Journal of the Learning Sciences

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlns20

A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data

Bruce Sherin [a]

[a] School of Education and Social Policy , Northwestern University

Accepted author version posted online: 05 Sep 2013.Published online: 29 Oct 2013.

To cite this article: Bruce Sherin (2013) A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data, Journal of the Learning Sciences, 22:4, 600-638, DOI: 10.1080/10508406.2013.836654

To link to this article: http://dx.doi.org/10.1080/10508406.2013.836654

THE JOURNAL OF THE LEARNING SCIENCES, 22: 600–638, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 1050-8406 print / 1532-7809 online
DOI: 10.1080/10508406.2013.836654

Routledge
Taylor & Francis Group

A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data

Bruce Sherin
School of Education and Social Policy
Northwestern University

A large body of research in the learning sciences has focused on students' commonsense science knowledge—the everyday knowledge of the natural world that is gained outside of formal instruction. Although researchers studying commonsense science have employed a variety of methods, 1-on-1 clinical interviews have played a unique 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. In this article, I explore the application of computational techniques to the analysis of this familiar type of data. I describe the success I have had using extremely simple methods from computational linguistics—methods that are based on rudimentary vector space models and simple clustering algorithms. These automated analyses are employed in an exploratory mode as a way to discover student conceptions in the data. The aims of this article are primarily methodological in nature: I attempt to show that it is possible to use techniques from computational linguistics to analyze data from commonsense science interviews in a manner that may provide convergent support for the work of human analysts. As a test bed, I draw on transcripts of a corpus of interviews in which 54 middle school students were asked to explain the seasons.

The rapid rise of the new field of learning analytics has been driven, at least in part, by the increasing power and omnipresence of computers. Researchers now have computers with power and storage that, not long ago, would have ranked them among the world's supercomputers. This development opens up the possibility

Correspondence should be addressed to Bruce Sherin, School of Education and Social Policy, Northwestern University, 2120 Campus Drive, Evanston, IL 60208. E-mail: bsherin@northwestern.edu

of bringing new types of computational techniques to the analysis of data about learning. Just as important, computers have made it possible to capture new kinds of *data* about learning. For example, when students carry out learning activities on a computer, it is relatively easy to capture a trace of their activities in the system. The result can be extremely large data sets—data sets that require the new methods of learning analytics.

Thus, discussions of learning analytics frequently encompass two interrelated changes: new data and new techniques for analyzing those data. In assessing the potential usefulness of learning analytics for the field of the learning sciences, I believe we should consider these two innovations separately. In this article, my focus is on the second of these two innovations: I describe new computational techniques, but I apply these new techniques to familiar data, and in the service of familiar research goals. In doing so, I hope to help the field better understand whether the techniques of learning analytics might allow us to address research concerns that have, to date, been the focus of the learning sciences.

A FOCUS ON COMMONSENSE SCIENCE KNOWLEDGE

In this article, I apply learning analytic techniques to the study of *commonsense science* knowledge, the science-related knowledge that individuals acquire prior to and outside of formal instruction in a discipline. Research in commonsense science has played an important role in the learning sciences, as well as in cognitive studies of education more broadly. It thus provides a sensible platform on which to base this work. More specifically, I draw on a corpus of interviews in which students were asked to explain the earth's seasons, a topic that has been highly studied (e.g., Atwood & Atwood, 1996; Lee, 2010; Lelliott & Rollnick, 2010; Newman & Morrison, 1993; Sadler, 1987; Trumper, 2001).

Although researchers studying commonsense science knowledge have employed a variety of methods, one-on-one clinical interviews (Clement, 2000; diSessa, 2007; Ginsburg, 1997) 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. In simple terms, this coding by human analysts can be seen as consisting of two steps: (a) The analyst seeks to identify a set of "conceptions" in terms of which the reasoning exhibited by all students in the corpus can be understood; (b) this set of conceptions is employed to understand (and perhaps code) specific portions of transcribed interviews.

In the work described in this article, I am looking to computationally automate parts of this analysis. One possibility would be to focus on automating only the second step of the analysis. In that case, human analysts would first identify a set of conceptions and then the computer would code transcripts in terms of those conceptions. However, I attempt to show here that it is possible to automate *both* steps—the identification of common conceptions and the coding of individual transcripts.

There are many reasons to believe that this type of automation should be extremely difficult. It would seem to require that we solve the full problem of natural language processing (NLP). 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. Finally, inducing a coding scheme of conceptions—as opposed to merely applying one—would seem to require an ability to look across the breadth of a data corpus, with the larger aims of the research in mind, in order to recognize relevant patterns in the data. All of this work can pose difficulties even for human analysts.

THE GOALS OF AN AUTOMATED ANALYSIS

One might assume that the great strength of computational methods is that they can reduce the labor required of human analysts. If that were the case, then our goal would be to replicate the work of human analysts and to do so in such a manner that the work required by human analysts is minimized. However, I do not believe that this is the goal to which we should aspire. At present, the analyses I describe do not obviously reduce the amount of work required by human analysts. As I lay out in future sections, performing and interpreting the computational analyses still requires, as a preliminary, much of the work associated with a traditional, human-based, qualitative analysis.

Furthermore, there is a loftier goal to which we can aspire: We can seek to use human-based and computational methods in tandem, in a manner that increases our confidence in both. Our existing human-based analyses are not so uncontested that we should be satisfied with replicating them computationally. Similarly, we should not be content with using computational techniques to apply existing theoretical frameworks under the assumption that those frameworks are uncontroversial.

It is worth elaborating how the use of computational methods might increase our confidence in traditional methods and existing theory. In a typical workflow, a team of researchers works together to derive a coding scheme for a data corpus. Then multiple researchers apply this 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.

In 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 this type of convergent support—this possibility for triangulation—that I believe is the biggest potential contribution of the computational techniques. If we accept this goal as the ultimate aim of computational analyses, then we are faced with a difficult bootstrapping endeavor. We would like to use the new computational methods as a somewhat independent means of confirming prior analyses and testing theories. But we have to first develop some certainty in the computational methods. To do so, the only references we have available are existing analyses and theoretical frameworks.

THE APPROACH, IN BRIEF

My approach in this article is to go as far as possible with simple techniques. This approach has two virtues. First, it makes good sense as a research strategy; we will better understand the merits of different methods if we start simply and only introduce more complex approaches when there is a clear payoff. Second, starting simply allows me to present a set of techniques that can be described, in complete detail, for a wide audience.

The methods employed are drawn from *statistical NLP.* 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. The algorithms that underlie techniques from statistical NLP can be much easier to implement than true NLP.

More specifically, I explore the use of *vector space models* augmented with *cluster analysis.* These choices make sense for a number of reasons. First, vector space models have a conceptual simplicity that make them relatively easy to reason about as we attempt to understand what an analysis is capturing. Second, one type of vector space model, *latent semantic analysis* (LSA), has already been employed, with substantial success, in applications relevant to educational research (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 a vector space–based approach to my research team's data proved promising and thus justified further exploration (Dam & Kaufmann, 2008).

The remainder of this article proceeds as follows. In the first section, I review some of the relevant literature, including research on commonsense science knowledge and applications of vector space models in education research. Then, in the section that follows, I familiarize the reader with the data corpus employed in this work. Following this section is a discussion of how I judge the success of the computational methods I describe.

The next section is the heart of my presentation. There I describe the computational techniques employed in some detail and the results obtained with those techniques. Next, I briefly describe some variations on this analysis in order to give the reader a sense for the larger space of analyses that I explored. In the final section, I conclude the article and reflect on the prospects of the techniques described here, both for research on commonsense science and for the field in general. In all sections, I attempt to describe the algorithms employed so that readers can program these algorithms on their own and so that they can judge whether the techniques presented in this article are likely to be useful for learning scientists beyond the narrow circumstances described in this work.

LITERATURE REVIEW

Research on Commonsense Science

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 (see Duit, 2009).

The large 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), or 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. It is for these reasons that *interviews*—and clinical interviews, in particular—are frequently the method of choice for researchers studying prior science 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. In this perspective, 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, then answers given might be generated, in the moment, out of the elements that compose 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 (Sherin, Krakowski, & Lee, 2012). This sets a high bar that must be met by the computational analyses. To be successful, the computational analyses must be able to produce results that are consistent with, and can support, a KiP-style analysis. This requires seeing through the data—the words spoken by students—to uncover a possibly large number of underlying knowledge elements.

Vector Space Models

The problem faced here is essentially one of NLP: We want to give a computer the ability to understand the natural language spoken by participants in interviews. The techniques employed here are drawn from statistical NLP, a subfield of NLP that uses a variety of statistical and probabilistic methods to solve some of the thornier problems in NLP. Analyses from traditional NLP are frequently concerned with the parsing of individual sentences in order to extract their meaning. In contrast, analyses in statistical NLP are usually concerned with analyzing statistical properties of a large corpus of text. (For a lucid introduction to statistical NLP, see Manning & Schütze, 1999.)

The analyses in this article draw primarily on one class of techniques from statistical NLP, those that are based on *vector space* models. In these models, the meaning of a block of text—a word, paragraph, essay, and so on—is associated with a vector, usually in a high dimensional space. Two blocks of text are then understood to have the same meaning to the extent that their vectors are the same. In this way, a vector space analysis makes it possible to compute the similarity in meaning between any pair of words or blocks of text.

Vector space models were initially developed to solve problems of *information retrieval;* they were intended as a means of retrieving information from large electronic databases. In the prototypical situation, there was a corpus of documents, numbering in the hundreds or even thousands. The goal was for users to be able to type queries into a computer-based system, which would respond by returning the documents in the corpus most closely related to the query.

Outside of information retrieval, some of the earliest and most persistent applications of vector space–based approaches have been in applications related to education. Many of these educational applications have made use of one type of vector space approach, *LSA* (Berry, Dumais, & O'Brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA incorporates two primary innovations that distinguish it from a more basic vector space analysis. First, LSA makes use of an auxiliary *training corpus,* which typically consists of a large collection of texts that make use of words that appear in the corpus to be analyzed. Second, LSA reduces the dimensionality of the vector space, which has the effect of uncovering latent relations among terms and speeding up the computations required.

We can understand the range of uses of vector space–based approaches in education by recognizing that they provide us with one important type of capability: They allow us to compare the meanings of two blocks of text. This one tool has been shown to have many uses in education-related applications.

- *Essay evaluation.* LSA has been employed, in multiple ways, to evaluate essays written by students (Foltz, 1996; Foltz, Britt, & Perfetti, 1996; Foltz, Gilliam, & Kendall, 2000; Foltz, Laham, & Landauer, 1999; Hastings, Hughes, Magliano, Goldman, & Lawless, 2011; Landauer, Laham, & Foltz, 2003). The essays are evaluated by using LSA to compare student essays to reference texts.
- *Assessing reader strategies.* LSA has been used as a tool to automate the identification of strategies used by readers (Kurby, Wiemer-Hastings, Ganduri, Magliano, & Millis, 2003; Magliano & Millis, 2003; Magliano, Wiemer-Hastings, Millis, Munoz, & McNamara, 2002; Millis et al., 2004). In these applications, students may be asked to read a source text one sentence at a time. After reading each sentence, they "self-explain" the sentence, either verbally or by typing into a computer interface. The analysis then focuses on comparing these self-explanations to various benchmarks, such as parts of the source text.
- *Judging of the appropriateness of texts for readers.* LSA has been used as a tool for determining the appropriateness of a text for specific learners. In some of these applications, students are matched to a text by looking at their prior knowledge and comparing it to the information contained in a text (e.g., Wolfe et al., 1998). In other applications, LSA is used in order to measure the *coherence* of texts by comparing sentences within the text (Foltz et al., 1998; Graesser, McNamara, & Kulikowich, 2011; Graesser, McNamara, Louwerse, & Cai, 2004).
- *Use in intelligent tutors.* LSA has been used within computer-based instructional systems. One prominent example is Summary Street (Wade-Stein & Kintsch, 2004), which provides students with feedback as they learn to write summaries of source texts. Another prominent example is AutoTutor (D'Mello, Graesser, & King, 2010; Graesser, Lu, et al., 2004; Graesser, Wiemer-Hastings, & Wiemer-Hastings, 2000). AutoTutor

provides a general tutorial architecture, applicable to many domains, in which students are presented with a series of tasks. In AutoTutor, LSA is used as part of a component of the system that is used to understand the responses typed by students.

Comparison to the present work.

The work to be presented here differs in several important respects from the educational applications described previously. The first three classes of applications all look at the relationships between readers and source texts. In contrast, in the present work, there is nothing that plays the role of the source text. Instead, students respond to short questions posed by the interviewer. This is important because it means that there is not a preexisting source text that can serve as a reference against which we can compare the answers given by students.

The use of vector space methods within AutoTutor is, in some respects, a better match to the present work. As in the present work, AutoTutor asks a student to respond to a short question. AutoTutor has even been applied to science content (Graesser, Lu, et al., 2004). However, there are important differences. In AutoTutor, student responses are compared to metrics that are specified in advance, including misconceptions. In the present work, I describe techniques for automatically inducing a set of conceptions from the data themselves. Finally, I am not using LSA; instead, I believe it makes sense to begin with simpler techniques and then to pursue more sophisticated methods as it seems necessary.

THE INTERVIEWS

The Data Corpus

The data used in this work were drawn from a larger corpus collected by the National Science Foundation–funded Conceptual Dynamics Project.[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 54 interviews in which students were asked about issues pertaining to the earth's climate and seasons. Although the interview protocol was standardized, the corpus was assembled opportunistically, across a few research contexts. The students were all interviewed while they were in seventh or eighth grade or in the summer following eighth grade. Eighteen of the 54 interviews were conducted at the end of a researcher-designed curriculum on climate change. However, the explanation of the seasons was never explicitly discussed during this curriculum.

[1] National Science Foundation Grant No. REC-0092648. Conceptual dynamics in complex science interventions (B. Sherin, principal investigator).

Transcripts of all 54 interviews served as inputs to the analysis that is reported here. However, in this article, detailed analyses are presented for only four of the interviews. A supplementary set of analyses is available in online supplemental material, discussed later.

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 his or her 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.

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 (e.g., Atwood & Atwood, 1996; Lee, 2010; Lelliott & Rollnick, 2010; 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.

The difficulty of explaining the seasons, coupled with the diversity of responses, are some of 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. Ultimately, we 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.

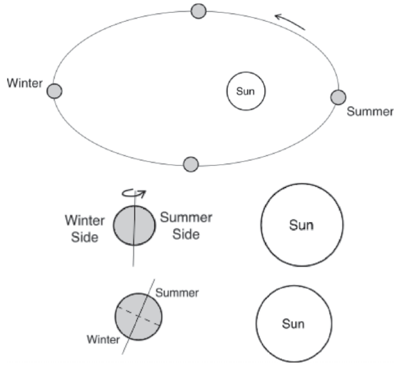
An Overview of the Contents of the Data Corpus

In prior work with our seasons data, conceptual dynamics researchers have adopted a strongly KiP perspective (Sherin et al., 2012). 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 is cued. The student then gives 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* (DMC). For the purpose of the present work, it is reasonable to think of a DMC as a student's current working explanation of the seasons. Thus, throughout this article, I use the terms *DMC* and *explanation* interchangeably.

Furthermore, in prior work, conceptual dynamics researchers have attempted to understand the dynamics of interviews at the level of knowledge elements; we have attempted to identify specific elements and to describe in some detail how DMCs are constructed in specific interviews (Sherin et al., 2012). I cannot fully recapitulate that analysis here. But it is necessary for the reader to have a sense for that analysis in order to understand the automated results discussed later.

The explanations of the seasons given by the students we interviewed varied along a number of dimensions, and each explanation had elements that made it unique. It is helpful, nonetheless, to begin with a number of reference points in the form of a few categories of explanations (DMCs). The first category, *closer–farther,* is illustrated by the diagram in Figure 1a . In closer–farther explanations, the earth is seen as orbiting (or moving in some other manner) 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 is 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 usually focused on the rotational motion of the earth rather than its orbital motion. In side-based explanations, the earth rotates so that first



(a) Closer-farther explanation of the seasons.

(b) Side-based explanation of the seasons.

(c) Tilt-based explanation of the seasons.

FIGURE 1 Three categories of dynamic mental constructs.

one side, then the other, faces the sun. The side facing the sun 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. The hemisphere that is tilted toward the sun experiences summer, and the hemisphere that is tilted away experiences winter. This category of explanation includes (but is not limited to) the normative explanation. In the normative explanation, the earth maintains a fixed tilt relative to its plane of orbit. For this reason, first one hemisphere, then the other, is tilted toward the sun. The hemisphere that is tilted toward the sun is struck more directly by sunlight and is therefore warmer.

As discussed in Sherin et al. (2012), each of these DMCs should be understood to be constructed out of multiple elements of knowledge. For example, when constructing a closer–farther DMC, a student might recall that the earth's orbit is not a perfect circle and that it takes a year to orbit the sun. He or she might further reason that when the earth is closer to the sun, it should be warmer and that, because this follows a 1-year cycle, it constitutes a plausible explanation of the seasons.

Tilt-based explanations are built out of somewhat different elements. Many of the students we interviewed seemed to recall that the earth's axis is tilted, and many recalled that this is somehow important to an explanation of the seasons. Some of the students we interviewed gave tilt-based explanations that were fully consistent with the normative explanation. However, we also saw tilt-based explanations with nonnormative characteristics. For example, some students reasoned that the hemisphere tilted toward the sun is warmer because the tilt makes that hemisphere closer to the sun. To be successful, the computational analysis should be able to capture differences of this sort; it should be able to capture similarities and differences at the level of the elements that compose DMCs.

Across the interviews, we saw many commonalities in the knowledge students seemed to possess. For example, most students had at least rudimentary knowledge about the earth and sun and their relationship:

- Every interviewee seemed to know that the earth and sun are large celestial objects.
- Most knew that the earth moves with respect to the sun in some manner and that the earth orbits the sun and rotates on its axis.
- Many knew that the earth's orbit is somewhat elliptical.
- Many mentioned the fact that the earth's axis is tilted (and seemed to know that the tilt is in some way critical to an explanation of the seasons).
- Many mentioned the fact that the earth has two hemispheres divided by an equator.
- Many seemed to understand that the earth's day/night cycle must somehow be linked to the movements of the earth, sun, and moon.
- All students assumed that the sun is the primary source of heat for the earth.

Students also invoked some basic understanding of the behavior of heat sources. For example:

- Many asserted (or implied) that the effect of a heat source will be felt less strongly farther from the source.
- Some asserted (or implied) that when light impinges more directly on a surface its effects will be stronger.

Finally, all students had a great deal of basic world knowledge pertaining to the phenomenology of the seasons. For example, they knew that it was cold in the winter and warm in the summer, that there is more snow in the winter, that leaves change their colors in the fall, and that days tend to be longer in the summer and shorter in the winter.

In Sherin et al. (2012) we also presented data to support the claim that student explanations often changed as each interview unfolded. We were able to see students gradually assemble explanations out of parts such as those discussed previously. We could also see students shift among dramatically different explanations. This is another feature of our analyses that the computational analysis should capture.

Example Interviews

Now I briefly discuss a few specific interviews that are used as reference points throughout this article. I begin with a discussion of one student, Marcus. Like many students, Marcus's first response to the interview question took the form of a closer–farther explanation, with the earth orbiting the sun in such a way that it is sometimes closer and sometimes farther from the sun. (In the transcript, I = Interviewer and M = Marcus.)

I: Okay. All right, so my first question is, um, why is it warmer in the summer and colder in the winter?

M: Because of like the alignment, like in, like the sun, I think it has to do with. Like the sun, like the plan—like, like the course [moves finger in circular motion], like the sun is here [gestures vaguely], and like the planets, you know, go like around the sun and so like it depends like where, cause like the earth might be, like farther away from the sun, or something. And like in like, in the winter, and in the summer it might be closer, or, yeah, or um, like, yeah.

I: So how is—say that again, how is the earth moving?

M: It, it goes like around the sun.

I: Uh-uh.

M: And, like, and so like once a, one time around is one year.

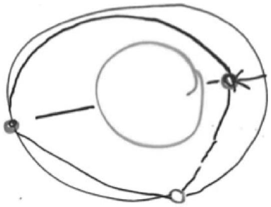


FIGURE 2 Marcus's drawing.

At this point in the interview, Marcus asked if he could draw a picture. He then made the drawing shown in Figure 2, commenting as he drew.

M: Let's say like this is the sun [draws the large central circle], and then, sorry, and then this is the earth [makes a circle for the earth], it goes like around, like it has a path that it goes on [draws outer orbit]. And like, let's say this is, pretend this is like here or something [draws small circle closest to sun], like summer, I think, I'm like I'm not really sure, but maybe like summer might be like closer cause this is closer, and then this might be winter, or, in the, I, I don't know.

I: And, so where's the earth in the summer?

M: We'd be like, like closer, like this is summer cause it's closer [indicates closer earth], or like, I don't know, like this is winter cause it's farther [indicates small earth at the left]. And like, closer is warmer, I don't know.

We thus see that Marcus has built an explanation from some of the resources discussed previously; he knows, for example, that the earth orbits the sun in a path that is noncircular, that the sun is hot, and that the effect of a heat source is stronger closer to the source.

When a student gave an explanation that implied that the entire earth experiences the same season at a given time, our interviewers were instructed to follow up with a specific challenge; namely, we asked if the student had heard that when it is summer locally it can be winter in other parts of the world. (I refer to this as the *different seasons challenge.*) When challenged in this manner, Marcus seemed to immediately recognize that it posed a problem for his explanation.

I: Well that makes a lot of sense, uh, but have you heard, um, that when it's winter here it's summer in other places?

M: Oh yeah. Um.

I: Is that a problem for your explanation?

M: Yes [laughs].

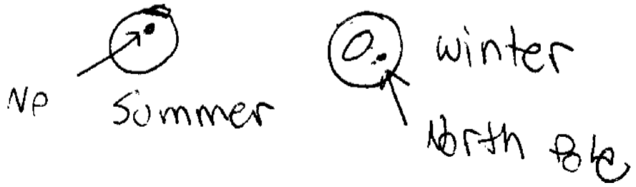


FIGURE 3 Zelda's drawing.

I: So could you just explain out loud why it's a problem?

M: It's a problem because if earth is here in winter [points to drawing] and it has, like if earth, if I said earth was here in winter, and in summer earth is here, well in winter like China would be like summer, because China's on the opposite side. But if earth is here and this is winter, then it all has to be winter, but it's not all winter at the same time.

The interviewer then offered Marcus the chance to refine his explanation. In response, he attempted to construct a side-based explanation based on the fact that the earth rotates, so that different sides are facing the sun at a given time. He began by saying:

M: Winter, um, it, maybe because [sighs] how, how like maybe because when the earth moves, like how it like rotates, it like, one side is more facing, well no that doesn't make sense at all. Um, I don't know.

Then, over the next few minutes, he struggled to construct a side-based explanation that was consistent with various known facts. He never succeeded in producing an explanation with which he was satisfied.

I want to briefly introduce interviews with two other students from the corpus. My purpose here is to show how two DMCs that are grossly similar—they are both varieties of tilt-based explanations—can differ in the elements out of which they are composed. The first student, Zelda, gave a tilt-based explanation very close to the accepted scientific explanation (refer to Figure 3). According to Zelda, because of the earth's tilt, one "side of the earth is tilting toward the sun, or it's facing the sun or something so the sun shines more directly on that area, so it's warmer."

I: Why do you think, what is, could you tell me your best guess, why it's warmer in the summer and colder in the winter?

Z: Because, I think because the earth is on a tilt, and then, like that side of the earth is tilting toward the sun, or it's facing the sun or something so the sun shines more directly on that area, so it's warmer.

I: Can you draw a picture? It doesn't have to be artistic or anything.

Z: So that was the sun, and like the earth, if this is the top it's like tilted so the sun shines on like the bottom part, it's tilted back.

In saying that the earth's tilt causes one part of the earth to receive more direct sunlight, Zelda is giving an explanation that is very close to the accepted explanation. However, Zelda's talk about directness is also combined with talk about the side of the earth facing the sun. These are features that are not part of a strictly normative explanation and are perhaps more indicative of a side-based explanation.

Caden also gave an explanation in which tilt featured prominently. But in Caden's explanation, the tilt of the earth affects temperature because the hemisphere tilted toward the sun is *closer* to the sun, and the hemisphere tilted away is *farther* from the sun.

I: So the first question is why is it warmer in the summer and colder in the winter?

C: Because at certain points of the earth's rotation, orbit around the sun, the axis is pointing at an angle, so that sometimes, most times, sometimes on the northern half of the hemisphere is closer to the sun than the southern hemisphere, which, change- changes the temperatures. And then, as, as it's pointing here, the northern hemisphere it goes away, is further away from the sun and get's colder.

I: Okay, so how does it, sometimes the northern hemisphere is, is toward the sun and sometimes it's away?

C: Yes because the at—I'm sorry, the earth is tilted on its axis.

I: Uh uh.

C: And it's always pointed towards one position.

Thus, Caden and Zelda both gave tilt-based explanations, but they differed in how exactly the tilt of the earth affects the seasons. For Caden the tilting causes one hemisphere or the other to be closer to the sun. For Zelda, the tilting primarily causes parts of the earth to receive the sun's rays more or less directly.

HOW WE WILL JUDGE SUCCESS

Before beginning a discussion of the computational analysis, it is essential to clarify what I am hoping to achieve and how I propose to judge whether the effort is successful. As discussed earlier, my goal is not to replicate human analysis as a means of saving labor. Instead, I am hoping to provide a new method, one that we can use in tandem with human analysis in order to increase our confidence in both. Because this is the ultimate aim, it means that we should view this work as in the early stages of a bootstrapping endeavor. I would like to use the new computational methods to test existing theory and analyses. But first the computational methods must be tailored for the present context, and I must establish that they provide meaningful results. Doing so requires that we use existing theory and human analyses as they currently exist.

For these reasons, my approach is to look only for indications that the bootstrapping endeavor has a good chance of ultimately being successful. To do this, I employ two broad types of criteria. First, I ask the following: Are the computational analyses capable of resolving features of the data that are necessary to test the theoretical frameworks under examination? In order to provide convergent evidence for the theoretical frameworks that are being tested, the computational analyses must be able to make contact with the theories and to provide analyses in some of the same terms as the human analysis. Of course, in the long run, we must be open to the possibility that the theories in question are incorrect. But in this stage of the work, we must assume some confidence in the prior state of our theorizing. More specifically, I ask the following:

1. Can the computational analyses produce results that are interpretable in terms of entities that appear in our theory? In particular, can they identify entities that can be aligned with the types of knowledge elements identified by our KiP analysis?
2. Are the identified knowledge elements at an appropriate grain size and level of abstraction?
3. Can the computational analysis identify configurations of elements of the sort identified by the theory? In the human analysis, not only do we see individual knowledge elements, we see them as combined into DMCs.
4. Can the computational analysis capture interview dynamics? According to our theory we expect that, as an interview unfolds, a student's DMC may shift and evolve.

Second, in looking to judge the success of the computational analysis, we will want to do more than check to see whether the analysis resolves features that are key to the theory. We also want to see whether the computational analyses produced align with human analyses. In the long run, we will not want to assume that the human analysis provides a gold standard against which the computational analyses can be checked. But at present we have more confidence in human analysis. Thus, we will look for alignment with prior analyses along the dimensions listed previously; we will look to see whether there are similarities in the knowledge elements identified as well as in the analyses of detailed dynamics of individual interviews.

VECTOR SPACE ANALYSIS OF THE SEASONS CORPUS

In this section I describe how I used a combination of vector space models and clustering to automate analyses of the data described previously. This work takes off from the foundational work done by Gregory Dam and Stefan Kaufmann (2008), which employed an LSA-like approach to apply a given coding scheme to an earlier subset of our data corpus. The work described in this article extends the work of Dam and Kaufmann in several respects. I already mentioned one important difference: I am exploring the use of simpler vector space models. There are two other important differences. First, Dam and Kaufmann's analysis did not *discover* student conceptions in the data corpus. Instead, it began with the conceptions identified by human analysts and used those conceptions to code transcript data.

Second, Dam and Kaufmann were primarily concerned with coding at the level of the student; each student was coded, by the computer, in terms of one of the three categories of explanations listed in Figure 1. The success of this analysis was judged by comparison to an analysis of these same transcripts by human coders, restricted to the same set of three explanations. However, this type of analysis represented a drastic simplification over our earlier qualitative analyses of the corpus, as it ignored interview dynamics and did not produce an analysis at the level of elements. Nonetheless, Dam and Kaufmann obtained a moderately good alignment between computer and simplified human codes.

At the time Dam and Kaufmann performed their analyses, we only had a portion of the current corpus. As our data corpus has grown, the simplification employed by Dam and Kaufmann has become increasingly untenable. A significant fraction of our corpus now contains interviews in which students clearly shift among explanations. Thus, analysis at the level of the transcript no longer yields good agreement between automated and any type of simplified human coding procedure. Therefore, in this new work, all of the analysis is done at a finer time scale; I look to identify student ideas only in small segments of text.

Converting Text to Vectors

In a vector space model, every passage of text is mapped to a single vector, each of which consists of a list of numbers. The direction in which this vector points is taken to be a representation of the meaning of the passage. If the vectors for two passages point in roughly the same direction, then the vectors are understood to have similar meanings. More precisely, the similarity between two passage vectors is quantified as the cosine of the angle between the two vectors (or, equivalently, the dot product of the vectors if the vectors are of unit length).

In the most rudimentary forms of vector space models, the mapping between text and vector is accomplished in a straightforward manner. First the analysis

TABLE 1
Vocabulary and Vector for Marcus's First 100 Words

Vocabulary	*Raw Count*	*Weighted Count*	*Normalized Vector*
sun	6	2.8	0.81
earth	1	1	0.29
side	0	0	0.00
away	1	1	0.29
tilted	0	0	0.00
closer	1	1	0.29
axis	0	0	0.00
day	0	0	0.00
farther	1	1	0.29
time	0	0	0.00

looks across the entire corpus of text included in the analysis and compiles a vocabulary—a complete list of all of the words that appear in the corpus. This vocabulary is then pruned using a so-called stop list of words. The stop list typically consists of a set of highly common noncontent words, such as *the, of,* and *because.* For the corpus used in this work, the full vocabulary contained 1,429 words, the stop list consisted of 782 words, and the resulting pruned vocabulary contained 647 words. If this resulting vocabulary is sorted from the most common to least common words, the top 10 words correspond to the list shown in the left hand column of Table 1 .

This vocabulary can be used to compute a vector for a passage from an interview transcript as follows. First the analysis takes the transcript and removes everything except the words spoken by the student. For example, when this is done, the first 100 words from the interview with Marcus are as follows:

> because of like the alignment like in like the sun i think it has to do with like the sun like the plan like like the course like the sun is here and like the planets you know go like around the sun and so like it depends like where cause like the earth might be like farther away from the sun or something and like in like in the winter and in the summer it might be closer or yeah or um like yeah it it goes like around the sun and like and so like once a one

This passage of 100 words can now be converted to a vector. To do this, the analysis goes through the entire vocabulary, counting how many times each word in the vocabulary appears in the passage. When this is done for this small block of text, the result is a list of 647 numbers. For illustration, the values for the 10 most common words are listed in the "Raw Count" column in Table 1. It is worth emphasizing that the resulting vector depends only on the words that appear somewhere in the passage of text; it does not depend, in any manner, on the order in which these words appear. For this reason, models of this sort are called *bag of words models,* as they treat text as just an unordered collection of words.

Two final complications remain. In most vector space analyses, the raw counts are modified by a weighting function. In the analyses reported in this article, each count is replaced with (1 + log[count]). This has the effect of dampening the impact of very frequent words. (Raw counts of zero are just left as zero.) Appropriately weighted values are shown in the third column of Table 1.

Finally, all of the vectors used in the present work are *normalized.* This means that every entry in the vector is divided by a constant (the length of the vector) so that the resulting vector has a length of 1. If the vector shown in the third column of Table 1 was the entirety of the vector for this segment of text, then the normalized vector would appear as shown in the fourth column of the table.

Using Passage Vectors to Discover Meanings in the Data Corpus

The procedure described in the preceding sections provides a means of mapping a passage of text to a vector consisting of 647 numbers. This capability can now be used to discover units of meaning that exist across the 54 interviews that compose the data corpus. This process involves four steps that I now discuss: (a) preparing and segmenting the corpus, (b) mapping segments to vectors, (c) clustering the vectors, and (d) interpreting the results.

Preparing and segmenting transcripts.

First, as described previously, the transcripts are reduced so that they include only the words spoken by the student during the interview. This means that all interviewer utterances are removed, as are any annotations, punctuation, or symbols that are included in the transcript. The result is a transcript consisting of just a long list of words.

Next, the analysis requires a means of attaching meanings to small parts of an interview transcript. This requires a means of segmenting a transcript into smaller parts. In keeping with my attempt to begin with as simple an analysis as possible, I simply broke each transcript into 100-word segments. (This size produces segments that are small enough to resolve the dynamics of interviews while still being large enough to be interpreted meaningfully by the algorithms.) In order to lessen problems that might be caused by the fact that this introduces arbitrary boundaries, I chose to employ overlapping 100-word segments, with the start of each segment beginning 25 words after the start of the preceding segment. So the first segment of a transcript would include Words 1–100, the second Words 26–125, the third Words 51–150, and so on. When all of the 54 interview transcripts were segmented in this manner, I ended up with 794 segments of text.

Mapping segments to vectors.

The next step in the analysis is to map each of these 794 segments to a vector. To accomplish this, I employ precisely the

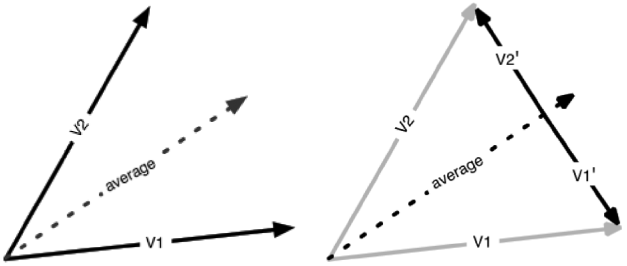


FIGURE 4 How to compute deviation vectors.

method described previously. The result is 794 vectors, each consisting of a list of 647 numbers.

However, there is one inherent problem with applying vector space models to an analysis of this sort of data. Recall that vector space models such as LSA were originally developed as a means of finding documents in a large corpus that pertain to a given topic. They were thus not developed for finding fine distinctions in meaning among documents pertaining to very similar topics; their purpose was to identify documents pertaining to one topic from among a larger set of documents on a range of different topics. That purpose is quite different than the one we face here. All of the documents involved in this analysis are about very similar subject matter; they all explain the seasons, and they almost all do so by talking about the position and motion of the earth in relation to the sun.

In fact, the clustering analysis (described in the next section) does not yield meaningful results if I use the document vectors that are produced using only the method described in the preceding section. I also need a means of modifying the vectors so that they highlight their more unique features—the features that, on average, tend to differentiate a segment from the other 793 segments of text.

For that purpose, I compute what I call *deviation vectors.* To compute the deviation vectors for two vectors *V1* and *V2,* I 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 (refer to Figure 4). The perpendicular components, V1' and V2', are the deviation vectors. If we use these deviation vectors in place of the original vectors, the result is that V1 and V2 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. For the next step of the analysis, I replaced the 794 segment vectors in just this way: I found their average and then replaced each vector with its deviation from this average. There are some more traditional methods that are employed to achieve a similar purpose. However, those methods were not successful when used in the present application. The success of the approach based on deviation vectors was discovered through an iterative process of tinkering—trying multiple algorithms in an attempt to get the vectors to cluster in a meaningful way. Ultimately it would be better to place the deviation algorithm on a more sound theoretical and mathematical footing, to better understand why it works and why other methods fail. A more extended discussion of deviation vectors and alternatives can be found in the Appendix.

Clustering the vectors. At this stage of the analysis, each of the 794 segments has been mapped to a vector that is understood to represent the meaning of that segment. The next step is to identify common meanings among these segments. To do that, I look for natural clusterings of the 794 vectors.

To cluster the transcript vectors, I employed the very general technique called *hierarchical agglomerative clustering.* In hierarchical agglomerative clustering, the analysis begins with each of the items in its own cluster. Thus, we begin with a number of clusters equal to the total number of items. Then two of those clusters are combined into a single cluster containing two items, thus reducing the total number of clusters by one. The process then iterates: Two clusters are combined 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. The results I describe 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).

Determining the number of clusters. The result of the clustering analysis can be thought of as a table with 794 rows. At the top is a row in which each segment is in its own cluster. At the bottom is a row in which all of the segments are in a single cluster. Table 2 displays the results for just a part of this large table. The bottom row, for example, shows the results when the segments are grouped into three clusters containing 271 segments, 279 segments, and 244 segments, respectively. Moving up the table, the number of clusters grows, and the size of each cluster shrinks. For example, in moving from the three-cluster row to the four-cluster row, the cluster that consisted of 244 segments has been split into two clusters, each with 122 segments. (It is just an accident that, in this case, the cluster has split precisely in half.)

In each row of Table 2, clusters contain segments that have been grouped together because, from the point of view of our vector space model, they have

TABLE 2
Number of Segments in Each Cluster for Various Cluster Numbers

No. of Clusters	*Sizes of the Clusters*
10	19 72 9 68 140 62 44 122 136 122
9	19 72 68 62 44 122 136 122 149
8	19 72 68 44 122 136 122 211
7	72 68 44 122 122 211 155
6	68 44 122 122 211 227
5	68 122 122 211 271
4	122 122 271 279
3	271 279 244

similar meanings. This means that each row in Table 2 constitutes a candidate coding scheme. In selecting one of these candidate schemes, there is a tradeoff to be negotiated. When the number of clusters is high, there is a better fit to the data. However, this better fit is obtained at the expense of a more complex model. Because each cluster is described by a list of 647 values, any of the usual methods for determining the appropriate number of clusters tends to always prefer the smallest number of clusters. Here, I therefore make my choice in a purely heuristic manner. Across multiple analyses, I have found that working with a set of about seven clusters strikes a workable balance. With seven clusters, it is possible to resolve interesting features of the data while producing results (in the form of graphs) that are not overly difficult to interpret. Furthermore, using anything between five and nine clusters leads to a story that is not substantially different.

What do the clusters mean?.

In order to make contact with theory and prior analyses, it is necessary to develop a means of associating the seven clusters selected with “ideas”—perhaps something akin to knowledge elements. To do so, we think of each of the seven clusters as defined by its centroid vector. These centroids are each described by a list of 647 entries, each of which corresponds to one of the words in the vocabulary. One way to attempt to understand the meaning of the clusters, then, is to look at the words that have the largest value in each centroid vector, as these are the words that are most characteristic of that cluster.

When this is done I obtain the results shown in Figure 5. For each cluster, I have listed the 10 words that are most strongly associated with that cluster. The second column in each table has the value from the centroid vector corresponding to this word. The third column in each table lists the total number of times that the word appears across the entire corpus.

The reader might note that these lists are already suggestive of interpretations of the clusters. But it will simplify our task if these lists are further pruned; namely, we will choose to simply ignore words that do not appear frequently in the larger

Cluster 1		
tilted	0.767	82
towards	0.199	40
away	0.186	83
tilt	0.166	21
north	0.098	30
part	0.084	46
kidding	0.083	3
pole	0.08	19
guess	0.077	31
toward	0.064	16

Cluster 2		
earth	0.4	395
spinning	0.366	37
china	0.295	14
united	0.202	23
spins	0.2	38
time	0.198	65
doesn	0.193	16
long	0.154	11
takes	0.138	17
axis	0.121	77

Cluster 3		
hemisphere	0.603	47
northern	0.522	31
sunlight	0.278	29
southern	0.229	19
direct	0.154	15
equator	0.146	26
colder	0.119	52
shorter	0.111	7
turning	0.106	18
facing	0.106	46

Cluster 4		
side	0.722	95
fall	0.273	26
spring	0.271	27
draw	0.107	21
sort	0.101	12
half	0.091	10
facing	0.091	46
earth	0.085	395
part	0.068	46
united	0.053	23

Cluster 5		
directly	0.382	27
rays	0.293	33
hitting	0.236	23
south	0.219	19
hit	0.212	27
north	0.197	30
angle	0.194	31
light	0.188	41
america	0.173	20
chicago	0.163	45

Cluster 6		
day	0.415	75
moon	0.398	52
night	0.377	63
rotates	0.178	54
make	0.089	18
turn	0.088	21
mind	0.077	6
rotating	0.068	32
planet	0.067	7
daytime	0.063	10

Cluster 7		
farther	0.413	71
closer	0.403	82
away	0.379	83
point	0.222	29
colder	0.216	52
cold	0.118	27
snow	0.107	9
sun	0.103	545
far	0.094	13
space	0.089	6

FIGURE 5 Top words associated with each cluster. The second column in each table displays the weight of the word within the centroid vector. The third column gives the total number of occurrences of the word across the entire corpus.

corpus. If we ignore words that appear less than 30 times in the corpus, we obtain the lists in Figure 6, which differ just slightly from those in Figure 5.

Interpreting the clusters based on the word lists. These word lists are the first results that we should examine in order to judge the success of the computational analysis. Here we can apply two of the criteria listed earlier.

Cluster 1		
tilted	0.767	82
towards	0.199	40
away	0.186	83
north	0.098	30
part	0.084	46
guess	0.077	31
closer	0.044	82
warmer	0.042	40
sun	0.03	545
farther	0.017	71

Cluster 2		
earth	0.4	395
spinning	0.366	37
spins	0.2	38
time	0.198	65
axis	0.121	77
seasons	0.068	30
tilted	0.031	82
angle	0.017	31
north	0.014	30
chicago	0.006	45

Cluster 3		
hemisphere	0.603	47
northern	0.522	31
colder	0.119	52
facing	0.106	46
closer	0.043	82
farther	0.035	71
warmer	0.023	40
axis	0.021	77
away	0.02	83
rays	0.018	33

Cluster 4		
side	0.722	95
facing	0.091	46
earth	0.085	395
part	0.068	46
chicago	0.018	45
guess	0.008	31
seasons	-0.008	30
time	-0.01	65
heat	-0.025	30
rotates	-0.026	54

Cluster 5		
rays	0.293	33
north	0.197	30
angle	0.194	31
light	0.188	41
chicago	0.163	45
sun	0.134	545
heat	0.076	30
towards	0.045	40
warmer	0.02	40
side	0.019	95

Cluster 6		
day	0.415	75
moon	0.398	52
night	0.377	63
rotates	0.178	54
rotating	0.068	32
earth	0.055	395
spins	0.05	38
facing	0.048	46
light	0.046	41
seasons	0.046	30

Cluster 7		
farther	0.413	71
closer	0.403	82
away	0.379	83
colder	0.216	52
sun	0.103	545
warmer	0.064	40
rotates	0.033	54
time	0.028	65
heat	0.02	30
rotating	0.013	32

FIGURE 6 Top words associated with each cluster, ignoring words that appear fewer than 30 times in the larger corpus. The second column in each table displays the weight of the word within the centroid vector. The third column gives the total number of occurrences of the word across the entire corpus.

1. *Can the computational analyses produce results that are interpretable in terms of entities that appear in our theory?* Several of the clusters seem to align with the three broad classes of seasons explanations discussed previously. For example, it seems natural to associate Cluster 1, which starts with the words *tilted, toward,* and *away,* with tilt-based explanations. Similarly, it seems natural to associate Cluster 4 (*side, facing*) with side-based explanations and Cluster 7 (*farther, closer*) with closer–farther explanations of the seasons. However, to make contact with the theory, we would prefer to associate the clusters with knowledge elements and to see explanations as built out of those elements. Looking at the lists in Figure 6, there are some ways in which one could imagine seeing clusters combined in the explanations given by students. Recall that side-based explanations were typically linked to the rotation of the earth. Thus, we might expect to see Cluster 4 (*side, facing*) together with Cluster 2 (*earth, spinning*) as part of a side-based DMC. It might also appear in conjunction with Cluster 6 (*day, night*), as the rotation of the earth explains our day/night cycle. Similarly, we might see Cluster 1 (*tilted, towards*) together with Cluster 7 (*farther, closer*), as in Caden's explanation. Cluster 1 might also be seen in tandem with Cluster 5 (*rays, north, angle, light*) in a tilt-based explanation that focuses on the angle of impact of rays of light.
2. *Are the identified knowledge elements at an appropriate grain size and level of abstraction?* The elements identified by these clusters seem to be close to the grain size identified by human analysts, though it might be more natural to associate them with small clusters of the elements identified by human analysts. For example, Cluster 7 (*closer, farther*) does seem to combine multiple components that would appear together in a typical closer–farther explanation of the seasons. The fact that the analysis only produces seven clusters means that it is not possible to capture all of the elements that appear in the human analysis.[2] Nonetheless, we will see that this is enough to capture some of the important nuances in the analysis of specific interviews.

Application to Segmented Transcripts

The next step in the analysis must be to introduce a means of producing analyses of individual interviews that can be compared with the human analyses. Here we are particularly interested in whether the computational analyses can capture dynamics such as shifts between DMCs. In order to accomplish this, I begin by

[2] Pushing the analysis to a larger number of clusters did not significantly add to the interpretable clusters. It is not clear at this point whether this is a limitation of the analysis or a result of the limited size of the data corpus.

preparing each of the interview transcripts precisely as before; the transcripts are reduced so that they include only the words spoken by a student, then they are broken into 100-word segments using a moving window that steps forward by 25 words. Next I compute the vector for each of these segments, again using the same techniques described earlier. Finally, each of these vectors for the segments is compared to the seven centroid vectors corresponding to the seven clusters (by taking the cosine of the angle between the vectors and each centroid).

I begin my discussion of the results with Zelda, because her analysis produces a graphic that is relatively easy to read. In Figure 7, we see that Zelda's transcript has been broken into five overlapping segments. Each of these segments is associated with seven bars, one bar for each of the seven clusters. Recall that Zelda's explanation focused on the earth's tilt and that it centered on the idea that the changing tilt causes parts of the earth to experience more or less direct sunlight. This is captured by the plot. The largest bar in all seven segments is the bar associated with Cluster 1 (*tilted, towards*). The bar associated with Cluster 5 (*rays, north, angle*) is also relatively large in a number of segments of the interview. This makes sense, as it captures the changing angle of rays of sunlight and hence their directness. Finally, the bar associated with Cluster 4 (*side, facing*) is comparatively large in the first segment. As noted earlier, Zelda's initial explanation also included talk about the side facing the sun: "I think because the earth is on a tilt, and then, like that side of the earth is tilting toward the sun, or it's facing the sun or something so the sun shines more directly on that area, so it's warmer."

The interview with Caden provides an interesting contrast. Recall that Caden also gave a tilt-based answer. However, unlike Zelda, he maintained that the changing tilt causes one hemisphere or the other to be closer to the sun. When Caden's transcript is analyzed using the segment centroids, we get eight segments with the bars shown in Figure 8. This plot is quite different than Zelda's. The largest bar across most of the segments is the one associated with Cluster 3 (*hemisphere, northern*). However, Cluster 1 (*tilted, towards*) is relatively large in

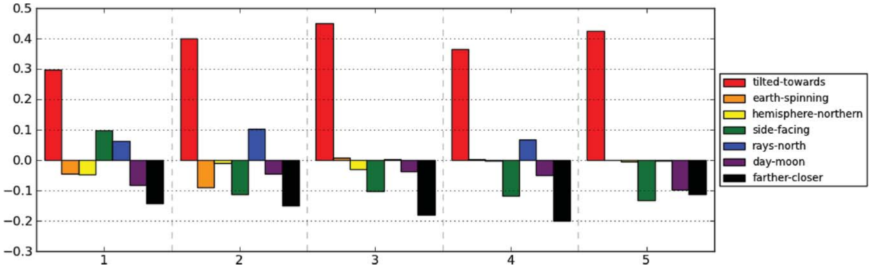


FIGURE 7 Segmenting analysis of Zelda's transcript (color figure available online).

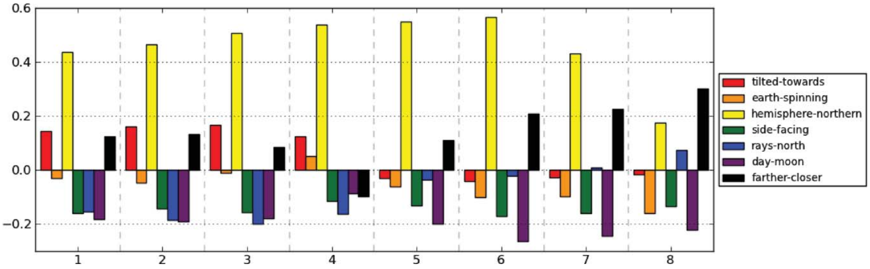


FIGURE 8 Segmenting analysis of Caden's transcript (color figure available online).

several segments. In addition, Cluster 7 (*closer, farther*) is large in some segments. Taken together, these features capture the essence of the human-based qualitative analysis and also capture the major differences between Caden and Zelda: most notably that the changing tilt of the earth leads to parts of the earth being close to or farther from the sun, rather than changing the directness of the sun's rays. Thus, we are seeing here some evidence that the computational analysis is capable of capturing combinations of elements and in a way that resolves important differences captured by human analyses (Criteria 3).

Both Zelda and Caden were relatively stable in the explanations that they gave. It is critical to determine whether the automated analysis can capture shifts that occur in some interviews (Criteria 4). Here we return to the interview with Marcus. Looking at Figure 9, it seems clear that the interview has at least two major parts. Note that the first part of the interview is dominated by Cluster 7 (*farther, closer*). The end of the interview is dominated by Cluster 2 (*earth, spinning*). The middle portion of the interview is jointly dominated by Cluster 2 and Cluster 4 (*side, facing*).

These results align well with the description of this interview presented earlier. Recall that Marcus began by giving a closer–farther explanation of the seasons

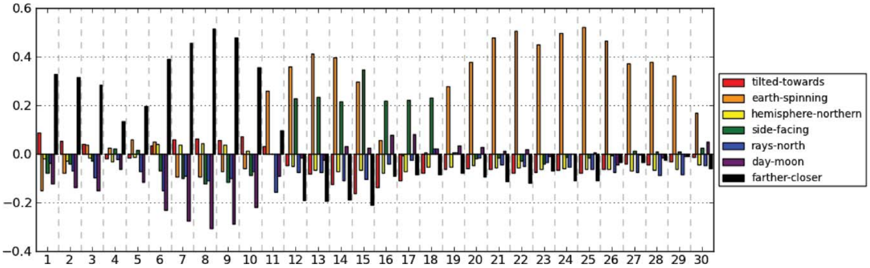


FIGURE 9 Segmenting analysis of Marcus's transcript (color figure available online).

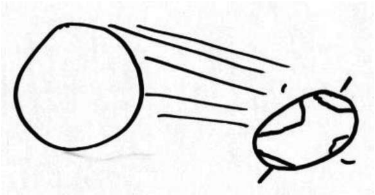


FIGURE 10 Edgar's drawing.

that included an emphasis on the orbital motion of the earth. Consistent with this, the bar corresponding to Cluster 7 (*closer, farther*) dominates during the first part of the interview.

When Marcus was challenged, he abandoned that explanation, and he attempted to construct a new explanation based on the fact that the earth rotates, so that one side or the other faces the sun (though he was never successful in constructing such an explanation). This challenge first appears in Segment 11 of the transcript.[3] Consistent with this, the plot in Figure 9 shows a clear shift at Segment 11. Over the segments that follow, the computational analysis is dominated by Cluster 2 (*earth, spinning*) and Cluster 4 (*side, facing*). Finally, starting at the end of Segment 17, Marcus began work on a new drawing. Here the emphasis was on trying to work out how the earth moves, with some emphasis on a comparison between the United States and China. In the computational analysis, we see that Cluster 2 (*earth, spinning*) dominates during this last part of the interview. This makes sense because there is a focus on the motion of the earth here. Note also that, as shown in Figure 5, the words *china* and *united* are strongly associated with Cluster 2.

We thus have some evidence that the computational analysis is capable of capturing interview dynamics (Criteria 4). Here, however, we are seeing shifts that occurred primarily because of relatively strong intervention from the interviewer. It is possible that such shifts would be particularly easy to capture, perhaps because the interviewer could change the words that are in play. Here I present one final example in which a student shifted explanations but without strong prompting from the interviewer. In this example, a student, Edgar, began by giving an explanation focused on the fact that the earth rotates, and he stated that light would hit more directly on the side facing the sun. He made the drawing shown in Figure 10 as he commented:

E: Here's the earth slanted. Here's the axis. Here's the North Pole, South Pole, and here's our country. And the sun's right here [draws the circle on the left], and

[3] Because segments include overlapping portions of the transcript, the challenge appears in more than one segment.

> the rays hitting like directly right here. So everything's getting hotter over the summer and once this thing turns, the country will be here and the sun can't reach as much. It's not as hot as the winter.

When Edgar recalled that the earth orbits the sun in addition to rotating, he shifted to a closer–farther type explanation:

> I: Let's say we're here and it's summer, where is it, where will the earth be when it's winter?
>
> E: Actually, I don't think this moves [indicates earth on drawing] it turns and it moves like that [gestures with a pencil to show an orbiting and spinning earth] and it turns and that thing like is um further away once it orbit around the s- earth- I mean the sun.
>
> I: It's further away?
>
> E: Yeah, and somehow like that going further off and I think sun rays wouldn't reach as much to the earth.

The automated analysis for Edgar's interview is shown in Figure 11. The interview is clearly split into two parts, with a change occurring at Segment 5. Looking at the segmented transcript reveals that this corresponds to the point at which Edgar shifted his explanation. Thus, once again, the analysis seems to be capturing dynamics within an interview in a way that aligns with human analysis.

However, there are some respects in which the alignment is not perfect. As one would expect, the latter part of the plot is strongly dominated by Cluster 7 (*farther, closer*). However, the first part is dominated only by Cluster 5, which has to do with rays striking the earth's surface. This is reasonable, but given the nature of Edgar's explanation, we might also expect to see contributions from Cluster 2 (*earth, spinning*) and Cluster 4 (*side, facing*). Thus, the first part of Edgar's interview indicates a mild failure of the computational analysis to capture a potentially important feature of student reasoning. (In this case, it is likely in part due to the fact that the computational analysis lacks access to drawings and gestures.)

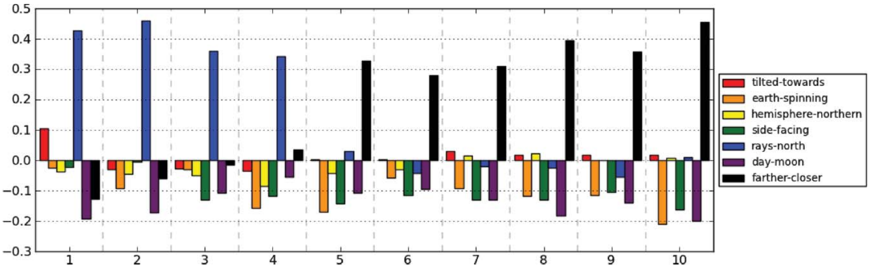


FIGURE 11 Segmenting analysis for Edgar's transcript (color figure available online).

A Look at the Larger Corpus

A final step is to show that the four interviews described here are representative of the larger corpus of 54 interviews. This is difficult to do because each interview can have its own complex story. In the online supplemental materials, I have included short analyses of 24 of the 50 interviews not discussed here. To narrow the group, I excluded the 18 interviews that were conducted following a curriculum on related subject matter. In addition, I excluded interviews that were longer than 25 segments because of the difficulty of reproducing the plots in a manner that is easily readable.[4]

The short analyses in the online supplemental materials provide further examples of the features described here. There are multiple cases in which simple DMCs are captured (e.g., Angela, Cassie, Denise, Leslie, Lisa, Robbie, Samantha, Vanessa) and examples with DMCs consisting of combinations of multiple elements (e.g., Alex, Amanda, Beth, Blake, Jill, Ovadya, Richard). In addition, there are examples in which shifts in explanations or shifts in the topic under discussion are captured (e.g., Alex, Angela, Cassie, Candice, Kelly, Samantha).

There are, however, some cases in which all or part of the results of the computational analysis are difficult to interpret (e.g., Ali, Deidra, Kelly, Kim, Kurt, Mark, Randy, William). Most of these cases were equally unclear for human analysts (e.g., Ali, Deidra, Kelly, Kim, Kurt, Randy). In some cases, weak English was a factor (Kim, Mark, William).

Finally, there were some failures of the computational analysis to capture features of interviews that were clear and important to the interviews. This was most true for explanations that were based around unusual motions or in which the students used wrong or idiosyncratic words to describe their spatial model (e.g., Blake, Kim, Richard, Robbie). For example, Blake used the word *side* to refer to the hemispheres. Robbie described an unusual motion, in which the earth's axis is in the plane of its orbit and the earth spins so that first one pole, then the other, is pointed toward the sun.

FURTHER EXPLORATIONS

There are some important respects in which the account presented in the previous section was idealized: I presented the results I obtained with one set of algorithms and with a single set of input parameters. In reality, some search was required in order to discover algorithms and parameters that produced interpretable results.

[4] Although detailed analyses of individual interviews for the remaining 26 transcripts are not presented here or in the online supplemental materials, the reader is reminded that these transcripts nonetheless figured in the analysis that produced the centroid vectors reported in Figure 5 and Figure 6.

In this section, I want to briefly give a sense for the results produced by alternative approaches, including some that did not produce interpretable results.

The analysis followed the following plan: (a) The transcripts were pruned so that they only contained the words spoken by the interviewee; (b) the resulting documents were broken into overlapping segments; (c) a vector was computed for each segment, using a particular weighting function and ignoring words in my stop list; (d) the resulting vectors were replaced with their deviation vectors; and (e) the vectors were clustered using hierarchical agglomerative clustering. In seeking to explore more broadly, we could consider small changes to the procedure I followed—changes that are still consistent with this five-step plan. We can also consider more significant changes.

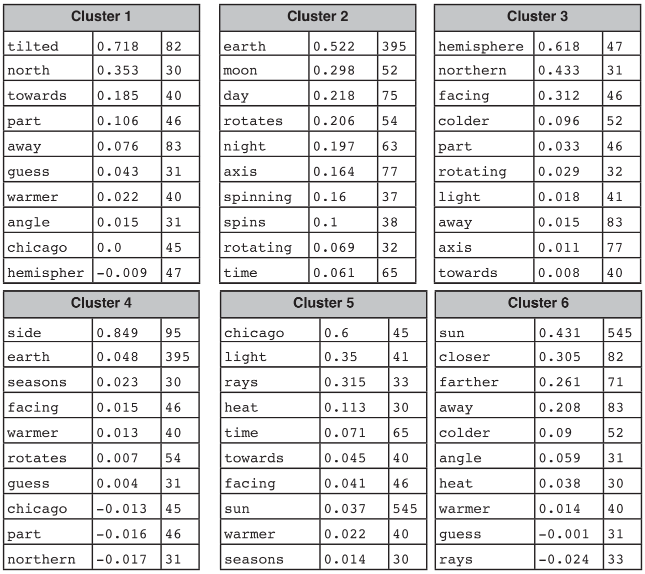
First let us consider smaller changes. Even in the first step—the pruning of transcripts—there were many choices to be made. For example, I had to decide what to do with word fragments. I also could have opted to stem words, that is, to reduce them to their base or root. I found that these changes had little impact on the results I obtained.

It is illustrative to consider, in some detail, what happens if the second step is modified. Recall that I broke each of the transcripts into 100-word segments with a step size of 25 words. Here, for comparison, I look at what happens if I perform the same analysis but employing 50-word segments and a step size of 10. When the transcripts are segmented in this new way, I obtain 2,320 segments. These 2,320 segments can then be clustered as before, producing the results shown in Table 3 . Like Table 2, this table shows the sizes of the clusters when the vectors are grouped into different numbers of clusters.

In my earlier analysis I selected the row of Table 2 that contained seven clusters. However, note that in Table 3 this row has a cluster with only a single segment. Thus, this cluster does not correspond to a conception that appears with any frequency in the corpus. For that reason, it makes sense to select the row with

TABLE 3
Sizes of Clusters for Selected Clusterings

No. of Clusters	*Sizes of the Clusters*
10	1 78 88 160 156 11 628 235 137 638
9	1 88 160 156 11 628 235 638 215
8	1 160 11 628 235 638 215 244
7	1 160 235 638 215 244 639
6	160 235 638 215 639 245
5	235 638 215 639 405
4	638 639 405 450
3	638 450 1,044



Cluster 1		
tilted	0.718	82
north	0.353	30
towards	0.185	40
part	0.106	46
away	0.076	83
guess	0.043	31
warmer	0.022	40
angle	0.015	31
chicago	0.0	45
hemispher	-0.009	47

Cluster 2		
earth	0.522	395
moon	0.298	52
day	0.218	75
rotates	0.206	54
night	0.197	63
axis	0.164	77
spinning	0.16	37
spins	0.1	38
rotating	0.069	32
time	0.061	65

Cluster 3		
hemisphere	0.618	47
northern	0.433	31
facing	0.312	46
colder	0.096	52
part	0.033	46
rotating	0.029	32
light	0.018	41
away	0.015	83
axis	0.011	77
towards	0.008	40

Cluster 4		
side	0.849	95
earth	0.048	395
seasons	0.023	30
facing	0.015	46
warmer	0.013	40
rotates	0.007	54
guess	0.004	31
chicago	-0.013	45
part	-0.016	46
northern	-0.017	31

Cluster 5		
chicago	0.6	45
light	0.35	41
rays	0.315	33
heat	0.113	30
time	0.071	65
towards	0.045	40
facing	0.041	46
sun	0.037	545
warmer	0.022	40
seasons	0.014	30

Cluster 6		
sun	0.431	545
closer	0.305	82
farther	0.261	71
away	0.208	83
colder	0.09	52
angle	0.059	31
heat	0.038	30
warmer	0.014	40
guess	-0.001	31
rays	-0.024	33

FIGURE 12 Top words associated with each cluster. The second column in each table displays the weight of the word within the centroid vector. The third column gives the total number of occurrences of the word across the entire corpus.

six clusters. Figure 12 shows the top words associated with these six clusters. As before, words that appeared less than 30 times in the corpus are omitted.

Finally, we can compare the word lists in Figure 12 and Figure 6. Although there are many differences, it is not difficult to discern an alignment. Clusters 1, 3, 4, and 5 in the new analysis seem to be similar to the corresponding clusters in the original analysis. Cluster 6 seems to be similar to Cluster 7 in the original analysis. Lastly, Cluster 2 in the new analysis is similar to both Cluster 2 and Cluster 6 in the original analysis. This makes sense because, in the original analysis, Clusters 2 and 6 both pertain to the rotation of the earth. Thus, although there are differences, the qualitative picture produced by this new analysis seems to bear a close resemblance to the original analysis.

It is of course possible to consider explorations that depart more significantly from the plan of analysis described previously. We could, for example, employ a different clustering algorithm. I tried several, without getting better results.[5] In the Appendix I show what happens if I omit the use of deviation vectors. There I show that this modified analysis does not produce interpretable results.

[5] See, for example, Manning et al. (2008) for a discussion of some of these alternative clustering techniques and similar applications in computational linguistics.

Why Does This Work?

For many reasons I believe that we should be surprised by the success of the computational methods presented here. As I noted earlier, even human analysts can have difficulty understanding what students say during clinical interviews. Student utterances are often halting and ambiguous, and they can use gestures and diagrams in a manner that can be difficult to understand. Furthermore, the computational analyses begin with a significant handicap. They have no access to students' facial expression, gestures, and drawings. In addition, the specific techniques I employed discard *additional* information; they discard most information about word order. Furthermore, from among the bag of words models that are employed by linguists, I chose to begin with one of the simplest possible models.

For all of these reasons, we would like to understand how it is that a simple algorithm can produce meaningful results. Some further reflection and exploration can help provide a sense for why the algorithms work. First, we should keep in mind that my analysis algorithms do not need to deconstruct and interpret *any* utterances in *any* context. Student speech was restricted to a discussion of the seasons and was only a few minutes long. Some of this simplification is highlighted if we look at sample transcripts and how they were processed. Earlier I presented the first 100 words of Marcus's interview, stripped of interviewer utterances, punctuation, and annotations. But, prior to further analysis, these 100 words are reduced even further, as words that do not appear in the vocabulary, pruned by the stop list, are simply ignored. The result is the following reduced transcript:

> alignment sun sun plan sun planets sun depends earth farther away sun closer sun

Note that the text has already been reduced to a set of words that seem to have a close relationship to explanations of the seasons.

Second, I have described my analysis as "automated." However, the human analyst still plays an essential role. I called on the reader to interpret the list of words associated with each cluster, and I hoped that the reader would agree that these lists are suggestive of certain interpretations—interpretations that aligned with the work of human analysts. But the leap from word lists to interpretations is nontrivial; it is real and important work that is done by the human analyst.

Thus, the magic of the automated analyses is not just located in the computational algorithms. Instead, it is a property of the experimental design as a whole. The interviews are set up so as to radically constrain the nature of the utterances students will make. Furthermore, our own understanding of the nature of the experimental setup constrains our interpretation of the lists of words produced by the clustering analysis and the analysis we obtain when the clusters are applied to specific transcripts.

Thus, by looking closely at how the analysis plays out, it is possible to greatly reduce one's surprise that the analyses produce meaningful results. Nonetheless, I believe that the relative success of the computational methods still raises interesting questions for all types of analysis of clinical interview data. For example, it might help us better understand the interpretive task faced by human analysts as they seek to interpret interview data.

DISCUSSION

In this article, I described my attempts to apply computational techniques to the analysis of data that have been a traditional focus in the learning sciences. This work differs, in many respects, from prior applications of vector space text analysis methods in education. Prior work has most typically used vector space methods to evaluate student responses, or to apply preexisting coding schemes, based on presumed-to-be established theory. In contrast, I did not attempt to evaluate responses or to apply an existing coding scheme; I attempted to *discover* ideas in the corpus and to capture the dynamics in student reasoning.

Furthermore, I did not adopt the stance that the principal promise of these methods is their potential to reduce the labor of human analysts. Instead, my claim has been that there is a more important potential benefit, the potential for the computational methods to provide convergent evidence, supporting the work of human analysts.

The implication of this stance is that we must see ourselves as engaged in the first steps of a bootstrapping program. Thus, my aim in this article was limited to establishing the plausibility that this program would ultimately be successful. To do so, I presented four broad criteria against which to evaluate the success of the methods employed. I looked to see whether the computational methods produced results that (a) were interpretable in terms of the theory, (b) captured knowledge at an appropriate grain size, (c) captured combinations of elements, and (d) captured the dynamics of individual interviews. In all cases, the results I presented are only suggestive. Nonetheless, I believe that they are sufficient to establish that it is worthwhile to pursue this bootstrapping program further. The success of the bootstrapping program will confirm the usefulness of the computational methods. In addition, it may ultimately contribute to our confidence in the larger theoretical and empirical program: the use of a KiP approach to understand the construction of scientific explanations during interviews. The success reported in this article is one step toward that goal.

Open Issues and Next Steps

A large number of problems remain unsolved, some of which I have been careful to highlight and others of which I have glossed over. One important unsolved problem is how to determine the number of clusters when using the clustering algorithms. Unfortunately, there does not seem to be a simple solution to this problem. As I discussed earlier, we are 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 model complexity. Ultimately, I believe that the best that I might be able to do will be to introduce a tunable parameter that represents a judgment about how to make this tradeoff (as discussed in Manning et al., 2008).

Future work could explore the range of potential algorithms beyond what I have already tried. For example, it should be possible to replace my rudimentary technique for segmenting transcripts with a method that finds boundaries in a more principled manner (e.g., Hearst, 1997). Moreover, my simple vector space models could be replaced with more sophisticated models that make use of a training corpus, such as LSA. Finally, there are methods of analysis that depart still more dramatically from the core approach applied here, such as probabilistic latent semantic indexing (Hofmann, 2001) and latent Dirichlet allocation (Blei, Ng, & Jordan, 2003).

In addition to refining the computational analyses and trying alternative methods, future work should also seek to test the methods in new ways. One next step will be to employ 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 discussions of the seasons, a small number of key words (e.g., *tilt*) can do a lot of the work of discriminating among explanations.

Moreover, my presentation of results in this article was not fully systematic. Ultimately, it will be necessary to seek an analysis in which I systematically compare the output of the automated analysis to codes produced by human analysts. As I discussed earlier, the earlier work by Dam and Kaufmann (2008) performed an analysis of this sort. But their analysis assigned a single code to every student. This simplification worked reasonably well on a smaller corpus, but it has become less and less tenable as our data corpus has grown. Thus, we need to look toward a match between human and automated analyses that is at a smaller grain size.

Finally, it remains to be seen whether methods of the sort described here can be used in the service of real research questions. My goals in this article were primarily methodological: I attempted to give the reader a sense of what is possible with these new computational methods. But I did not attempt to answer new research questions. If and when we learn something new and surprising using these methods, that will provide the most convincing evidence of their power and utility.

REFERENCES

Atwood, R. K., & Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching*, *33*, 553–563.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, *37*, 573–595.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In R. Lesh & A. Kelly (Eds.), *Handbook of research methodologies for science and mathematics education* (pp. 341–385). Hillsdale, NJ: Erlbaum.

Dam, G., & Kaufmann, S. (2008). Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods*, *40*(1), 8–20.

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

diSessa, A. A. (2007). An interactional analysis of clinical interviewing. *Cognition and Instruction*, *25*, 523–565.

D'Mello, S., Graesser, A. C., & King, B. (2010). Toward spoken human-computer tutorial dialogues. *Human-Computer Interaction*, *25*(4), 289–323.

Duit, R. (2009). *Bibliography: Students' and teachers' conceptions and science education*. Kiel, Germany: Leibniz Institute for Science Education.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197–202.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996, January). *Reasoning from multiple texts: An automatic analysis of readers' situation models*. Paper presented at the 18th Annual Cognitive Science Conference, La Jolla, CA.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*(2), 111–127.

Foltz, P. W., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*, 285–307.

Foltz, P. W., Laham, D., & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). Retrieved from http://imej.wfu.edu/articles/1999/2/04/index.asp

Ginsburg, H. P. (1997). *Entering the child's mind: The clinical interview in psychological research and practice*. New York, NY: Cambridge University Press.

Graesser, A. C., Lu, S., Jackson, G. T., & Mitchell, H. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods*, *36*(2), 180–192.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*(2), 193–202.

Graesser, A. C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 129–147.

Hastings, P., Hughes, S., Magliano, J., Goldman, S., & Lawless, K. (2011). Text categorization for assessing multiple documents integration, or John Henry visits a data mine. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: Proceedings of the 15th International Conference, AIED 2011* (pp. 115–122). Berlin, Germany: Springer.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, *23*(1), 33–64.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1), 177–196.

Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., & Millis, K. K. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 244–250.

Landauer, T., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T., Laham, D., & Foltz, P. W. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy and Practice*, *10*(3), 295–308.

Lee, V. R. (2010). How different variants of orbit diagrams influence student explanations of the seasons. *Science Education*, *94*, 985–1007.

Lelliott, A., & Rollnick, M. (2010). Big ideas: A review of astronomy education research 1974-2008. *International Journal of Science Education*, *32*, 1771–1799.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*(3), 251–283.

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Munoz, B. D., & McNamara, D. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 181–188.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 289–324). Hillsdale, NJ: Erlbaum.

Millis, K., Kim, H.-J. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 213–221.

Newman, D., & Morrison, D. (1993). The conflict between teaching and scientific sense-making: The case of a curriculum on seasonal change. *Interactive Learning Environments*, *3*, 1–16.

Sadler, P. M. (1987). Alternative conceptions in astronomy. In J. D. Novak (Ed.), *Second International Seminar on Misconception and Educational Strategies in Science and Mathematics* (Vol. 3, pp. 422–425). Ithaca, NY: Cornell University Press.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science*, *21*(2), 147–177.

Shapiro, A. M., & McNamara, D. S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research*, *22*(1), 1–36.

Sherin, B. (2001). How students understand physics equations. *Cognition and Instruction*, *19*, 479–541.

Sherin, B., Krakowski, M., & Lee, V. R. (2012). Some assembly required: How scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching*, *49*(2), 166–198.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3*, 115–163.

Trumper, R. (2001). A cross-college age study of science and nonscience students' conceptions of basic astronomy. *Journal of Science Education and Technology*, *10*(2), 192–195.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction*, *22*, 333–362.

Wellman, H. M., & Johnson, C. N. (1982). Children's understanding of food and its functions: A preliminary study of the development of concepts of nutrition. *Journal of Applied Developmental Psychology*, *3*, 135–148.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, *25*, 309–336.

APPENDIX

Deviation Vectors Explained

Here I describe the calculation of deviation vectors in more detail, drawing out some of the less obvious effects. The procedure begins with a set of vectors, $\bar{v}_i$, that have been normalized. (Normalizing a vector results in a vector that points in the same direction but has a length of 1.) To deviationalize these vectors, we first find their sum, $\bar{v}_{sum}$:

$$\bar{v}_{sum} = \sum_i \bar{v}_i.$$

Then we normalize this vector to produce a vector I call $\bar{v}_{avg}$:

$$\bar{v}_{avg} = norm\left(\bar{v}_{sum}\right).$$

Next we take each original vector $\bar{v}_i$ and subtract its component along $\bar{v}_{avg}$. Then we normalize the resulting vector to produce the deviationalized vector, $\bar{v}'_i$:

$$\bar{v}'_i = norm(\bar{v}_i - (\bar{v}_i \cdot \bar{v}_{avg})\bar{v}_{avg})$$

This final step has some important implications that might not be initially obvious. When a vector $\bar{v}_i$ is orthogonal or almost orthogonal to $\bar{v}_{avg}$, then it will be left nearly unchanged. However, vectors that point in a direction that is close to $\bar{v}_{avg}$ will be dramatically altered. For example, consider two vectors $\bar{v}_1$ and $\bar{v}_2$ that both point in nearly the same direction as $\bar{v}_{avg}$ and hence are very similar to each other. When we subtract the components of $\bar{v}_1$ and $\bar{v}_2$ along the average using $\bar{v}_i - (\bar{v}_i \cdot \bar{v}_{avg})\bar{v}_{avg}$, we obtain two new vectors that are both small. Now let us imagine that these two new vectors point in exactly opposite directions. When the vectors are then normalized, the result will be two vectors of unit length in opposite directions and thus very far apart. For this reason, the deviationalizing procedure can behave like a threshold effect. The effect can, in some cases, be to dramatically separate vectors that differ only in a small way from the average.

Table A1 shows the results that are produced when I do not compute deviation vectors before clustering the 794 segments from the original analysis. Note that in each of the candidate clusterings shown, I obtain one very large cluster, containing most of the segments, and several very small clusters. When I look at earlier stages of the clustering (which are not shown in the table) I see the following behavior: Initially the segments are all clustered into a large number of relatively small clusters, then the these small clusters begin to agglomerate, one at a time, into one large cluster. In the final stages, which we see in Table A1, some small remaining

TABLE A1
Sizes of Clusters for Selected Clusterings

No. of Clusters	*Sizes of the Clusters*
10	2 2 3 4 3 4 9 6 11 750
9	2 2 3 4 3 4 6 11 759
8	2 2 3 4 3 4 6 770
7	2 2 3 3 4 6 774
6	2 3 3 4 6 776
5	2 3 3 4 782
4	2 3 4 785
3	3 4 787

clusters are swept up into the large cluster. In short, this analysis does not seem to discover a small number of moderately sized clusters that we can associate with conceptions. Thus, it does not seem that we can simply eliminate the use of deviation vectors.

The question remains as to whether there are other more standard approaches that might improve on the results shown in Table A1. In particular, it is standard practice to use judiciously chosen weighting functions as a means of accentuating the differences among documents. Recall that the counts in my vectors were all weighted by (1 + log[count]), where "count" is the number of times a word appears in a given document. We can modify this function so that it weights words that appear across many documents less strongly than words that appear in just a few documents. I tried several such weighting functions, including variants of the so-called tf-idf weighting. In all cases, I obtained results that looked like Table A1.