NLP: Getting Computers to Understand Discourse

Danielle S. McNamara, Laura K. Allen, Kathryn S. McCarthy, & Renu Balyan

Natural language processing (NLP) is the automated analysis of human language using computers. It is called *natural* to make the contrast to the analysis of computer languages (e.g., computer science), which dominated the field when NLP emerged in the mid-1990s (Jurafsky & Martin, 2008). NLP tools provide the means to extract various aspects of language, where the unit of analysis might be the words themselves, or features of the words (e.g., part-of-speech, frequency, concreteness), the sentences (e.g., syntactic complexity), or the text as a whole (e.g., cohesion). Prior to the use of NLP, these types of linguistic and semantic features were examined using hand-coded discourse analysis, usually targeting a few specific features at a time.

Discourse analysis has provided a wealth of information about language, behaviors, psychological processes, cognition, and social interactions, but it is laborious, time consuming, and subject to human biases. NLP is automated, and therefore quick and reliable. It provides the means to *get computers to understand discourse* by extracting features of language from text, and then applying algorithms (i.e., mathematical combinations of linguistic and semantic features) to develop predictions about humans’ meaning and intentions conveyed through their language (for a review, see McNamara, Allen, Crossley, Dascalu, & Perret, 2017).

This purpose of this chapter is to describe how NLP has been used to tap into students’ deep comprehension of complex concepts. Here, we define deep comprehension as a students’ understanding of a concept that goes beyond the surface-level components provided in a text or other form of language. To deeply comprehend, students must establish relations between pieces of information that are presented to them, as well as to their prior knowledge of related concepts. In this chapter, we focus on how NLP has been used to measure deep comprehension of concepts in the context of intelligent systems that interact with humans for educational purposes, as well as for the purpose of automating discourse analysis to better assess and understand human behavior.

**Intelligent Systems**

ELIZA was the first computer program that prompted its users to ‘*tell me more about that*’ (Weizenbaum, 1966). The program was designed to play the part of a Rogerian psychotherapist in a dialogue. Many of the people who interacted with ELIZA began to anthropomorphize the agent. Using surface and syntactic cues from language (i.e., words), “she” simulated understanding human discourse regarding daily life and potential emotional problems. ELIZA never successfully treated significant clinical problems, such as depression or neuroses, nor was she able to successfully pass the Turing test. Nonetheless, the success of this program demonstrated the power of using relatively simple NLP in lulling its users into believing there was a *human behind the box*. Three decades later in the late 1990s, Art Graesser and his team were the first to take this approach further -- to lull *students* into believing there was an *instructor* behind the box (Graesser et al. 1999, 2004a; see Figure 1). Using NLP in combination with cognitive and pedagogical principles, his team developed principles of discourse interaction that revolutionized the field of discourse-interactive intelligent tutoring systems (ITS) and applied NLP.
AutoTutor was the first pedagogical agent behind the box. A pedagogical agent engages students in interactive dialogues driven by a combination of NLP and discourse rules with the objective of providing instruction to learn some content or skill. A full discussion of the cognitive and pedagogical underpinnings of AutoTutor is well beyond the scope of this chapter (see Cai & Hu, chapter 11; Olney chapter 12). In brief, AutoTutor is an ITS that provides tutoring on various concepts (e.g., physics, computer literacy, critical thinking) through student-agent dialogue within a computer-based environment. Drawing upon constructionist approaches proposed by Graesser and colleagues (1994), AutoTutor’s dialogues encourage students to generate questions and inferences that aid in the construction of more elaborated and durable mental models which in turn lend to deeper, conceptual understandings of complex material (Graesser, VanLehn, Rosé, Jordan, & Harter, 2001). AutoTutor mimics human tutoring interactions using instructional methods such as hints, prompts, and elaborations based on student models of learning progressions (Graesser et al., 2004a). 

AutoTutor was revolutionary in NLP for intelligent tutoring in that it leveraged latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998). LSA was important to advancing intelligent systems in general because it incorporates mathematical algorithms to extract the semantic relations between words. Thus, the introduction of LSA allowed researchers to go beyond shallow, word-based comprehension to model key components of deep learning processes – the integration of new to-be-learned information with prior knowledge, which relies on inference generation. AutoTutor accomplished this by using LSA to compare the semantic similarity between a student’s response and benchmark responses (i.e., previously collected responses that are expert-coded in terms of accuracy, type of response, or quality). For example, if a student’s response in AutoTutor is semantically similar to a benchmark response reflecting a shallow or incorrect understanding of a targeted topic, and less similar to a response reflecting a deep understanding, then AutoTutor might provide the student with a hint and a prompt.

While word-based (non-latent) analyses may be able to detect when students’ responses use information from a particular source, the words alone cannot distinguish new words as elaborating on ideas in the text versus completely irrelevant ideas (see Magliano, Higgs, & Millis, chapter X). LSA affords assessments of the extent that students bring in new, but
conceptually-related information, which is indicative of students’ generation of high-quality inferences, and deep learning.

LSA assumes that words that appear together or in similar documents are semantically related and therefore represent similar concepts. For example, the words cake and pie are not synonymous, but they are likely to appear together in texts or appear in similar contexts because they are conceptually related to baking. The word cake may also be found more often in texts that contain words such as birthday or ice cream, but probably less so in texts including the words traffic or blockade. LSA generates a matrix of all word counts across paragraphs and texts and uses a mathematical procedure called singular value decomposition (SVD) to construct vectors (i.e., a list of values) for each word, indicative of contexts in which the word appears, and in turn its semantic meaning. As such, LSA is not limited by exact words, which affords responses to students’ natural language, which can be vague or unexpected and contain poor grammar and syntax.

Extracting the latent meanings of words is a key ingredient of getting computers to understand discourse (McNamara, 2011). But that is not sufficient to drive pedagogical agents. To discern students’ understanding and performance using their natural language responses, algorithms must combine multiple aspects of language, discourse, and cognition. The unique challenge lies in the fact that NLP tutors must understand language to the extent that they can provide accurate evaluations of open-ended responses and subsequently provide accurate and meaningful feedback. Such complex moves require more than simple production rules used by early conversational agents (e.g., ELIZA, PARRY) and, thus, rely on the implementation of more sophisticated NLP tools. AutoTutor was the first pedagogical system to use non-structural approaches to language processing, relying critically on words and statistics rather than specific grammars or sets of logic rules (Boonthum, Levinstein, McNamara, Magliano, & Millis, 2009). As such, Graesser’s creation of AutoTutor was revolutionary. Indeed, the incorporation of LSA and the development of discourse principles for agent-based pedagogical dialogue spawned multiple generations of natural language tutoring systems (Figure 2).

![AutoTutor Family tree](image)

Figure 2. AutoTutor Family tree (from Nye, Graesser, & Hu, 2014)
Building on the foundation set by Graesser, NLP has been incorporated into a number of ITSs, particularly those that prompt the student to generate verbal responses (e.g., iSTART: McNamara, Levinstein, & Boonthum, 2004; Writing Pal: McNamara et al., 2012; Roscoe & McNamara, 2013). For example, iSTART provides students with instruction and practice on strategies for reading complex informational texts. Students are prompted to generate self-explanations, which restate what a text means in their own words (paraphrasing) and then link the explicit text to prior text (i.e., making bridging inferences) or connect the text with information outside of the text (i.e., generating elaborations). iSTART provides instruction on five comprehension strategies: comprehension monitoring, paraphrasing, predicting, bridging, and elaboration. The quality of students’ self-explanations is then assessed using an algorithm that relies on both word-based and LSA indices. Word-based indices are used to identify particular metacognitive expressions indicative of comprehension monitoring (e.g., “I understand that”, “I am confused”) or that a self-explanation is too short or too similar to the text. LSA indices are used to assess the extent that the student generates inferences and constructs a deeper understanding of the text. The algorithm produces a quality score from 0 to 3 and provides feedback for the student to improve. An advantage of LSA is that the algorithms can be text general. That is, researchers can use the same algorithm for analysis of responses to multiple texts and topics. In iSTART, this means that researchers and educators can easily add their own texts to the system and the algorithm provides appropriate evaluation and feedback across a wide range of texts and topics (Jackson, Guess, & McNamara, 2010; McNamara, Boonthum, Levinstein, & Millis, 2007).

In sum, NLP is a key to intelligent adaptivity in educational systems that include students’ natural language responses (McNamara, Crossley, & Roscoe, 2013). Learning can take place in and across a variety of contexts -- when students read their textbooks, while engaging in class lectures and group discussions, and in the context of online discussions and blogs. In the majority (if not all) of these scenarios, students are required to process complex language and ultimately communicate to-be-learned material to themselves and others. While often ignored, language is an undeniably crucial factor during learning processes, and in turn educational research and applications. With the implementation of NLP, educational technologies gained the capacity to move beyond point-and-click problem-solving tutors affording conversational dialogues and interactions that better simulate real-world learning environments, and also enhance learning by increasing students’ generation of questions and answers, with feedback.

**Using NLP to Understand Discourse Processes**

One important objective that has been tackled by discourse processing researchers is to understand the factors that support a deep and robust understanding of content. Comprehension can be described in terms of multiple levels. Researchers often distinguish between literal and inferential comprehension or shallow and deep comprehension. A shallow (literal) comprehension of text is generally focused on the surface-level features of the text itself as compared to a deep (inferential) comprehension characterized by a more integrated and coherent representation of the text material. A coherent understanding requires an understanding of information explicit in the text, inferences that connect disparate parts of the text and text content to prior knowledge, as well as explanations or questions about why particular events occur (Graesser, Singer, & Trabasso, 1994; Kinstch & van Dijk, 1978).
Discourse researchers aim to capture differences in readers’ levels of understanding as well as investigate how these differences emerge. The amount and what a reader ultimately comprehends from a text is limited or enhanced by a number of factors, including the linguistic properties of the text, readers’ prior knowledge of the text content, and specific strategies employed during reading (e.g., McNamara, et al., 2014; McNamara & Kintsch, 1996). Multiple strategies such as self-explanation (explain the texts to oneself), bridging (making connections between ideas in a text), and question asking can enhance readers’ levels of comprehension (McNamara, 2007).

To assess these varying levels in comprehension, researchers rely on a variety of language-based responses including think-alouds, free recalls, short answer questions, and essays McCarthy, Kopp, Allen, & McNamara, in press; Wiley & Guerrero, chapter X). While these assessments can be based on human scoring, NLP has provided an additional, multi-dimensional window with which to analyze and understand discourse. A critical assumption of NLP discourse analysis is that one single linguistic feature cannot on its own capture the depth and complexity of discourse processes. Brown and Fraser (1979) emphasize that focusing on specific, isolated linguistic markers can be misleading. As a result, it is critical that we take into account systematic variations involving co-occurrence of sets of markers and analyze texts along multiple dimensions to develop the most robust and complete picture of discourse processes.

Multi-dimensional (MD) or multi-feature Analysis (Biber, 1986, 1991) is a corpus-based approach that assesses linguistic variations within texts. The MD analytic approach also assumes that different types of texts differ linguistically and functionally, and that multiple parameters of variations are generally operative in any discourse domain. As such, it is a quantitative approach synthesizing functional, macroscopic and microscopic approaches. In MD analysis, the underlying “dimensions” of variation are the key to analyzing linguistic co-occurrence. Dimensions are continuous scales of variation and the co-occurrence patterns underlying dimensions are identified quantitatively (Douglas, 1992). MD analyses have been applied to a number of areas such as language disabilities (Gregg, Coleman, Stennett, & Davis, 2002), bilingual creativity (Baker & Eggington, 1999), world Englishes (Xiao, 2009), and argumentative essay writing (Crossley, Allen, & McNamara, 2014). Importantly, MD analyses can provide more fine-grained information to inform theories of discourse processes. A significant breakthrough in the automatic assessment of discourse at multiple levels was the development of Coh-Metrix (Graesser et al., 2004b; McNamara et al., 2014). Coh-Metrix provides information on text properties ranging from the characteristics of the words, the complexity of the sentence structures, to cohesive devices that serve to explicitly convey implicit relations (Britton & Gülgöz, 1991; Graesser, Cai, Louwerse, & Daniel, 2006; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996). There are two major contributions of Coh-Metrix. The first is practical – Coh-Metrix was developed to have linguistic indices “all in one place” rather than having to pass texts through multiple analysis tools. The second was the ability to assess a text’s cohesion. Cohesion is a property of text that reflects the degree to which information is meaningfully organized. Texts that are more cohesive are easier to understand because they include more discourse markers that indicate what information is relevant and where to connect different ideas within the text. Thus, cohesion serves as an index of a text’s difficulty in a more sophisticated and theoretically-motivated way than traditional readability formulas (Duran, McCarthy, Graesser, & McNamara, 2007; Graesser et al., 2004b; McNamara et al., 2014).

Coh-Metrix is used in various research contexts.
Further, assessing the cohesion of a reader’s response indicates the coherence of the readers’ mental model of that concept or text. That is, more cohesive responses suggest that the student has not only identified the relevant ideas, but that the ideas are interconnected. For example, Graesser and colleagues (2007) demonstrated that students with high-knowledge on the topic (physics) appeared no different than their low-knowledge counterparts in terms of indices such as syntax, use of connectives, and causal cohesion. However, high-knowledge students had increased semantic and conceptual overlap suggesting that more prior knowledge is beneficial for aspects of deep comprehension (Graesser, Jeon, Yan, & Cai, 2007; see also Allen, Jacovina, & McNamara, 2016a).

Importantly, Graesser and McNamara (2011) leverage the MD analytic approach in their multilevel framework of discourse comprehension, which includes the (1) surface code, (2) the textbase, (3) the situation model, (4) the type and structure of the text, and (5) the pragmatic communication level (e.g., what the reader thinks the author knows). The purpose of this framework was to explain when comprehension succeeds and breaks down, and to identify the linguistic scaffolds available in text that can hinder or enhance comprehension. This was a significant advance in the field because it demonstrated the use of theoretically-driven NLP to inform theories of discourse processes. Specifically, along with the proposal of this framework, Graesser and McNamara describe how Coh-Metrix can be used to analyze texts along numerous dimensions, which are mapped to the first four of the five levels of their framework.

NLP can also be used to assess other dimensions of discourse. Linguistic Inquiry Word Count (LIWC; Pennebaker, Booth, & Francis, 2007), for example, analyzes a text or response by counting words that reflect different parts of speech as well as different psychological and emotional states. Speech act classifiers can be used to identify metacommunicative expressions (e.g., “Can you repeat that?”), metacognitive expressions (e.g., “I’m confused”), and various categories of questions (Louwerse, Graesser, & Olney, 2002; Olney, Louwerse, Matthews, Marineau, Hite-Mitchell, & Graesser, 2003).

NLP has also been used to analyze student writing. The majority of this work has focused on the development of automated essay scoring (AES) engines that provide scores and feedback on students’ essays (Allen, Jacovina, & McNamara, 2016b; Dikli, 2006; McNamara, Crossley, Roscoe, Allen, & Dai, 2015). Specifically, AES engines have been developed to model essay scores assigned by expert human raters through calculations that rely on the extraction of linguistic features. AES provides the means to automate feedback, and the research conducted in the development of these models has revealed important information about the linguistic properties of high-quality texts. For example, Crossley and McNamara (2011) found that high-quality essays contain more diverse and sophisticated word choices, as well as more complex sentence constructions (see also Crossley & McNamara, 2014).

Researchers have also begun to investigate how and whether individual differences among students can be modeled using the linguistic features calculated with NLP tools (Allen & McNamara, 2015; Allen, Perret, & McNamara, 2016). For instance, Allen and McNamara (2015) used indices related to the lexical properties of students’ essays to successfully model their scores on a vocabulary knowledge assessment. Similarly, Allen, Perret, and McNamara (2016) demonstrated that students’ working memory capacity can be modeled with the linguistic features of their essays, and moreover, that these effects were moderated by inferencing skills. Overall, research suggests that NLP techniques are an extremely powerful source of data that can serve to inform models and research on the writing process. Overall, these and other studies illustrate the power of NLP to inform theories of discourse processing.
**Discussion and Future Directions**

A primary goal of discourse analyses is to model the characteristics of language in order to develop more robust and fine-grained theories, as well as targeted instruction and feedback (Allen & McNamara, 2015). NLP provides a means to automatically analyze virtually any aspect of text that one can imagine (Crossley et al., 2014; Graesser et al., 2004b; McNamara et al., 2014). Ultimately, we can leverage this information for a variety of applied educational purposes, such as the delivery of automated feedback, personalized learning, and scaffolded support. Accordingly, researchers are more and more frequently relying on large and complex data sets, which require more sophisticated analysis techniques. We have argued that NLP is a critical piece of this puzzle, which should be leveraged to drive theoretical progress and more personalized learning environments (McNamara et al., 2017).

The development of a robust theory of discourse processing relies on the consideration of multiple approaches to data analysis. Discourse processing and production are highly complex processes that emerge from interactions among properties of texts, individual differences, and the task in which the learner is situated (McNamara et al., 1996). These elements are all then taken into account within the backdrop of a larger sociocultural context (Snow, 2002). Thus, the reliance on any single source or type of data to understand these processes is short sighted, particularly when so many sources of information are available to researchers. NLP simply provides one source of rich, multidimensional data that can be used to inform our understanding of discourse. However, the development of a more complete understanding will ultimately require an integration of multiple sources of data.

A further aspect of discourse processing research that is important to discuss is the dynamic nature of discourse processes. Many theoretical models of the comprehension or production of discourse implicitly or explicitly assume that these processes are dynamic and involve interactions across many different sources of information. As such, the relative role of these sources is expected to change over the time course of comprehension. While this a general assumption of most comprehension models, there is surprisingly little research on the dynamics of comprehension.

Some researchers have focused on reading times and eye tracking, which have provided some insight into these dynamic processes of discourse comprehension. However, few studies have synchronously examined reading as the features of the text change from sentence to sentence. That is, there has been little work exploring reading dynamics as they unfold over time. Further, there has been significantly less research conducted on students’ online writing processes compared to their reading processes. Most of this research has focused on students’ finished essays and not their moment-by-moment writing process (cf. van den Bergh & Rijaarsdam, 2007; Van Waes, Leijten, Lindgren, & Wengelin, 2015). We have argued that further investigations of these dynamic processes can be used to model important characteristics of students’ and the language they produce that are impossible to assess with product measures alone (McNamara & Allen, 2017).

To this end, tools have been developed recently to facilitate recording individual keystrokes pressed by individuals during discourse production tasks (Allen, Mills, Jacovina, Crossley, D’Mello, & McNamara, 2016c; Bixler & D’Mello, 2013; Van Waes et al., 2015). These tools have allowed researchers to glean significant insights about the dynamics of discourse production, such as differences between L1 and L2 writers (Sullivan & Lindgren, 2006) and the affective states observed during writing (Allen et al., 2016c; Bixler & D’Mello, 2013). Recent research is working to integrate these keystroke analyses with NLP techniques to
glean information about how the linguistic properties of texts unfold dynamically throughout discourse tasks. Ultimately, these types of analyses will help to fill in gaps in our understanding of cognition, to provide a more complete picture of students’ learning processes. Multidimensional text and discourse analyses can be used to identify areas of potential difficulties for students’ discourse processes (e.g., reading, writing, communication), model students’ strengths and weaknesses, and drive more adaptive instruction and feedback in educational technologies.

**Conclusion**

Discourse is of central importance for our ability to form thoughts, express ourselves to others, and engage with the world around us. Not surprisingly, then, analyses of language can provide important theoretical insights related to how individuals comprehend complex information and subsequently communicate those ideas to others. Analyses of text and discourse serve as one method for understanding the complex processes associated with these language production and comprehension processes. These analyses rely on systematic measures of the characteristics of spoken and written discourse, which can then be related to important behaviors, psychological constructs, and social interactions.

We encourage researchers to utilize NLP techniques as assessments to inform their theoretical perspectives and educational practice. Although NLP is not the only source of information that is important, it provides an important lens through which we can better understand discourse processes. Ultimately, our aim is for these NLP techniques to be integrated with other forms of assessment to provide a more thorough understanding of deep learning and the dynamics of discourse processing. This approach will be essential to integrating research from theoretical and applied perspectives, which will have the strongest impact on research and educational practice.

**References**


