tr>
<td>orbit</td>
<td>1.000</td>
<td>-0.026</td>
<td>0.030</td>
<td>-0.010</td>
<td>0.150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rotate</td>
<td>1.000</td>
<td>-0.119</td>
<td>0.141</td>
<td>-0.075</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slant</td>
<td>1.000</td>
<td>0.079</td>
<td>0.030</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spin</td>
<td>1.000</td>
<td>0.153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tilt</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Some sample dot products between words in the Word Space.

For illustration, Table 2 shows the similarities between a selection of words from the Word Space. Each word has a dot product of 1.00 with itself. (It has the same meaning as itself.) The word *tilt* is apparently closely associated with *angle* and *axis* but not with *rotate*. The word *rotate* is closely associated with *spin* but not with any of the other words that appear in the table.

### 4.2 Computing Document Vectors

The next step in the analysis (refer to Figure 4) is to employ the Word Space to compute vectors for each of the transcript documents. Before we can compute a vector for a transcript document, such as the transcript of the interview with Jill, we must strip out everything except the student’s speech, removing, for example, statements made by the interviewer, notations about gestures and drawings, and punctuation. In the case of Jill’s transcript, the resulting document has 269 words. Once this is done, the procedure for computing the document vector is simple – perhaps surprisingly simple. We take every word that appears in the document, write the 200-number vector from the Word Space that corresponds to that word, and then add up all these vectors, normalizing at the end.
To code the transcripts, we need reference vectors to which we can compare the transcript vectors. For that purpose, I constructed three idealized response documents, each corresponding to one of the three DMCs in the coding scheme: closer-farther (88 words), side-based (87 words), and tilt-based (127 words). For illustration, the text of the idealized closer-farther document is reproduced in Figure 9. A document vector was computed for each of these idealized answers, using precisely the same technique employed to compute a vector for transcript documents.

![The earth orbits around the sun. It takes one year for it to go around. The earth orbits in an ellipse so that sometimes the earth is closer to the sun and sometimes it's farther away from the sun. When the earth is over here, it's closer to the sun, it gets more heat, so that makes it warmer and it's summer. When the earth is over here, it's farther from the sun, it gets less heat from the sun, and it's colder. So that's when it's winter.](image)

*Figure 9. Text of the closer-farther idealized document*

4.3 Computing dot products to code transcripts

To code the transcript documents, we now determine the similarity between the transcripts and each of the idealized answers by computing the dot products. We then interpret the largest similarity value as the code to be assigned to the document. Table 3 and Figure 10 display the results when these dot products are computed for the document vector for Jill’s transcript. As shown in Table 3 and Figure 10, the greatest similarity is to the closer-farther idealized document, which agrees with the coding of Jill by human analysts.

<table>
<thead>
<tr>
<th></th>
<th>Closer-father</th>
<th>Side-based</th>
<th>Tilt-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jill</td>
<td>0.86</td>
<td>0.69</td>
<td>0.67</td>
</tr>
</tbody>
</table>

*Table 3. Similarities between Jill’s transcript and each of the three idealized answers.*

![Similarities between Jill’s transcript and each of the three idealized answers.](image)

*Figure 10. Similarities between Jill’s transcript and each of the three idealized answers.*

Although the results summarized in Table 3 and Figure 10 agree with the human coders, there is a respect in which they are not satisfying. The problem is that Jill’s transcript exhibited comparatively large alignments with all three idealized answers, not just closer-farther. In retrospect, this result is not surprising. LSA was developed as a means of locating documents that pertain to a given topic. However, all of the documents involved in this step of the analysis are about very similar subject matter; they all explain the seasons, and they all do so by talking about the position and motion of the earth in relation to the sun.
This observation suggests that our particular application of LSA might require some special techniques for differentiating among documents with similar subject matter. For that purpose, I compute what I call *deviation vectors*. As illustrated in Figure 11, to compute the deviation vectors for two vectors \( V_1 \) and \( V_2 \), we first find their average, and then break each vector into two components, one that lies along the average, and another that is perpendicular to the average. The perpendicular components, \( V_1' \) and \( V_2' \), are the deviation vectors. If we use these deviation vectors in place of the original vectors, the result is that \( V_1 \) and \( V_2 \), have each been replaced by the component that defines its unique piece – a piece that characterizes how it differs from the average. The same procedure can be employed with any number of vectors.

![Figure 11. How to compute deviation vectors.](image)

<table>
<thead>
<tr>
<th></th>
<th>Closer-father</th>
<th>Side-based</th>
<th>Tilt-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jill</td>
<td>0.31</td>
<td>-0.11</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

*Table 4. Similarities between Jill’s transcript and deviation vectors for the idealized answers.*

As a next step, I now compute deviation vectors for the three idealized response vectors, and thus replace each with the vector that characterizes how it differs from the average of the three.\(^7\) Then these deviation vectors are compared to Jill’s transcript vector. When this is done, I obtain the results given in Table 4 and Figure 12. Now the closer-farther explanation is the clear winner, in line with the judgments of human analysts.

\(^7\) Dam and Kaufmann (2008) used a different procedure to achieve a similar result. They removed words from the idealized answer documents that appeared in more than one document, thus leaving only the words that were unique to that document. This method gave very good results, but it does not generalize to a case in which there are many documents (which we will encounter in a later section). If there are many documents, then removing words that appear across multiple documents would leave many documents with no words at all.
4.4 Results for the full set of transcripts

The same machinery used to code Jill can now be applied to the full set of 21 transcripts. First, as for Jill, each of the transcripts was stripped down so that it only included the student’s speech. The resulting documents have an average of 300 words, and range from 59 to 717 words. When these documents are coded, the results are as shown in Figure 13. This figure uses a format I will use throughout the sections that follow. As with Jill, each student is represented by a set of three bars, one for each of the three codes: black for closer-farther, white for side-based, and grey for tilt-based. The longest of these three bars is the code for the associated student.

![Figure 13. Results when the transcripts vectors for all 21 students in the data corpus are compared to the idealized answer (deviation) vectors.](image)

<table>
<thead>
<tr>
<th>Student-only transcript</th>
<th>Agree</th>
<th>Disagree</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>4</td>
<td>0.62 (good agreement)</td>
</tr>
</tbody>
</table>

| Student + Interviewer   | 14    | 2        | 0.81 (very good agreement) |

Table 5. Results of comparison to human coders.

These results can be compared to the coding by human analysts. As shown in the first row of Table 5, the LSA-based analysis agrees with human coders on 12 of 16 codable transcripts, and with a Cohen’s Kappa of .62. (In this computation, I am ignoring the 5 students that were coded as shift by the human coders.) Table 5 also presents results from one other analysis. In this second analysis, both the student and the interviewer utterances are included in the transcript documents, rather than only student utterances. When that is done, the LSA-based analysis agrees on 14 of the 16 codable transcripts, with a Kappa of 0.81.

4.5 Discussion

To this point, I have been applying the program for analysis developed by Dam and Kaufmann (2008), with only a few refinements. Like their results, the findings reported thus far, while not perfect, are nonetheless encouraging. On the one hand, the results fall short of the high reliability achieved by human analysts with our coarse coding scheme. On the other hand, prior to the work of Dam and Kaufmann, one might have thought that this approach would not produce any meaningful results. As I have emphasized, the LSA-based analysis is working without much of the information available in traditional coding – it knows nothing, for example, about diagrams, gestures, or facial expressions. Furthermore, the LSA-based analysis discards additional information, such as word order, that is available in the raw transcripts. Nonetheless, this limited
stream of information seems to be enough for the simple computational algorithms employed here to make some sense of the data.

I conclude this section with one final note. As shown in Table 5, better agreement with human coders was achieved when the interviewer utterances were included in the transcript. However, I am not ready to conclude that these analyses are best performed by including interviewer utterances along with student utterances. As I will discuss in Section 8, the inclusion of the interviewer was just one of a very large number of parameters that I was able to vary in performing the analysis. Across the broad space of parameter choices, including the interviewer sometimes led to better results, and sometimes to worse results. More generally, it was difficult to draw general conclusions about what choices of parameters lead to better results. In the concluding section of this manuscript, I will discuss what this observation implies for the larger endeavor.

5 Analysis of segmented transcripts

To this point, my analysis of transcripts has been coarse, with each transcript associated with a single DMC code. In many ways, this is a great simplification. As noted above, the human analysts coded 4 or 5 of the 21 transcripts as involving a shift in DMC. These were cases in which the interviewee clearly espoused multiple explanations of the seasons. Furthermore, even when the human analysts could assign a single code to a transcript, this single code frequently hid a great deal more complexity. For example, across the corpus of interviews, there was frequently an initial phase in which a student would consider multiple ideas, or struggle, before settling on an explanation.

<table>
<thead>
<tr>
<th>Interv.</th>
<th>First thing I want to know here is why is it warmer in the summer and colder in the winter?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leslie</td>
<td>Um::: well, um, you know times savings? // you know? Like in the summer, when you have –</td>
</tr>
<tr>
<td>Interv.</td>
<td>// mm-hmm</td>
</tr>
<tr>
<td>Interv.</td>
<td>– Daylight savings time?</td>
</tr>
<tr>
<td>Leslie</td>
<td>Yeah, daylight savings time. Um, in the summer we have more time, like, with, like, daylight and that’s why it gets warmer. And like just with the circulation of the earth and like the axis that it’s on just has to do with like summer and winter. And it depends on where we are on the earth. Like if you look at, umm, India, it’s like toward the equator, you know? And so it’s like always hot. And like if you go up north then it gets colder because there’s just, like, I can’t really say less sun, but it kind of has to do with that and there’s just a lot of snow and, like ice cause it’s colder up there.</td>
</tr>
</tbody>
</table>

*Figure 14. First part of interview with Leslie.*

For illustration, Figure 14 shows the first few moments from the interview with Leslie. Leslie was coded as side-based by both of the human analysts, as well as by the computational analysis described in Section 4. And, indeed, Leslie did seem to settle relatively quickly on a side-based explanation. However, if we look closely at the first few moments of her interview, we see Leslie considering multiple ideas, perhaps in an attempt to figure out what she knows that might be relevant. For example, as we can see in the partial transcript in Figure 14, she mentions daylight savings time, the “circulation” of the earth, and the climate in India. However, most of these ideas ultimately played no role in the explanation she constructed.
This sort of behavior is exactly what we would expect from a KiP perspective; in this perspective, we expect that, in general, DMCs will be constructed out of more basic knowledge.

I believe that if the computational techniques described in this paper are going to be viable as tools for researchers in commonsense science, then they will need to be able to capture these dynamics that exist at a finer timescale. That is the purpose of this section; I will extend the second analysis reported by Dam and Kaufmann (2008), in which they broke transcript documents into a few segments, and separately coded each of those segments.

In setting out to extend Dam and Kaufmann’s segmenting analysis, I had hoped to be able to code some students as shift, in a manner that aligned with the work of the human analysts. Furthermore, in the cases where students shifted, I hoped to be able to code the multiple DMCs constructed in a manner that aligned with the work of human analysts. However, I have not yet succeeded in achieving either of these goals. Nonetheless, the results to be described in this section do suggest that it is possible to capture information at a finer timescale using the LSA-based techniques.

5.1 Segmenting Jill

I begin my discussion of the segmenting analyses by returning to the interview with Jill. As discussed above, human analysts coded Jill as constructing a closer-farther DMC, and this is consistent with the result produced using the LSA-based techniques presented in Section 4. I now want to break Jill’s transcript into segments. As a first step, I break Jill’s transcripts into overlapping, 100-word segments as follows: I begin with a window spanning words 1 through 100 in the transcript, and this window steps forward 25 words at a time. So the first segment has words 1-100, the second 26-125, the third 51-150, and so on. This results in 8 separate documents. When each of these 8 segments is coded using the technique described in Section 4, the results are as shown in Figure 15; all 8 of the 100-word segments, from the beginning to the end of the transcript, are clearly coded as most similar to the closer-farther idealized answer.

![Figure 15](image)

Figure 15. Results when Jill’s transcript is cut into 100-word segments with a moving window that steps by 25 words.

What happens if Jill’s transcript is divided into even smaller segments? The results for some smaller window and step sizes are shown in Figure 16. Note that, even for segments as small as

---

8 The use of overlapping segments is intended to address one particular type of difficulty. When segmenting a transcript based solely on the number of words in a segment, we are making relatively arbitrary cuts. Depending where these cuts fall, we could miss important features of the data. Suppose, for example, that a student first speaks in a way consistent with side-based explanation, then moves briefly to a tilt-based explanation, and finally returns to a side-based explanation for the remainder of the interview. If a segment break falls in the middle of the tilt-based explanation, so that part of it is grouped with the preceding side-based explanation, and part with the following side-based explanation, then the tilt-based explanation might be missed. The use of overlapping segments is intended to ameliorate this difficulty.
10 words, the black bars (closer-farther) clearly predominate. This is an encouraging result. It suggests that the LSA-based computational techniques are able to extract useful information even for relatively short segments of text from the transcripts.

![Figure 16. Jill's transcripts coded in smaller and smaller segments.]

### 5.2 Segmenting other students

Thus far, we have seen that both the human and computational analyses coded Jill as consistently giving answers most similar to a single DMC. It is perhaps more interesting to look at interviews in which students exhibited greater variability in the DMCs they constructed. Some example analyses are shown in Figure 17. For all of these analyses, the transcripts were broken into 100 word segments, with a moving window that steps forward in 25-word increments.

Let us begin by looking at the analysis of Leslie. As discussed above, both the human analysts, and the computation analysis presented in Section 4, coded Leslie as side-based. However, when we looked at the first moments of Leslie’s interview, we saw that there was an initial phase during which Leslie was searching through her knowledge, working on constructing an explanation. Looking at the analysis of Leslie in Figure 17, we see that during about the latter two-thirds of the interview, the white bars dominate, consistent with a side-based DMC. However, there is a transitory initial phase, during which the black bars dominate (closer-farther). Furthermore, from my inspection of the transcript, the transition to side-based does seem to occur around the time that Leslie settled on her side-based explanation.
Figure 17. Segmented analyses for Leslie, Edgar, and Jacob. In each case, the transcript has been segmented into 100-word segments, with a window that steps forward 25 words.

Thus, once again, the results suggest that the computational techniques are able to capture information at a finer grain-size. However, the results also suggest some reasons for caution. Looking at Figure 17 might lead us to say that Leslie initially gave a closer-farther explanation, and then shifted to a side-based explanation. However, this is not in agreement with inspection of the early part of the transcripts by human analysts; in the initial phase, Leslie seems to be searching for an explanation, there is not a clear indication that she had settled, even briefly, on a closer-farther explanation.

Furthermore, suppose that we attempted to use the computational techniques simply to code whether a student had shifted explanations, and not necessarily what the DMCs were before or after the shift. In that case, we might be tempted to say that Leslie had shifted between two DMCs. However, again, this is not in agreement with the work of human analysts. Looking back, it is evident that our human coders tended to ignore portions of interviews in settling on a single code for a transcript. For example, transitory initial phases tended to be ignored, as did portions of transcripts that were clearly digressions, in which the interviewer and student briefly discussed another topic. Thus, if we want to use our automated techniques to code shifts in a manner that is analogous to human analysts, then it will be necessary to develop a means of ignoring some portions of a transcript.
Figure 17 includes analyses for two additional students, Edgar and Jacob. Edgar is an interesting case to consider because both human analysts coded Edgar as shifting explanations. Furthermore, although this was not captured in our formal coding, both analysts believed that Edgar shifted from a side-based explanation to a closer-farther explanation. As shown in Figure 17, the computational analysis also indicates a shift, and from my inspection of the transcript it appears that this shift occurs at roughly the same time identified by human analysts. However, the LSA-based analysis in Figure 17 shows Edgar as shifting from grey (tilt-based) to black (closer-farther), rather than from white (side-based) to black. So, as with Leslie, it seems that the analysis is sensitive to fine-grained information in the transcript, including some sensitivity to shifts. However, the match with human coders for the specific DMCs exhibited around shifts is far from perfect.

Some analyses of segmented transcripts are more difficult to interpret. This is the case for the third example in Figure 17, drawn from the interview with Jacob. The human analysts coded Jacob as side-based, as did the LSA-based analysis in Section 4. But in Figure 17 it appears that the white bars dominate only during the middle part of the interview. Looking at the complete transcript, this result is not too big a surprise. The interview with Jacob was longer than the others presented so far (as is evident from the fact that it consists of a larger number of segments). Furthermore, the interviewer took more liberties, and the interview ranged over somewhat broader territory.

5.3 Discussion
As stated above, I believe that, if these computational techniques are going to be viable tools for research on commonsense science, then they will need to be able to capture dynamics at a finer timescale. In this regard, there are some respects in which the first foray described in this section falls short. I had hoped, first, to be able to use these techniques simply to code whether a student had shifted DMCs during an interview. Furthermore, in the cases where students had shifted, I hoped to be able to code the multiple DMCs constructed in a manner that aligned with the impressions of human analysts. However, I was not able to develop an algorithm that was capable even of meeting the more modest goal of coding whether a shift had occurred.

Still, I believe that the results discussed in this section provide reasons to be optimistic that these techniques can be adapted to capture the finer timescale information that is my goal. We saw that, at least in some cases, the LSA-based analyses seemed to produce meaningful results with very short segments of text – as little as ten words. Prior to performing this analysis, it was not at all obvious that this would be possible. And where there were visible shifts in the output of the computational analysis, these did seem to align with important transition points in the transcript.

There are actually some good reasons to think that a somewhat different approach should be needed to capture finer timescale dynamics. Note that, when attempting to code individual segments of transcripts, I compared the vectors for those segments to vectors for the idealized response documents. These idealized vectors were intended to be associated with complete explanations of the seasons – a full DMC. However, it is not clear that this is an appropriate way to code segments. Even if a student, throughout an interview, answers in a manner that is consistent with a single DMC, there is no reason to expect that every segment of that interview should have a vector that encompasses the entire DMC. It might, instead, express a component that is just one piece of the DMC. Thus, it might not make sense to code small segments of transcripts in terms of one of our DMCs.
Furthermore, the same component may very well be common to more than one DMC. For example, every explanation given by students depended, in one way or another, on the fact that the sun is the earth’s source of heat. And most explanations mentioned the fact that the earth orbits the sun. Thus, the ideas mentioned in a brief segment might not allow us to reliably distinguish which explanation is being given by a student.

Given these observations, it is clear that the approach used in this section was probably overly crude. In fact, it is perhaps a bit surprising that our attempt to code segments in terms of DMCs worked as well as it did. If we are serious about capturing dynamics at a finer grain size, we will likely need a different approach, one that somehow attempts code segments in terms of a larger vocabulary of possible meanings. I will offer some attempts in this direction in Section 7.

6 Inducing a coding scheme by clustering transcript documents

In the preceding sections, my analyses all depended on the coding scheme developed by human analysts, which was embodied in the idealized response documents. In taking that approach, I was following the program of analysis developed by Dam and Kaufmann (2008). Furthermore, this is, broadly speaking, the approach that has been adopted in the vast majority of applications of LSA in education. As discussed in Section 2.3, the vast majority of those applications have involved the comparison of text to a reference document constructed or selected by the researcher.

![Figure 18. Overview of the clustering-based analysis.](image)

In this section, I will show how it is possible, within the confines of the present work, to eliminate the need for idealized response documents, or any sort of reference documents. The computational analyses described here instead discover the coding scheme by finding natural clusters in the data. This does more than simply save researchers work; it makes it possible to produce an alternative validation of researcher-devised coding schemes.
An overview of the new clustering-based analysis is presented, in schematic form, in Figure 18. In this new analysis, the processing of the training corpus is unchanged, and the transcript vectors are computed from transcript documents precisely as before. But instead of comparing the transcript vectors to idealized response vectors, the transcript vectors are fed into cluster analysis. The idea behind the cluster analysis is conceptually fairly simple. Each of the transcript vectors can be understood as associated with a point in a 200-dimensional space (corresponding to the tip of the vector). The cluster analysis seeks to find groups of these points that are near to each other—in short, it looks for clusters.

6.1 Hierarchical agglomerative clustering

To cluster the transcript vectors, I employed the very general technique called hierarchical agglomerative clustering (HAC). In keeping with the spirit of my overall presentation, I briefly describe HAC here, so that more readers can follow and judge the viability of the techniques described in this paper. In HAC, we begin by taking all of the items to be clustered, and placing each of these items in its own cluster. Thus, we begin with a number of clusters equal to the total number of items. Then we pick two of those clusters to combine into a single cluster containing two items, thus reducing the total number of clusters by one. The process then iterates; we again pick two clusters to combine, and the total number of clusters is decreased by one. This repeats until all of the items are combined into a single cluster. The result is a list of candidate clusterings of the data, with each candidate corresponding to one of the intermediate steps in this process.

A central issue in applying this algorithm is determining which clusters to combine on each iteration. In practice, there are many rules that can be applied. Throughout my discussions here, I will present results that were obtained using a technique called centroid clustering. At each step in the iteration, I first find the centroid of each cluster—the average of all of the vectors currently in the cluster. Then I find the pair of centroids that are closest to each other, and merge the associated clusters. An explanation of centroid clustering, including its application to LSA-produced vectors, can be found in Manning, Raghavan, and Schütze (2008).

My description of the clustering algorithm in the preceding two paragraphs glosses over one important detail. Prior to clustering the transcript vectors, I replace the entire set of transcript vectors with their deviation vectors, using the procedure described in Section 4.3.

Figure 19 shows the sequence of candidate clusterings that are generated when the hierarchical clustering algorithm is applied to the 21 transcript vectors. At the top of the figure, all of the students are in individual clusters. In each row that follows, a pair of clusters has been combined. (The newly formed cluster in each row is highlighted.) In the bottom row there are just three clusters, containing 9, 9 and 3 students respectively.

In each row, clusters contain documents that have been grouped together because, from the point of view of LSA, they have similar meanings. This means that each row in Figure 19 constitutes a candidate coding scheme—it is a scheme for sorting transcripts into categories. The puzzle, of course, is which row to select. One possibility is to select the bottom row, simply because our human analysts sorted the transcripts into three categories. But since we want to be able to

---

9 As I will discuss in Section 8, I explored the use of multiple clustering algorithms.
10 The clustering procedures do not seem to give meaningful results if this is not done.
**discover** the coding scheme, it would be preferable if we could computationally determine the appropriate number of codes. This is a problem I have not yet solved. In what follows, I will simply select a few rows from Figure 19 to discuss. In Section 8.4 I will return to the question of how the selection might be made computationally.

**Figure 19. Results of hierarchical aggregate clustering.**
6.2 Examination of the three-cluster set

I begin by jumping directly to the bottom row in Figure 19, the row in which the transcripts have been sorted into three clusters. A question we would like to answer is whether the coding implied by this clustering of the transcripts is in agreement with the coding of human analysts. However, at this point, each of the three clusters is identified solely by the list of students that comprise the cluster. In order to make further progress, we need some way to name the clusters or otherwise give them interpretations that we can compare to the human coding. I have tried several approaches to solve this problem. Here I will discuss two.

In the first approach, I begin by finding the centroid vector of each of the three clusters, by averaging the vectors for the students that appear in that cluster. I then compare each of these centroid vectors to the vector for every one of the 12,493 words that appears in the Word Space, looking for the ten words that are most similar to the centroid vectors. When this is done, I obtain the lists of words shown in Figure 20. For each of the three clusters – which I will now call Cluster 3-1, Cluster 3-2, and Cluster 3-3 – I have listed the 10 words that are closest in meaning to the centroid, with the most similar words at the top. In addition, for each word, the tables include the dot product of the word with the cluster centroid and the number of times that the word appears in the training corpus.

<table>
<thead>
<tr>
<th>Cluster 3-1</th>
<th>Cluster 3-2</th>
<th>Cluster 3-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>tilted</td>
<td>closer</td>
<td>side</td>
</tr>
<tr>
<td>0.485305</td>
<td>0.453565</td>
<td>0.493785</td>
</tr>
<tr>
<td>away</td>
<td>summer</td>
<td>day</td>
</tr>
<tr>
<td>0.419981</td>
<td>0.333442</td>
<td>0.399848</td>
</tr>
<tr>
<td>kind</td>
<td>winter</td>
<td>night</td>
</tr>
<tr>
<td>0.329259</td>
<td>0.314201</td>
<td>0.394530</td>
</tr>
<tr>
<td>meteoroids</td>
<td>truth</td>
<td>assistant</td>
</tr>
<tr>
<td>0.285616</td>
<td>0.240860</td>
<td>0.306860</td>
</tr>
<tr>
<td>towards</td>
<td>habitation</td>
<td>janitor</td>
</tr>
<tr>
<td>0.283286</td>
<td>0.230071</td>
<td>0.306860</td>
</tr>
<tr>
<td>deus</td>
<td>otal</td>
<td>inquisitive</td>
</tr>
<tr>
<td>0.278318</td>
<td>0.228579</td>
<td>0.303723</td>
</tr>
<tr>
<td>machina</td>
<td>mudslides</td>
<td>train</td>
</tr>
<tr>
<td>0.278318</td>
<td>0.227996</td>
<td>0.301719</td>
</tr>
<tr>
<td>card</td>
<td>actuality</td>
<td>staff</td>
</tr>
<tr>
<td>0.272152</td>
<td>0.219875</td>
<td>0.298649</td>
</tr>
<tr>
<td>stuff</td>
<td>northeasterly</td>
<td>time</td>
</tr>
<tr>
<td>0.262092</td>
<td>0.215121</td>
<td>0.298600</td>
</tr>
<tr>
<td>mentioned</td>
<td>brings</td>
<td>aptitude</td>
</tr>
<tr>
<td>0.249782</td>
<td>0.214194</td>
<td>0.292909</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 20. Labels for the set of three clusters. Each table lists the ten words most similar to the centroid of the cluster, followed by the dot product between the word vector and centroid vector, and the number of times that the word appears in the training corpus.

The lists in Figure 20 contain some words, such as “tilted” and “summer,” that clearly have to do with explanations of the seasons. However, some of the words, such as “janitor” and “card” seem to be out of place. Note that these out-of-place words tend to occur much less frequently in the training corpus. For example, “janitor” appears only once in the entire corpus and “card” only twice. The training corpus does not contain much information about these infrequently-occurring words. Thus, it is reasonable to restrict the lists in Figure 20 so that they only include words that appear in the training corpus above some minimum number of times. If we set the minimum threshold at 10 words, then we get the new lists shown in Figure 21. Although there are still a few odd words, the majority of the words in the lists now appear to bear some relation to explanations of the seasons.

The results in Figure 21 are clearly very suggestive: Cluster 3-1 has “tilted” as its top word, Cluster 3-2 has “closer,” and Cluster 3-3 has “side.” This suggests that we should align these clusters with the tilt-based, closer-farther, and side-based DMCs, respectively. The other words

---

11 This makes use of capabilities that are built into InfoMap.
that appear in each list also seem to generally support this interpretation. For example, the words “away” and “towards” are high up in the list for Cluster 3-1. We can imagine these words appearing in a tilt-based explanation in which the earth tilts towards the sun in the summer and away in the winter.

There are also notable words in Cluster 3-3. The words “rotate” and “rotation” appear on this list, which makes sense if this cluster is associated with side-based explanations; the rotational motion of the earth is central to most side-based explanations. Note also that the words “day” and “night” appear near the top of the list. These words might be there because, when students proposed side-based explanations, the discussion often drifted into discussion of the day/night cycle. Furthermore, explanation of the day/night cycle might be associated, in the training corpus, with discussion of the earth’s rotation.

So the words that appear in each of the lists in Figure 21 are consistent with an interpretation of the three clusters in terms of our three idealized DMCs. It is also worth noting that the lists in Figure 21 are disjoint; no word appears on more than one of the three lists. For example, it is not only the case that the word “tilted” appears at the top of the list for Cluster 3-1, it does not appear anywhere on the other two lists. We can also look at the dot product of individual words, such as “tilted,” and the centroids of other clusters. For example, the dot product of the word “tilted” with the centroids of clusters 3-2 and 3-1 are -.4556 and -.1347 respectively. Thus, it is not only the case that the word “tilted” is closely aligned with the centroid of Cluster 3-1; it is very much not in alignment with the other two clusters.

I want to take one last step to cement our interpretation of these three clusters. In Table 6 and Figure 22, I show the results of comparing the three centroid vectors to the vectors derived from the idealized response documents. The results clearly align with the interpretations suggested by the lists in Figure 21: Cluster 3-1 aligns with tilt-based, Cluster 3-2 with closer-farther, and Cluster 3-3 with side-based. This second technique for identifying the meaning of clusters is less desirable than the first technique, since it relies on the idealized response documents, rather than

<table>
<thead>
<tr>
<th>Cluster 3-1</th>
<th>Cluster 3-2</th>
<th>Cluster 3-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>tilted</td>
<td>0.485305</td>
<td>0.453565</td>
</tr>
<tr>
<td>away</td>
<td>0.419981</td>
<td>0.333442</td>
</tr>
<tr>
<td>kind</td>
<td>0.329259</td>
<td>0.314201</td>
</tr>
<tr>
<td>towards</td>
<td>0.283286</td>
<td>0.214194</td>
</tr>
<tr>
<td>mentioned</td>
<td>0.249782</td>
<td>0.199429</td>
</tr>
<tr>
<td>angles</td>
<td>0.220155</td>
<td>0.199373</td>
</tr>
<tr>
<td>facing</td>
<td>0.2196</td>
<td>0.194245</td>
</tr>
<tr>
<td>hemisphere</td>
<td>0.219496</td>
<td>0.193123</td>
</tr>
<tr>
<td>axis</td>
<td>0.21242</td>
<td>0.191793</td>
</tr>
<tr>
<td>incident</td>
<td>0.212335</td>
<td>0.191387</td>
</tr>
</tbody>
</table>

Figure 21. Labels for the set of three clusters, now restricted to words that appear at least 10 times in the training corpus.
discovering DMCs only from the data. Nonetheless, the use of this technique here can help build our confidence in the first method.

<table>
<thead>
<tr>
<th></th>
<th>Cluster 3-1</th>
<th>Cluster 3-2</th>
<th>Cluster 3-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closer-Farther</td>
<td>-0.473</td>
<td>0.557</td>
<td>-0.104</td>
</tr>
<tr>
<td>Side-Based</td>
<td>0.037</td>
<td>-0.260</td>
<td>0.416</td>
</tr>
<tr>
<td>Tilt-Based</td>
<td>0.381</td>
<td>-0.276</td>
<td>-0.242</td>
</tr>
</tbody>
</table>

*Table 6. Dot products of the three cluster centroids with the idealized answer vectors.*

Finally, if we accept these interpretations of the three clusters, we can compare the results of the clustering analysis to the work of human coders. As shown in Table 7, the results agree with human coders for 10 students, and disagree for 6, with a Cohen’s Kappa of .46. This same clustering analysis can also be reproduced, in its entirety, using transcripts in which interviewer utterances are included. As shown in Table 7, this produces results that are slightly better.

The results in Table 7 perhaps deserve to be called “promising,” but they are not quite as good as those obtained using the idealized answer documents. At this point, it appears that the ability of the cluster analysis to derive a coding scheme is more impressive than its ability to assign codes to individual transcripts.

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student only transcript</td>
<td>10</td>
<td>6</td>
<td>0.46 (moderate)</td>
</tr>
<tr>
<td>Student+Interviewer</td>
<td>12</td>
<td>4</td>
<td>0.63 (good)</td>
</tr>
</tbody>
</table>

*Table 7. Comparison of clustering results to human coders.*

6.3 Examination of the four and six-cluster sets

Thus far, I have only looked at the bottom row of Figure 19, where the students were grouped into three clusters. I now want to look at some of the higher rows in Figure 19, in which students are grouped into a larger number of clusters. Here, I am interested in looking at whether the additional clusters that appear in these rows are suggestive of DMCs not captured by the three codes in the scheme developed by human analysts. In this discussion, I will work my way up from the bottom of the figure to the top, splitting clusters rather than merging them. This is, of course, the reverse of the process by which Figure 19 was constructed. The changes that occur as
we move up the bottom rows of Figure 19 are shown in Figure 23 (with the five-cluster row omitted).

I begin by looking one row up the figure, at the four-cluster set. Note that, in moving from the three-cluster to the four-cluster set, two clusters must remain unchanged, and one must be split into two clusters. As depicted in Figure 23, the two that remain the same are Cluster 3-2 (closer-farther) and Cluster 3-3 (side-based), and these become Cluster 4-3 and Cluster 4-4.

In contrast, the tilt-based cluster (3-1) has been split into two clusters (4-1 and 4-2). The only question that remains is what interpretation can be attributed to these new clusters. If I perform an analysis analogous to the one performed for the three-cluster set, I obtain the results shown in Figure 24 and Figure 25. Looking at Figure 25, it appears that Cluster 4-2 is still a tilt-based cluster. The dot product of this cluster with the tilt-based idealized answer is even higher than for Cluster 3-1 (compare Figure 25 to Figure 22). Thus, in some respects we have an even better tilt-based cluster. The list of words associated with Cluster 4-2 also are very much consistent with a tilt-based explanation (refer to Figure 24).

Cluster 4-1 is a different story however. Looking at Figure 25, it is clear that this cluster is not closely aligned to any of the idealized answers; all of the dot products are small. This suggests
that, if the cluster is picking out anything meaningful, it is a different meaning than is captured by any of the idealized answers. Based on the list of words in Figure 24, we can make some tentative guesses. Many of the words here – planets, planetary, and ecliptic – seem to pertain to the motion of planets in the solar system and are not necessarily associated with any explanation of the seasons. Indeed, students often did digress to talk about the motion of the planets in the solar system. However, we need to be very careful in constructing interpretations of this sort. We could perhaps construct plausible-sounding interpretations for just about any list constructed from common words in the training corpus. Some sort of principled restraint is required.

$$\begin{array}{|c|c|c|}
\hline
\text{Cluster 6-1} & \text{Cluster 6-2} & \text{Cluster 6-3} \\
\hline
\text{kind} & 0.486337 & 21 \\
\text{planets} & 0.256781 & 294 \\
\text{axis} & 0.249392 & 524 \\
\text{motions} & 0.245475 & 37 \\
\text{away} & 0.240321 & 248 \\
\text{movement} & 0.231895 & 60 \\
\text{ecliptic} & 0.223469 & 65 \\
\text{learned} & 0.216245 & 11 \\
\text{objects} & 0.214165 & 50 \\
\text{planetary} & 0.198111 & 80 \\
\hline
\text{Cluster 6-4} & \text{Cluster 6-5} & \text{Cluster 6-6} \\
\hline
\text{closer} & 0.446613 & 79 \\
\text{summer} & 0.352215 & 623 \\
\text{winter} & 0.318982 & 622 \\
\text{farther} & 0.226172 & 52 \\
\text{northwest} & 0.221865 & 10 \\
\text{perfect} & 0.219059 & 18 \\
\text{eccentricity} & 0.205018 & 91 \\
\text{ellipse} & 0.187969 & 25 \\
\text{reaches} & 0.184083 & 79 \\
\text{summers} & 0.180285 & 54 \\
\hline
\text{Cluster 6-2} & \text{Cluster 6-3} & \text{Cluster 6-6} \\
\hline
\text{tilted} & 0.666371 & 239 \\
\text{towards} & 0.374110 & 91 \\
\text{perpendicular} & 0.294423 & 62 \\
\text{striking} & 0.287892 & 21 \\
\text{hitting} & 0.287467 & 20 \\
\text{top} & 0.271488 & 136 \\
\text{tilts} & 0.260443 & 25 \\
\text{receives} & 0.254229 & 67 \\
\text{sunlight} & 0.246744 & 250 \\
\text{hit} & 0.243595 & 42 \\
\text{takes} & 0.443101 & 71 \\
\text{starts} & 0.347549 & 14 \\
\text{revolution} & 0.308196 & 70 \\
\text{complete} & 0.257569 & 44 \\
\text{time} & 0.242780 & 589 \\
\text{weaker} & 0.237101 & 10 \\
\text{reach} & 0.235824 & 70 \\
\text{wintertime} & 0.234189 & 13 \\
\text{zenith} & 0.232355 & 14 \\
\text{receiving} & 0.232154 & 12 \\
\text{side} & 0.493785 & 132 \\
\text{day} & 0.399848 & 550 \\
\text{night} & 0.394530 & 199 \\
\text{time} & 0.298600 & 589 \\
\text{moon} & 0.288719 & 234 \\
\text{rotates} & 0.271162 & 80 \\
\text{poster} & 0.263904 & 10 \\
\text{lunar} & 0.240465 & 15 \\
\text{shadow} & 0.229902 & 58 \\
\text{rotation} & 0.226524 & 258 \\
\hline
\end{array}$$

Figure 26. Labels for the set of six clusters.

I want to look at one more row from Figure 19. For this last analysis, I will jump a little farther up the table to the six-cluster set, so we can begin to see what happens as we group transcripts into a larger number of clusters. The results for the six-cluster set are captured in Figure 26 and Figure 27. In moving from the four-cluster to the six-cluster set there have been two changes (refer to Figure 23). First, one student, Mark, has been split out into his own cluster, Cluster 6-5. The list of words closely associated with this cluster is not particularly suggestive. However, comparing the centroid of this cluster to the idealized answer vectors reveals a moderately strong similarity to the tilt-based idealized vector, which is consistent with the code assigned to Mark by human coders.\footnote{This, of course, is precisely how it should be. Since Cluster 6-5 only contains Mark, comparing its centroid to the idealized vectors is the same as comparing Mark’s transcript vector to the idealized vectors.}

The other change between the four-cluster and six-cluster set is a bit more interesting. Cluster 4-2 – the one associated with a tilt-based explanation – has been split into two clusters, Cluster 6-2 and 6-3. Cluster 6-2 is even more strongly aligned with the idealized tilt-based answer, and it has an extremely strong association with the word “tilted.” Cluster 6-3 is a bit more difficult to
interpret. It aligns most strongly with the side-based answer, though this alignment is weaker than for Cluster 6-6, which has been carried forward unchanged from the four-cluster set. The list of words associated with Cluster 6-3 is not immediately suggestive, though the words on this list seem to have to do with the earth’s hemispheres.

There are only two students in Cluster 6-3, and these two students were actually among the more difficult for our human coders. One of these students was Robbie, who gave an explanation in which the earth rotates, so that first one part of the earth faces the sun, then the other. This is consistent with the usual side-based explanation. However, in Robbie’s explanation, the earth rotated so that first one pole and then the other faced toward the sun. Thus, the hemispheres experienced seasons at different times. This suggests that Cluster 6-3 might be beginning to pick up real differences in explanations that simply were not captured by the coarse coding scheme embodied in the three idealized DMCs. Once again, however, we must be restrained in making such interpretations.

As with the three-cluster set, we can compare the analyses embodied in the four and six-cluster sets to the analysis produced by human coders. As shown in Table 8, the results are slightly better for these larger cluster sets, and they approach the results obtained using the idealized answers for comparison. The results for these same analyses, when interviewer utterances are included in the transcripts, are given in Table 9.

Table 8. Comparison of clustering results from the four and six-cluster sets to human coders, with transcripts that only include student utterances.

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four Clusters</td>
<td>11</td>
<td>5</td>
<td>0.54</td>
</tr>
<tr>
<td>Six Clusters</td>
<td>12</td>
<td>4</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 9. Comparison of clustering results from the four and six-cluster sets to human coders, with transcripts that include both student and interviewer utterances.

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four Clusters</td>
<td>12</td>
<td>4</td>
<td>0.63</td>
</tr>
<tr>
<td>Six Clusters</td>
<td>12</td>
<td>4</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Note that, in order to make the comparisons contained in Table 7, I needed to associate each cluster with one of the three idealized DMCs. In all cases, I simply associated the cluster with the idealized vector with which its
6.4 Discussion

In this section, I took steps toward more fully automating the coding of transcripts. By using clustering, I was able to simultaneously induce and apply a coding scheme. When we look at how individual transcripts are coded, and compare these codes to the work of human analysts, 10 to 12 out of the 16 transcripts are in agreement, with Kappa values of .5 or .6. This level of agreement is, roughly speaking, comparable to the results reported in Section 4, in the analysis that made use of idealized response documents. In that respect, the analysis reported in this section does not seem to have produced a new level of achievement.

However, I believe that the ability of the clustering analyses to induce a coding scheme is striking. It is striking, first, that the cluster analysis can find any type of meaningful clusters in the transcript vectors. But, still more striking is the fact that the cluster analysis seems to identify precisely the same groupings that were pulled out in our earlier human-based analysis of DMCs. There are many ways in which our human analysts could have chosen to group and code transcripts. But, it seems that the one that we did select aligns with the same features picked out by the LSA-based clustering analysis.

The techniques that I employed to interpret clusters show some promise. But there is clearly work to do. One of my techniques relied on the idealized response documents. Ultimately, we will want to fully wean ourselves of any dependence on the idealized responses. The other technique I described produced suggestive lists of words. But we still need to devise a disciplined way of interpreting these lists.

7 Clustering segments of transcripts

Ultimately, the knowledge-in-pieces perspective calls for an analysis that is substantially different than all of the analyses in the preceding section. As KiP researchers, we are not interested only in sorting the explanations that students give (i.e., the DMCs they construct) into categories. Instead, in the long run, what we want is an analysis that can get at the knowledge out of which DMCs are constructed. Before concluding this paper, I want to offer one small step in that direction.

The partial success of the analysis of segmented transcripts, described in Section 5, suggests that it might be possible to use a clustering analysis to find meanings in segments of transcripts. Thus, for this analysis, I began by breaking all of the 21 transcripts into overlapping segments, just as I did for the segmenting analysis presented in Section 5. Then I performed a cluster analysis on all of these segments, using the same techniques I used for clustering transcripts. The purpose of this analysis was to see if we can take these little snippets of transcripts and cluster them into something like smaller-scale ideas – the sort of ideas that might be closer to the elements of knowledge out of which DMCs are constructed.

To date, I have tried this with the transcripts segmented in two different ways. In one analysis, I broke the transcripts into 100-word segments, with a moving window that stepped forward by 25 words; in the other analysis, I used 25-word segments and a step size of 10. These two analyses

---

centroid had the highest dot product. This means, for example, that clusters 4-1 and 6-1 were treated as closer-farther, even though the associations were relatively weak.
produced extremely similar results. Here I will present only the results from the analysis that used 25-word segments.

When all of the transcripts were broken into 25-word segments, this resulted in a total of 609 segments. When LSA vectors were computed for these 609 segments, 3 segments produced vectors with a magnitude of zero (because they contained no words that appeared as rows in the Word Space). These three zero-magnitude segments were dropped from the analysis, leaving 606 segments to be clustered. As in the transcript clustering, the vectors associated with each segment were replaced by their deviation vectors prior to clustering.

When all of the transcripts were broken into 25-word segments, this resulted in a total of 609 segments. When LSA vectors were computed for these 609 segments, 3 segments produced vectors with a magnitude of zero (because they contained no words that appeared as rows in the Word Space). These three zero-magnitude segments were dropped from the analysis, leaving 606 segments to be clustered. As in the transcript clustering, the vectors associated with each segment were replaced by their deviation vectors prior to clustering.

![Figure 28. Moving from the three-cluster, to the four-cluster, to the six-cluster sets of segments.](image)

### 7.1 Examination of the three-, four-, and six-cluster sets

I will now look closely at the three-cluster, four-cluster, and six-cluster sets. When the 606 segments were grouped into three clusters, the clusters were comprised of 225, 153, and 228 segments (refer to Figure 28). To interpret the meaning of these clusters, I employed the same techniques used in Section 6. As shown in Figure 29 and Figure 30, the results are reminiscent of transcript clustering; the three clusters seem to each align with the three idealized DMCs. The list of words that are associated with Cluster 3-2 are evocative of a tilt-based explanation, and this cluster does have a strong alignment with the tilt-based idealized response vector. However, the other two lists of words are less easy to interpret. Cluster 3-1 is most similar to the closer-farther answer document, but the list of words seems to pertain to the orbit of the earth around the sun, and not to any particular explanation of the seasons. Finally, Cluster 3-3 has the weakest alignment to any of the idealized answer documents, and it is difficult to see how the list of words might be associated with a side-based explanation, though the words “day” and “night” are on the list.

### Table 28: Words Associated with Each Cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Words</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-1</td>
<td>sun</td>
<td>0.557854</td>
</tr>
<tr>
<td></td>
<td>orbit</td>
<td>0.315278</td>
</tr>
<tr>
<td></td>
<td>earth</td>
<td>0.286494</td>
</tr>
<tr>
<td></td>
<td>~</td>
<td>0.196327</td>
</tr>
<tr>
<td></td>
<td>solar</td>
<td>0.179732</td>
</tr>
<tr>
<td></td>
<td>ellipse</td>
<td>0.176945</td>
</tr>
<tr>
<td></td>
<td>distances</td>
<td>0.168432</td>
</tr>
<tr>
<td></td>
<td>celestial</td>
<td>0.166542</td>
</tr>
<tr>
<td></td>
<td>evening</td>
<td>0.165143</td>
</tr>
<tr>
<td></td>
<td>morning</td>
<td>0.164609</td>
</tr>
<tr>
<td>3-2</td>
<td>tilted</td>
<td>0.475581</td>
</tr>
<tr>
<td></td>
<td>kind</td>
<td>0.460649</td>
</tr>
<tr>
<td></td>
<td>striking</td>
<td>0.349523</td>
</tr>
<tr>
<td></td>
<td>concentrated</td>
<td>0.304906</td>
</tr>
<tr>
<td></td>
<td>towards</td>
<td>0.280476</td>
</tr>
<tr>
<td></td>
<td>sunlight</td>
<td>0.279226</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>0.278318</td>
</tr>
<tr>
<td></td>
<td>hotter</td>
<td>0.267920</td>
</tr>
<tr>
<td></td>
<td>hit</td>
<td>0.257450</td>
</tr>
<tr>
<td></td>
<td>angle</td>
<td>0.254329</td>
</tr>
<tr>
<td>3-3</td>
<td>winter</td>
<td>0.668527</td>
</tr>
<tr>
<td></td>
<td>summer</td>
<td>0.624533</td>
</tr>
<tr>
<td></td>
<td>hemisphere</td>
<td>0.312125</td>
</tr>
<tr>
<td></td>
<td>southern</td>
<td>0.291326</td>
</tr>
<tr>
<td></td>
<td>northwest</td>
<td>0.287605</td>
</tr>
<tr>
<td></td>
<td>northern</td>
<td>0.269531</td>
</tr>
<tr>
<td></td>
<td>day</td>
<td>0.265189</td>
</tr>
<tr>
<td></td>
<td>autumn</td>
<td>0.239734</td>
</tr>
<tr>
<td></td>
<td>night</td>
<td>0.229601</td>
</tr>
<tr>
<td></td>
<td>brings</td>
<td>0.208714</td>
</tr>
</tbody>
</table>

![Figure 29. Labels for the set of three clusters.](image)
Figure 30. Dot products of the three cluster centroids with the idealized answer vectors.

None of these results are necessarily problematic or even surprising. There is no reason to expect that, when transcripts are broken into small segments and then clustered, the resulting clusters would align with DMCs. Indeed, the hope was that these small fragments of transcripts might be found to express alternative, smaller-scale ideas that comprise DMCs.

Figure 31. Labels for the set of four clusters.

I will now move on to a discussion of the four-cluster set. In moving from three to four clusters, Cluster 3-3 – the one most similar to a side-based DMC – splits into two clusters comprised of 142 and 86 segments (refer to Figure 28). As shown in Figure 32, one of the new clusters, Cluster 4-4, is much more similar to the side-based idealized answer document than Cluster 3-3. Furthermore, looking at Figure 31, we can see that the list of words associated with this cluster is now much more evocative of a side-based DMC. Thus, it seems that cluster 3-3 did contain the nucleus of a side-based cluster, but masked by the 142 segments that were split into cluster 4-3.

The other new cluster, Cluster 4-3, has very small dot products with all of the idealized answer documents, and the associated list of words has no obvious interpretation in terms of our established coding categories. However, the words “winter” and “summer” have very strong associations with this cluster. In addition, words pertaining to the northern and southern
hemisphere are high on the list. What we might be seeing here is talk about winter and summer, with an emphasis on the fact that these seasons occur at different times in the northern and southern hemisphere.

I now jump to the six-cluster set, which consists of clusters comprised of 197, 28, 110, 43, 142, and 86 segments. The changes as we move from the four-cluster to the six-cluster set are shown in Figure 28. In moving from the four-cluster to the six-cluster set, two clusters, 4-3 and 4-4, remain unchanged and become clusters 6-5 and 6-6. (Cluster 4-4 was the side-based cluster and 4-3 was the winter/summer and hemispheres cluster.)

![Figure 33. Labels for the set of six clusters.](image)

![Figure 34. Dot products of the six-cluster centroids with the idealized answer vectors.](image)

One of the clusters that changes is Cluster 4-1, which splits into clusters 6-1 and 6-2. Looking at Figure 33 and Figure 34, we can see that the results for Cluster 6-1 are extremely similar to Cluster 4-1 (which was identical to 3-1); it aligns most closely with the closer-farther idealized explanation, but with an emphasis on the orbital motion of the earth. Cluster 6-2, in contrast, appears to be something very new; it seems to be strongly associated with words having to do
with *time*. Note that this cluster is comprised of a small number of transcripts segments (28) compared to the other clusters.

Finally, Cluster 4-2 splits into Clusters 6-3 and 6-4. Like Cluster 4-2, 6-3 is strongly tilt-based, and is associated with words that strongly evoke a tilt-based DMC. Cluster 6-4, in contrast, is very difficult to interpret. Like Cluster 6-2, this cluster has a smaller number of segments.

To further give the reader a sense for what is going on in the six-cluster analysis, Appendix A has a table of 25-word segments drawn from this analysis. For each of the six clusters I have randomly selected 10 segments. The interested reader might want to peruse this list to get a sense for the contents of the segments that populate each cluster.

### 7.2 Using these clusters as the basis of a coding scheme

I want to briefly describe one final kind of analysis, in order to give a sense for how we might produce one type of analysis that is more in keeping with a KiP perspective. Once we have clusters of segments from the analyses described above, there are a variety of potential uses for these clusters. One interesting possibility is to use the centroids of clusters from the above analyses as comparison vectors, in place of the idealized response vectors used in my earlier analyses. We could use these centroids to code entire transcripts (as in Section 4) or to code segments of transcripts (as in Section 5). To give a sense for the possibilities here, I will just present some of the results of using cluster centroids to code entire transcripts.

![Figure 35. Analysis using clusters 3-1, 3-3, and 3-2 as comparison vectors.](image)

To start, let us use the centroids of clusters 3-1, 3-2, and 3-3 as a basis of comparison. In line with my earlier discussion, we will think of Cluster 3-1 as closer-farther, 3-2 as tilt-based, and 3-3 as side-based. When each of the 21 transcript vectors are compared to these centroids, we obtain the results in Figure 35. As usual, the black bar is closer-farther, the white side-based, and the gray tilt-based. Comparing these results to the work of human analysts, I find that the analysis agrees for 9 students, and disagrees for 7 (refer to Table 10).

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three cluster set</td>
<td>9</td>
<td>7</td>
<td>0.32</td>
</tr>
<tr>
<td>Six cluster set</td>
<td>12</td>
<td>4</td>
<td>0.62</td>
</tr>
</tbody>
</table>

*Table 10. Results when cluster centroids are used to code the 21 transcripts.*

Next, we can perform the same type of analysis, but using the three clusters from the six-cluster set that are most closely aligned with the idealized answer vectors: Cluster 6-1 (closer-farther),
Cluster 6-3 (tilt-based), and Cluster 6-6 (side-based). The results of this analysis are shown in Figure 36. As summarized in Table 10, this agrees with human analysts for 12 of the 16 students who are codable. Note that we are now approaching the level of agreement that was obtained using the analysis, described in Section 4, that made use of idealized answer documents.

![Figure 36. Analysis using clusters 6-1, 6-6, and 6-3 as comparison vectors.](image)

7.3 Discussion

This analysis started from 606 small snippets of text, and clustered them in a simple, brute-force manner. The most important – and I believe intriguing – result of this analysis is that, in many cases, the clusters produced by this analysis were readily interpretable.

One feature of the results merits additional attention. Recall that, in discussing the analysis of segmented transcripts in Section 5, I noted that it is surprising that segments of transcripts can be aligned with the DMCs embodied in my idealized response documents, since these DMCs were intended to capture the meaning of complete explanations, not little pieces of explanations. The same is true of the analyses in this section; I believe it is surprising that it is possible to align some clusters of transcript snippets with full DMCs. It seems that, in some cases, the full meaning of an explanation associated with a DMC is somehow carried within many of the small segments that comprise that explanation; it’s as if the full explanation is somehow built into the DNA of the little segments of a transcript’s text.

8 General discussion

I began this paper with the observation that research on commonsense science knowledge typically focuses on data derived from one-on-one clinical interviews. Furthermore, to date, researchers in this field have generally used humans as “instruments” for coding data. I believe that we have done so because of some tacitly-held beliefs: we have always assumed (at least before the work of Dam & Kaufmann, 2008) that, to code clinical interview data, it is necessary to have an instrument with an ability to understand natural language. I also believe we have tacitly assumed that deriving a coding scheme – rather than just applying one – is even farther out of reach; it would seem to require not only the ability to comprehend natural language, but also a deep understanding of the nature and purposes of the research, along with the ability to make leaps that look across the breadth of a data corpus.

And there are other apparent difficulties that I have mentioned previously. The language of students in interviews can be halting and ambiguous. Furthermore, student gestures and diagrams often play a prominent role, and it can be essential to pay attention to them in order to fully understand student utterances.
In an attempt to automate the coding of clinical interview data, I have explored the use of some techniques from Statistical NLP. In comparison to true NLP, these techniques are relatively simple to implement. Stated crudely, the statistical techniques rely primarily on “counting words,” and generally do not involve an effort to parse individual sentences. In adopting this statistical approach, I am thus ignoring much of the information contained in the transcripts that constitute my data. Some of this information – such as the order in which the words appear – seems, a priori, to be important, perhaps even essential, to understanding student utterances. In short, there was every reason to think that the types of analyses described here would not be very successful. Thus, I feel that the most important conclusion of this work is that this is not the case; some relatively simple computational techniques can capture important features of clinical interview dialogue. The implication is that, as a field, we should be open to trying techniques from Statistical NLP, including techniques that just “count words.” That said, most of my presentation in this manuscript has had an exploratory character. My goal has been only to begin to map the boundaries of what might be possible with a family of relatively simple computational techniques.

8.1 Summary of the paper

The analyses I presented were divided into four sections – Sections 4 through 7 of this manuscript. In the first two sections, my analysis was dependent on the use of idealized answer documents as a basis of comparison. In the latter two sections, I derived the coding scheme directly from the data, thus decreasing my reliance on the idealized answer documents. More specifically, the four analysis sections contained the following content:

Comparison of entire transcripts to idealized response documents (Section 4). In this section, I described the LSA-based algorithms employed in some detail, and I presented the results obtained when comparing entire transcripts to idealized response documents. This was done for transcripts that only included student utterances (Kappa of .62) and transcripts that also included interviewer utterances (Kappa of .81).

Analysis of segmented transcripts by comparison to idealized response documents (Section 5). Here I described my efforts to use idealized response documents to code segments of transcripts, and to capture how student DMCs change over the course of an interview. I was not successful in achieving this aim. However, the results suggested that it is, in some cases, possible to use the LSA-based techniques to capture meaning in segments of transcripts that are as small as 10 words.

Induction of a coding scheme by clustering transcript documents (Section 6). In this section, I described how I discovered a coding scheme by clustering the vectors associated with transcript documents. I described in some detail how the clustering algorithm works. Strikingly, the categories induced by the clustering analyses aligned clearly with categories induced by human analysts. For the coding of specific individual transcripts, the results obtained agreed with human coders at a level that approached that of the analysis described in Section 4.

Induction of a coding scheme by clustering segments of transcripts (Section 7). Finally, I discussed the results obtained when all of the transcripts were split into overlapping 25-word segments and these segments were clustered. The clusters obtained could, in most cases, be interpreted in a meaningful way. Furthermore, each set of clusters included some
clusters that aligned with the three idealized DMCs. In addition, I used centroids of the clusters obtained to code entire transcript documents. Once again, the results obtained agreed with human coders at a level that approached that of the analysis that made use of idealized response documents.

8.2 What do these computational techniques buy us?

What role might these computational techniques play in the toolkit of researchers, especially researchers who use clinical interviews to study the conceptions of science students? I presented many results that were intriguing, but all of my conclusions were conclusions about the methods themselves; I didn’t use the methods in the service of any scientific agenda. So what, in the long term, might these techniques buy us?

One question we can ask is whether the computational techniques can and should replace human coders, at least during the final stages of coding. Although it might ultimately be possible to automate the coding of data, I believe it is clear that that the techniques described here are not yet sufficiently accurate in replicating human analysis. As I discussed earlier, the type of coding described here can be done very easily, and with high reliability, by human coders. In contrast, the automated analyses I described generally agreed with humans around 11 or 12 times out of 16 students. Although we can be encouraged by this level of agreement, I do not think this would be an acceptable substitute for human coders.

Moreover, even if we were to decide that that this level of accuracy is sufficient, it must be noted that I never came close to fully replacing human analysts. Most of my analyses required that I had spent a substantial amount of time applying traditional analytic techniques to this data corpus. This was necessary in order to derive the coding scheme embodied in the idealized response documents. These documents were an essential element of the analyses described in Section 4 and Section 5. But, just as importantly, my familiarity with the corpus greatly added to my confidence in interpreting the results produced by the clustering analyses. Furthermore, although I took steps to begin to reduce dependency on the idealized response documents, my interpretation of the clusters still relied, in a substantial way, on these documents.

Thus, even if we can improve the accuracy of these computational techniques, the reduction in effort by human analysts might not be substantial. Nonetheless, the computational techniques described here might play a useful role in our toolkits. In fact, I believe that the biggest and most immediate contribution will be in the support computational techniques can provide for traditional kinds of analysis. Typically, we have multiple researchers apply a coding scheme to a data corpus as a way to assess the reliability of the scheme, and thus to improve our confidence in the results obtained. But reliability is not the same as validity; reliability alone does not ensure that a coding scheme is capturing anything meaningful in a data corpus. I believe that the primary contribution of the computational techniques will be in their ability to provide a type of triangulation that helps us to establish the validity of our analyses. This point is worth some elaboration.

When two humans code the same data in order to establish the reliability of a coding procedure, they are, in a fundamental way, doing the same thing. Thus, what we have are two sets of measurements, both essentially performed with the same type of experimental apparatus. In

\[14\] Of course, humans could be wrong and the computational analyses right.
contrast, if we can find a way to obtain confirmatory results, using a very different type of apparatus, then that should more profoundly increase our confidence in the validity of our results. It is that type of support that I believe is the biggest potential contribution of the computational techniques. Even if we do not have enough confidence in the computational techniques to replace human coders, the fact that we can approximate human-obtained results using very different techniques, even in a crude way, should increase our confidence in those results.

This is true for analyses in which we employ computational techniques to apply a human-derived coding scheme to a data corpus. However, it is even more true for computational analyses that induce a coding scheme. In fact, paradoxically, I believe that the techniques for inducing a coding scheme – the ones that rely on clustering – will likely provide the most powerful support for human analyses, both in the short term and long term. This is paradoxical because it seems, intuitively, that deriving a coding scheme from scratch should be more difficult than applying one. Nonetheless, the results of the clustering analyses already seem to consistently replicate the human-derived scheme. This provides support for the claim that the coding scheme captures something real in our data.\(^\text{15}\)

### 8.3 Prospects for extension to KiP-style analyses

From the perspective of KiP, most of the analyses presented here were very coarse. If we really want to adopt a KiP perspective – as our project team has done with other work on this corpus – then there are a couple of respects in which our analyses should be made more fine-grained. First, except in Section 5, whenever I coded transcripts, I did so at the level of entire transcripts. However, I believe that, in general, DMCs are constructed over a few minutes in an interview, and sometimes shift from one moment to the next. Thus, we would like our analysis to capture changes on that timescale. Second, it is the essence of the KiP perspective that DMCs are constructed, during interviews, out of more basic knowledge resources. From this perspective, what we would really like to do is to identify these more basic resources.

Based on what I have done to date, I can speculate about the prospects of extending the computational techniques to analyses that are more thoroughly consistent with a KiP perspective. First, even if I can never do more than confirm analyses at the level of DMCs, and only in a coarse manner, that might still be helpful. KiP-style analysis poses many difficulties for researchers. It requires that we look across many transcripts, immersing ourselves in the data. And we often can support our arguments in publications only by presenting lengthy examples from transcripts. Thus, even if the computational techniques can only be used to provide independent support for some small parts of a KiP-style analysis, that might still provide a boost to the whole program.

Second, I did take some small steps toward more KiP-style analyses. I showed that it seems to be possible to code very small segments of texts, and that, when coding these segments, there is sensitivity to at least some of the shifts and changes that are visible to researchers. I think we can

---

\(^\text{15}\) This argument rests on the claim that, in using the computational techniques, I am using a “different apparatus.” However, in some respects the apparatus I employed is still the same. In particular, all of the apparatus that went into producing the transcripts – the utterances that are the data – is the same in both cases. Thus, when using computational techniques to replicate my human analyses, I cannot, for example, address the possibility that some results are artifacts of the design of the interviews.
be confident that some version of the segmenting analysis presented in Section 5 will ultimately be able to capture shifts in DMCs.

My attempt to cluster segments into smaller “ideas” was also an attempt to push the analysis in the direction called for by the KiP perspective. That analysis did seem to be able to capture meanings beyond the set of three DMCs. However, I do not believe that this type of analysis will ever allow us to capture, computationally, the more basic knowledge resources out of which DMCs are constructed. The underlying problem is that, in general, there will not be a simple relationship between utterances – or even fragments of utterances – and the underlying knowledge that generates them. Instead, changing configurations of knowledge will work in concert in the generation of utterances. To see this underlying knowledge, I think our computational analyses will need to have another layer added, a layer that can model underlying cognitive states.

8.4 Outstanding issues

A large number of problems remain unsolved, some of which I have been careful to highlight, others which I have glossed over. Here I will mention a few.

8.4.1 The large parameter space

Throughout this manuscript, I have presented analyses that were based on one set of choices in what is essentially an enormous parameter space. There were many different approaches I could have adopted, and many choices about parameters that I was forced to make along the way. However, because the space of possible choices is so large, I was only able to explore a tiny portion of the range of possible parameters. Furthermore, most of the small explorations I did conduct produced inconclusive results as to which choices are optimal. For example, the important choices I faced included:

- Composition of the training corpus. In assembling a training corpus, I had an essentially limitless set of degrees of freedom. I made choices about what types of documents to include, and how many to include. For example, I included documents that were both informal and technical. As an experiment, I constructed a corpus that excluded the technical documents. The results were not substantially different, though the comparison to human coders was a little worse. I will say a little more about the composition of the training corpus below.

- Composition of the stop list. The stop list I used – which specifies words to be exclude from the Word Space – contains 765 words. I conducted some experiments with a significantly reduced stop list. The results with respect to the comparison to human coders were somewhat worse.

- Weighting factors used in building the co-occurrence matrix. Recall that when building the co-occurrence matrix, entries were multiplied by a weighting factor, and then the square root of entries was taken. This post-processing of matrix entries can be done differently. However, I have not tried any alternative approaches.

- Number of columns in the reduced co-occurrence matrix. All of the analyses described here were based on a Word Space with 200 columns, meaning that all of the vectors I worked with contained 200 numbers. But a range of values are possible. I experimented with values ranging from 50 to 500. The results varied in quality but with no obvious pattern. The use
of values between 100 and 200 seems to be typical in the literature. Other parameters also control how the processing of the training corpus is carried out, including the window size (I used 15 words) and the initial number of columns used prior to the reduction (I used 1000).

The idealized response documents. Each of the idealized response documents embodies many decisions. To date, I have not conducted any systematic investigations of variations in these documents.

Inclusion of interviewer utterances. As discussed in earlier sections, I sometimes replicated analyses on transcripts that included the utterances of the interviewer. In the cases presented in this manuscript, the inclusion of interviewer utterances led to better alignment with human coders. However, I nonetheless hesitated to say that analyses that include interviewer utterances are superior. The reason for this hesitation is that, across the larger space of parameters employed (including analyses not presented in this paper), it was not consistently true that better results were obtained when including interviewer utterances.

Cleaning of the transcripts. There are decisions to be made about how to clean up transcripts before analysis. For example, I needed to decide what to do about word fragments, whether to leave them, delete them, or complete them. I needed to decide what to do about “ums” and “uhs.” More dramatically, I could have opted to stem words, that is, to reduce them to their base or root. (If so, the same should perhaps be done for the training corpus.) In the end, I left the transcripts largely unchanged. I just ran some scripts that removed punctuation, formatting, and any annotations that had been added. In effect, this left many decisions to the work of the transcribers, who had transcribed the documents for other research purposes. Note, however, that words that do not appear in the Word Space do not figure into the computation of a vector for a transcript or snippet in which those words appear. This means that these words only matter for my analyses to the extent that they affect how transcripts were segmented.

Clustering algorithms employed. There are many different clustering algorithms that one could employ. Recall that I used a version of hierarchical agglomerative clustering in which, at each step, clusters were combined based on the similarity of their centroids. There are many variants of HAC that use different algorithms for deciding which clusters to combine. I have tried several, including group average agglomerative clustering. There are also clustering algorithms not based on HAC. I tried one prominent alternative, k-means. These algorithms all produced different clusterings. The algorithm I reported on here, HAC with centroid clustering, produced the best clusterings. (See the next section for a discussion of what constitutes the “best” clustering.).

The use of overlapping segments. In all of my analyses that segmented transcripts, I employed a moving window that produced overlapping segments. I used overlapping segments because of my intuition that this would help to minimize the impact of breaking transcripts at arbitrary locations. However, I have not explored the effects of this approach.

In sum, the space of choices available to explore was extremely large, and it was only feasible to explore a small part of this space. My impression is that this is the typical state of affairs in

---

16 See, for example, Manning et al. (2008) for a discussion of some of these alternative clustering techniques and their applications in computational linguistics.
research that makes use of LSA and related techniques. Where the goals of the application are practical (e.g., search and retrieval of documents), then it may be sufficient just to know that the algorithms work in a way that is satisfactory to users. But, when the goals are scientific – when we are trying to find out something about the world – then these decisions are perhaps of more import.

The situation would have been simplified if I was able to discern some clear rules about what choices of parameters worked best. For example, if there was clearly an optimal number of columns for the Word Space, then that would have helped to somewhat focus my choices. But in most cases I could not discover rules of this sort. Sometimes more columns was better, sometimes worse. Sometimes it worked well to include the interviewer, sometimes not.

In the end I chose to settle on choices that seemed, from my reading of the literature, to be relatively standard. For example, values of 100 to 200 seem to be standard for the number of columns in the Word Space, and I used the list of stop words that is provided when downloading the InfoMap software. In this way, at least, I was careful not to just “cherry pick” the best results for each analysis. In accord with this approach, I set values consistently across all analyses. For example, I used 200 columns across every analysis presented. In contrast, if I selected the set of parameters that produced the best results for each analysis, then I could have reported nearly perfect agreement with human coders.

8.4.2 The number and interpretation of clusters

Another important set of problems has to do with the number and interpretation of the clusters produced by the clustering algorithms. I begin with the problem of determining the appropriate number of clusters. Recall that each of my clustering analyses produced a set of candidate clusterings, ranging from one large cluster, to having each item in its own cluster. The difficulty is in deciding which clustering is the appropriate one to select. Unfortunately, there does not seem to be a simple solution to this problem. The best I can do here is to give a sense for the sort of exploration that might ultimately produce a workable solution.

![Figure 37. Residual sum of squares (RSS) as a function of the number of clusters.](image)

We can compute a quantity called the residual sum of squares (RSS). The RSS is the sum of the square of the distance between each element in a cluster and the centroid of that cluster. Figure 37 shows a plot of the total RSS as a function of the number of clusters for the transcript clustering analysis in Section 6 (refer to Figure 19). At the left extreme of this plot, all of the transcripts are in one big cluster; at the right extreme, each transcript is in its own cluster, so that
there are 21 total clusters. Note that the RSS decreases as the number of clusters increases. What this is saying is that, when we have more clusters, each item can be placed in a cluster that has a centroid with which it is a closer match. We are thus faced with a tradeoff: As we increase the number of clusters, the clusters have a better fit to the data. But we get this better fit at the expense of greater complexity.

Ultimately, we would like to have a metric that tells us how to select the best location in this space of tradeoffs. Note that there does seem to be diminishing returns as the number of clusters is increased. Initially, there is a very big drop in the total RSS when the number of clusters increases from 1 to 2, and the amount of improvement is less with a larger number of clusters. I believe that ultimately the best that I will be able to do will be to introduce a tunable parameter that represents a judgment about how to make the tradeoff embodied in this curve (as discussed in Manning, et al., 2008). It is not clear, however, that introducing a tunable parameter is better than simply making a judgment, in each case, about the right number of clusters to employ.

The interpretation of the clusters also poses problems that still remain to be solved. After obtaining clusters, we would like to be able to say something about what those clusters mean. My presentation in this paper relied heavily on the idealized response documents, which somewhat defeats the purpose of the clustering analysis.

8.4.3 The composition of the training corpus

As noted above, the training corpus embodies many choices. There are choices at a larger level about what types of documents, broadly speaking, should be targeted. And there are choices at a finer level, when making decisions about which specific documents to include.

It is worth emphasizing one feature of the broader approach I adopted in targeting documents. Note that I made use of a domain-specific training corpus; as mentioned above, I selected documents from the internet that pertained to the earth’s seasons and climate. Thus, the vectors that result for words such as “rotate” and “orbit” are likely to reflect their narrow uses in Astronomy and Earth Science. I could, instead, have selected a wider range of documents from the internet, in which words such as “rotate” and “orbit” are used in a more varied manner. In that case, the vectors for words such as “rotate” and “orbit” would have reflected these more varied uses.

It is not immediately obvious whether a domain-specific corpus is better for this work than a domain-general corpus. On the one hand, it seems sensible to tune the Word Space for the domain that is my focus. On the other hand, the students interviewed may have tended to use words in a manner that was more informal, and that reflects everyday rather than technical usage.

It is possible to obtain some guidance on this matter from existing research. As discussed in Section 2.3, Shapiro and McNamara (2000) had students read and summarize portions of psychology textbooks and used LSA to analyze the responses. In their analysis, these authors tested both domain-specific and domain-general corpora, and compared the results. They report that they achieved better results using a domain-specific corpus. Wolfe and Goldman (2003) also explored this issue, and they provide a general discussion of the conditions under which it may be preferable to employ a domain-specific or domain-general corpus. In the particular application that they studied, a corpus with an intermediate level of domain specificity seemed to provide the best results.
Thus, although there are no completely definitive guidelines, I believe that prior research does tend to suggest that a corpus that is at least moderately domain-specific is preferable for the type of work I conducted. Needless to say, the success of Dam and Kaufmann (2008), using a similar corpus, also gave me good reason to believe that this approach would work.

8.5 **Future work**

There are many obvious next steps, some that will be easy, some that will be hard. One obvious next step is to try some of these same analyses on different interview data, about topics other than the seasons. It is possible that there is something special about the seasons as subject matter. For example, it might be that, in this territory, a small number of key words (e.g., “tilt”) can do a lot of the work of discriminating among explanations. At the least, one thing that was special about this data is that it was possible to sort students into a small number of relatively distinct categories. This made it possible, for example, to describe the agreement with human results in a simple way. This would not be the case, for example, if there was a large number of highly variable DMCs, or if all of the students essentially gave the same explanation.

There are many ways in which the computational techniques here could be refined or extended. Some techniques are in relatively easy reach. For example, I believe that I am not too far from being able to code whether a student shifts among DMCs. To do so, I can make use of the techniques described here, which compared segments of transcripts to some externally derived reference set. Another possible technique is to compare each individual segment of a transcript to the segments that precede and follow it, looking for discontinuities. This is essentially the technique that has been used by other researchers to examine the internal coherence of texts (Foltz, et al., 1998).

Some extensions of the computational techniques described in this article will require substantially more effort. For example, the LSA-based techniques could be augmented with other approaches that pay some attention to the order of words. We could also develop a way to turn gestures or drawings into tokens that can be inserted into transcripts. At a still higher end are techniques that attempt to infer underlying states that generate the vectors captured by LSA.

Finally, perhaps the greatest puzzle raised by this research is the question of why these techniques work at all. In my view, this question is almost, on its own, worthy of a program of research. Are gestures and diagrams really so unimportant to understanding the explanations given in interviews of this sort? Are a few key words enough to understand what students are saying? Why did the clustering analysis pick out precisely the same set of categories as our human coders? Answering these questions may do more than tell us something about this new class of methods, it might lead to a deeper understanding of the very phenomena about thinking and learning that were are seeking to study.
9 References


## Appendix A

The tables below contain randomly selected 25-word segments, drawn from the six-cluster set, as described in Section 7.1.

### 6-1 (closer-farther)

<table>
<thead>
<tr>
<th>Segment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>actually it goes from no it goes from East to West Right when it moves around the sun I think it goes from east</td>
<td>cold cause it the farthest out Out of the planets cause we're like the third planet out from the sun that we're able to live</td>
</tr>
<tr>
<td>spins around and it the opposite It probably the other way around. But this would be day.</td>
<td>really causes night, somehow, a little more than how it causes the seasons. Well, when the moon comes around it blocks out the sun light, that</td>
</tr>
<tr>
<td>an hour conversation on this Yup Uh oh I don't know how much I learned about it When the earth orbits around the sun and</td>
<td>the earth moves around the sun and the earth is like at one point in the winter it like farther away from the sun and</td>
</tr>
<tr>
<td>the sun here Its also tilted more uh yeah I think that that would here it would be colder I think right now if we</td>
<td>sun It hits an angle say it the earth axis and sometimes the earth this is the sun and this is the earth and it</td>
</tr>
<tr>
<td>away from the sun and towards the summer it closer it near towards the sun Okay The sun in the middle and the earth kind</td>
<td>though yeah yeah yeah I think so because of the till away like over here it would be colder right where right here I think that its like</td>
</tr>
<tr>
<td>south pole cause it takes a long time and the equator is so hot The slant the slant looks like It kind of slants this</td>
<td>picture Yeah that is Um ok There is the sun Yeah I remember that now cause um it like as the earth is rotating as</td>
</tr>
<tr>
<td>here and probably like China somewhere having summer or something Yeah It wintertime since Chicago is right here Cause wintertime it takes a long time</td>
<td>Yeah I remember that now cause um it like as the earth is rotating as it orbiting it rotating too I guess I don't understand</td>
</tr>
<tr>
<td>Well I don't really know how it goes totally but; I don't know It just depends on the time of the year you know I'm</td>
<td>equator is so hot The slant the slant looks like It kind of slants this way if it in winter time and it slants say</td>
</tr>
<tr>
<td>it Yeah How Umm Okay then we were at the top of this the whole time I'm not sure I'm not sure how but wherever</td>
<td>every time at a certain time it will be right here covering part of the earth. It um, smaller. Not for the whole earth. Cause even</td>
</tr>
<tr>
<td>picture Yeah that is Um ok There is the sun Yeah I remember that now cause um it like as the earth is rotating as</td>
<td>it takes a long time and if it Chicago it should be right by the equator right there cause the sun shows No Chicago is</td>
</tr>
</tbody>
</table>
### 6-3 (tilt-based)

<table>
<thead>
<tr>
<th>Quickly and so in the summer it not as quickly but in the summer the sunlight is um the sunlight is coming on to us</th>
<th>Earth is tilted then like our hemisphere is tilted towards the sun but in the winter the southern one is tilted toward the sun and</th>
</tr>
</thead>
<tbody>
<tr>
<td>just colder there but I don't think its actually like ss I think the whole earth is at summer at one time like it goes</td>
<td>Maybe the sunlight just spreads all over and lighting everything and it just includes that area. I don't think so, but I'm not sure. Maybe</td>
</tr>
<tr>
<td>and then we would be like right here and you couldn't see it cause its like over tilted that way yeah this is like the</td>
<td>And it colder in the north pole and colder in the south pole cause it takes a long time and the equator is so hot the</td>
</tr>
<tr>
<td>way yeah this is like the north pole its like its still tilted you still get sunlight cause like during the day but its not as</td>
<td>Wouldn't hit it And it would be hotter uh colder it revolving around the sun Yeah Yeah it cause each circle is a day and</td>
</tr>
<tr>
<td>it colder in the north pole and colder in the south pole cause it takes a long time and the equator is so hot the</td>
<td>Is on, the area it shining on the most it will be the hottest area Okay. This would be the hot area. This area would</td>
</tr>
<tr>
<td>wouldn't hit it And it would be hotter uh colder it revolving around the sun Yeah Yeah It Cause each circle is a day and</td>
<td>The</td>
</tr>
<tr>
<td>the same axis here so if we were over here it would be colder and if we were over here it would be hotter because</td>
<td></td>
</tr>
</tbody>
</table>

### 6-4 (no clear interpretation)

<table>
<thead>
<tr>
<th>We're kind of tilted towards the sun And in the winter it not supposed to be bigger I just drew it that way and it</th>
<th>Earth goes around the sun it kind of tilted it turns too like every 24 hours it turns and so that changes our seasons So when</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sun is right here And all these little planets are over there I don't know And this is well it kind of tilted the earth</td>
<td>Spinning in a different way I'm not sure how I should say that it spinning kind of like that around and then as it goes</td>
</tr>
<tr>
<td>to its normal orbit Mm hmm Mm hmm So like here the sun The planets are kind of going out They kind of just turn</td>
<td>They kind of just turn around That why Pluto so cold cause it the farthest out Out of the planets cause we're like the third</td>
</tr>
<tr>
<td>in books I think I did I've seen pictures and stuff Yeah it does We kind of talked about this in class should say that it spinning kind of like that around and then as it goes it kind of starts spinning like that around that around</td>
<td>The other planets or something Oh The earth is like right here And there a couple of planets in here and it kind of goes</td>
</tr>
<tr>
<td>here the earth and it kind of at an axis And I guess this is where the sun is over here I don't know how</td>
<td></td>
</tr>
</tbody>
</table>

### 6-5 (winter/summer, hemispheres)

<table>
<thead>
<tr>
<th>Well here earth and here where we live And the sun is up over here And during summer it faces and hits it directly so</th>
<th>Summer is really close but how could you winter on the other side how could it be winter on the other side if it really</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter Uh the su the earth when it spins around it doesn't get closer to the equator And in the summertime it does Cause it hot</td>
<td>It does Cause it hot by the equator The earth This this is like the United States and it like doesn't get close to the</td>
</tr>
</tbody>
</table>
is the earth And there the equator and this is the northern hemisphere and this is the southern hemisphere We're in the northern hemisphere So it faces the sun it summer and then when it keeps turning it goes from fall to winter It kind of the earth is tilted doesn't get close to the sun It is it summer that why Um I don't know how to draw it. The sun is warm and will be over here So during our summer they'll be having winter if that theory is true I don't know if it true or not

Because it colder in the winter because the earth spins around and it like close to the equator in the winter Uh the su the world we have different seasons Say like we have winter here and probably like China somewhere having summer or something. Yeah It wintertime since Chicago

<table>
<thead>
<tr>
<th>6-6 (side-based)</th>
</tr>
</thead>
</table>
| is always cold, I think Antarctica is here, even there are like days on the daytime it very, they don't have very much daytime, the opposite. It will be daytime, I believe, at the opposite end. I believe this is day and this is night. Mm hmm. Well, the Yeah I think Rotating mm hmm The day would be hotter and the night will be cooler yeah Yeah but at night it will be to be Yeah Like it I don't really know how to draw a…no that like a box We're on this side Yeah cause we're on circle is a day and it revolves all around It will probably be around on this side But turning it around this would probably be where other side of the world is like summer on their side you know On the other side … of the earth We're here and we're the state so I think it get hotter and for winter it turns on other side and the sun not reaching it so it will everything and it just includes that area. I don't think so, but I'm not sure. Maybe it the opposite. This is night, and this is day basically on it lighting on it When it turns around and its summer it going to be cooler on the other side and it going guess it has to if this one or something I don't know Yeah Maybe Florida Did you mean a different country or something It would