Chapter 8: Natural Language Processing and Learning Analytics

Danielle S. McNamara[1], Laura K. Allen[1], Scott A. Crossley[2], Mihai Dascalu[3], Cecile A. Perret[4]

[1]Psychology Department, Arizona State University, USA
[2]Applied Linguistics and ESL Department, Georgia State University, USA
[3]Computer Science Department, University Politehnica of Bucharest, Romania
[4]Institute for the Science of Teaching and Learning, Arizona State University, USA

ABSTRACT

Language is of central importance to the field of education because it is a conduit for communicating and understanding information. Therefore, researchers in the field of learning analytics can benefit from methods developed to analyze language both accurately and efficiently. Natural language processing (NLP) techniques can provide such an avenue. NLP techniques are used to provide computational analyses of different aspects of language as they relate to particular tasks. In this chapter, the authors discuss multiple, available NLP tools that can be harnessed to understand discourse, as well as some applications of these tools for education. A primary focus of these tools is the automated interpretation of human language input in order to drive interactions between humans and computers, or human–computer interaction. Thus, the tools measure a variety of linguistic features important for understanding text, including coherence, syntactic complexity, lexical diversity, and semantic similarity. The authors conclude the chapter with a discussion of computer-based learning environments that have employed NLP tools (i.e., ITS, MOOCs, and CSCL) and how such tools can be employed in future research.

Keywords: Natural language processing (NLP), learning analytics, language, computational linguistics, linguistic features, automated writing evaluation, intelligent tutoring systems, CSCL, MOOCs

Language is one means to externalize our thoughts. It allows us to express ourselves to others, to manipulate our world, and to label objects in the environment. Language allows us to internally construct and reconstruct our thoughts; it can represent our thoughts, and allow us to transform them. It allows us to construct and shape social experiences. Language provides a conduit to understanding and interacting with the world.

Language is omnipresent in our lives – in our thoughts, our communications, what we read and write, and our interactions with others. Language is equally central to education. Our goal as instructors is to communicate information to students so that they have the opportunity to learn new information, to absorb it, and to integrate it. Students are tasked with understanding language used to communicate information, and then to connect that information with what they already know – to construct their understanding as individuals, in groups, and in coordination with each other and instructors.

Language plays important roles in our lives, and in education, and thus, it is important to recognize and understand those roles and outcomes. Text and discourse analysis provides one means to understand complex processes associated with the use of language. Discourse analysts systematically examine structures and patterns within written text and spoken discourse and their relations to behaviours, psychological processes, cognition, and social interactions. Indeed, text and discourse analysis has provided a wealth of information about language.

Traditionally, however, discourse analysis is laborious. First, for example, the meaningful units of language are identified and segmented (e.g., clauses, utterances) and then experts code those units (i.e., with respect to the particular analysis). The potential relations between the nature of those language units and outcomes are then assessed. In a world of *big data*, where there are thousands of utterances and exchanges between individuals, hand-coding language is nearly impossible. Large corpora of data open the doors to understanding language on a wider and even more meaningful scale, but traditional approaches to discourse analysis are simply not feasible. One solution derives from *natural language processing* (NLP).

Natural Language Processing Tools

Computational linguistics is a discipline that focuses on the development of computational models of language. NLP tools and techniques are often guided by theories, models, and algorithms developed in the field of computational linguistics, but the primary purpose of NLP tools is the automated interpretation of human language input. Such an endeavor calls upon interdisciplinary perspectives integrating disciplines such as linguistics, computer science, psychology, and education. While NLP has a history dating back to Turing (1950), the majority of current NLP algorithms have been developed using a combination of NLP tools and data mining. A clear distinction must be made from the beginning between the NLP software often used by computer or data scientists and the tools presented in this chapter. A large majority of NLP research has focused on surface-level text processing (e.g., machine translation), and the available tools consequently emphasize the central role of accurate word- and sentence-level text processing. Our aim in this chapter is specifically to focus on NLP within the context of *learning analytics*. Thus, we focus on tools developed to calculate linguistic indices that move beyond these surface-level tasks and provide information that may be more important within educational contexts. Notably, we describe a subset of NLP techniques that provide information about multiple levels of text. These tools begin from the words in the discourse, extract specific word features, and then go beyond the lexicon by considering semantics, as well as discourse structure. Our goal is to provide examples of a few common techniques, rather than an overview of all available methods. We group these methods into those that focus on the words directly as the units of analysis, and those that focus on features of the words.

The Words

One approach to NLP is to analyze the words used in the language directly. For example, calculating the incidence of specific types of words within a text can reveal a good deal about the nature and purpose of the language used in various contexts. This is often referred to as a "bag-of-words" approach. One tool that employs this approach is the *Linguistic Inquiry Word Count* (LIWC) system developed by Pennebaker and colleagues (Pennebaker, Booth, & Francis, 2007; Pennebaker, Boyd, Jordan, & Blackburn, 2015; see http://liwc.wpengine.com/). The 2007 version of LIWC provides roughly 80 word categories, but also groups these word categories into broader dimensions. Examples of the broader dimensions are linguistic forms (e.g., pronouns, words in past tense, negations), social processes, affective processes, and cognitive processes. For example, cognitive processes include subcategories such as insight (e.g., think, know, consider), causation (e.g., because, effect, hence), and certainty (e.g., always, never). LIWC counts the number of words that belong to each word category and provides a proportion score that divides the number of words in the category by the total number of words in the text.

A similar approach is to identify n-grams, such as groups of characters or words, where *n* refers to the number of *grams* included in the group (e.g., bi-grams refer to groups of *two* words). N-gram analyses calculate probability distributions of word sequences in text and can provide information about the words common to a group of texts, or distinctive for a specific text or sets of texts (e.g., Jarvis et al., 2012). Several advantages of n-gram analyses include their simplicity and the potential for providing information about the specific content of a text, the linguistic and syntactic features of a text, and relationships between those features (Crossley & Louwerse, 2007).

The Features of the Words

Calculating the occurrence of words and groups of words considers the explicit content of the text. An alternative approach involves the calculation of the features of the words and sentences in a text. One such technique is to derive the latent meaning behind the words (McNamara, 2011). There are numerous algorithms for doing so, but the most well known and perhaps the first was Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007; see lsa.colorado.edu). LSA emerged in the mid-1990s, providing a means to extract semantic meaning from large bodies of text, and to compare large and small text samples for semantic similarities. Such an approach provided a unique potential to revolutionize NLP. LSA is a mathematical, statistical technique that uses *singular value decomposition* to compress (i.e., factorize) a matrix representing the occurrence of words across a large set of documents. A principal assumption driving LSA is that the meanings of words are captured by the company they keep. For example, the word "data" will be highly associated with words of the same functional context, such as "computations", "mining", "computer", and "mathematics". These words do not mean the same thing as *data*. Rather, these words are related to *data* because they typically occur in similar contexts. By affording the computation of the semantic similarities between words, sentences, and paragraphs, LSA opened the doors to the simulation of meaning in text (McNamara, 2011). LSA can be considered the first word-based approach to successfully address the question of relevance (i.e., the degree to which a text is relevant to another text or to a core concept), a problem for which simple measures of word overlap are not sufficient. While there are multiple approaches that have gone beyond LSA (see McNamara, 2011, for an overview), LSA remains a common approach used across multiple contexts to model word meaning and to provide insights in terms of semantics and text cohesion (e.g., Landauer et al., 2007; McNamara, Graesser, McCarthy, & Cai, 2014).

One obvious feature of language is the meaning, but many other features can be derived from linguistic analyses, such as the parts of speech (e.g., verb, noun), syntax, psychological aspects (e.g., concreteness, meaningfulness), and the relations between ideas in the text (e.g., cohesion). Coh-Metrix is an example of an automated language analysis tool, first launched in 2003, that uses multiple sources of information about language to extract linguistic, psychological, and semantic features of text (McNamara et al., 2014; cohmetrix.com). Coh-Metrix adapts and integrates information about the English language from a variety of sources including *LSA*, the *MRC Psycholinguistic Database*, *WordNet*, and word frequency indices such as *CELEX*, as well as syntactic parsers. For example, the MRC Psycholinguistic Database provides psycholinguistic information about words (Wilson, 1988) and WordNet provides linguistic and semantic features of words, as well as semantic relations between words (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Coh-Metrix also calculates linguistic indices related to various aspects of language through simple features of text quality, such as word frequency and sentence length, as well as more complex features such as coherence and syntactic complexity, in order to produce a multi-dimensional analysis of written or spoken text (McNamara, Ozuru, Graesser, & Louwerse, 2006). Coh-Metrix can provide a relatively simple characterization of a text through descriptive indices (i.e., length of words, sentences, paragraphs). In addition, it offers various complex indices that describe a text's quality and readability. Among these indices are the five Coh-Metrix Text Easability Components, including narrativity, referential cohesion, syntactic simplicity, word concreteness, and deep cohesion (Graesser, McNamara, & Kulikowich, 2011; Jackson, Allen, & McNamara, 2016; see coh-metrix.commoncoretera.com).

Coh-Metrix has had a large impact on our understanding of language and discourse by making automated language analysis publicly available. While Coh-Metrix provides multiple measures of language, the primary, unique focus of Coh-Metrix has been on providing measures of cohesion in text. Cohesion is the overlap in features, words, and meaning between sentences (i.e., local cohesion) and larger sections of the text such as paragraphs (i.e., global cohesion) and the text overall (e.g., lexical diversity). While extremely useful, Coh-Metrix has had several shortcomings regarding facile and broad measurement of cohesion indices. First, it does not allow for the batch processing of text, and it is not housed on a user's hard drive (and thus it depends on an internet connection and an external server). Second, Coh-Metrix cohesion indices generally focus on local and overall text cohesion (i.e., average sentence overlap, lexical diversity), rather than global cohesion (e.g., semantic overlap between various sections of a text). Hence, the Tool for the Automatic Analysis of Text Cohesion (TAACO) and the Simple Natural Language Processing Tool (SiNLP) were developed to address these gaps (Crossley, Allen, Kyle, & McNamara, 2014; Crossley, Kyle, & McNamara, in press; http://www.kristopherkyle.com/taaco.html). TAACO is locally installed (as compared to an internet interface), allows for batch processing of text files, and includes over 150 indices related to local, global, and overall text cohesion. Similarly, SiNLP is locally installed and allows for batch text processing. However, SiNLP differs from TAACO in that it takes the "bag-of-words" approach to calculate information about multiple aspects of texts. Additionally, the tool is flexible and allows researchers to add their own categories of words to inform additional analyses.

Another example of a freely available NLP tool is the Tool for the Automatic Analysis of LExical Sophistication (TAALES; Kyle & Crossley, 2015; http://www.kristopherkyle.com/taales.html). TAALES focuses on providing extensive information about the level of lexical sophistication present in a text. This type of analysis is important because it provides information on the lexical demands of a text, as well as potential information related to the lexical knowledge of the author of the text (Kyle & Crossley, 2015). TAALES calculates over 130 classic and newly developed lexical indices to assess the breadth and depth of lexical knowledge used in a text. This tool is fast, reliable, and freely available for download. The measures for TAALES include word frequency, word and word family range, n-grams, academic lists, and word information indices that consider psycholinguistic components (Kyle & Crossley, 2015). These indices collectively provide extensive information on the complexity of word choices in text.

Dascalu, McNamara, Crossley, and Trausan-Matu (2016) also introduced *Age of Exposure* (*AoE*), a computational model to estimate word complexity in which the learning rate of individual words is calculated as a function of a learner's experience with language. In contrast to Pearson's calculation of word maturity (Landauer, Kireyev, & Panaccione, 2011), AoE is a reproducible and scalable model that simulates word learning in terms of potential associations that can be created with it across time or, more specifically, across incremental latent Dirichlet allocation (Blei, Ng, & Jordan, 2003) topic models. AoE indices yield strong associations (exceeding the reported performance of word maturity) with estimates of word frequency and entropy, as well as human ratings of age of acquisition and lexical response latencies.

Natural Language Processing and Learning Algorithms

NLP can be used to describe multiple facets of language from simple descriptive statistics, such as the number of words, n-grams, and paragraphs, to the features of words, sentences, and text (Crossley, Allen, Kyle, & McNamara, 2014). As depicted in Figure 8.1, multiple characteristics of language can be gleaned from the words (including n-grams and bags of words) and captured using both techniques for analyzing observable features (e.g., word frequencies, word-document distributions) and latent meaning from the text (McNamara, 2011). Information is provided by the features of the words, the sentences, and the text as a whole. This information can be analyzed using machine learning techniques such as linear regression, discriminant function classifiers, Naïve-Bayes classifiers, support vector machines, logistic regression classifiers, and decision tree classifiers. When these techniques are used to predict learning outcomes, algorithms can be derived that can then be used within educational technologies or applications. We discuss a number of these applications in the following sections.

Writing Assessment

The most common example of the use of NLP in the realm of education is for the development of automated essay scoring (AES) algorithms (Allen, Jacovina, & McNamara, 2016; Dikli, 2006; Weigle, 2013; Xi, 2010). AES systems assess essays using a variety of approaches. For example, the Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2003) primarily relies on LSA to assess the similarity of an essay to benchmark essays. By contrast, systems such as the e-rater developed at Educational Testing Service (Burstein, Chodorow, & Leacock, 2004), IntelliMetric Essay Scoring System developed by Vantage Learning (Rudner, Garcia, & Welch, 2006), and the Writing Pal (McNamara, Crossley, & Roscoe, 2013) rely on combinations of NLP techniques and artificial intelligence. AES systems process writing samples such as essays, and assess the degree to which the writer has met the demands of the task by assessing the quality of essays and their accuracy relative to the content. AES technologies are highly successful, reporting levels of accuracy generally as accurate as expert human raters (Attali & Burstein, 2006; Shermis, Burstein, Higgins, & Zechner, 2010; Valenti, Neri, & Cucchiarelli, 2003; Crossley, Kyle, & McNamara, 2015).

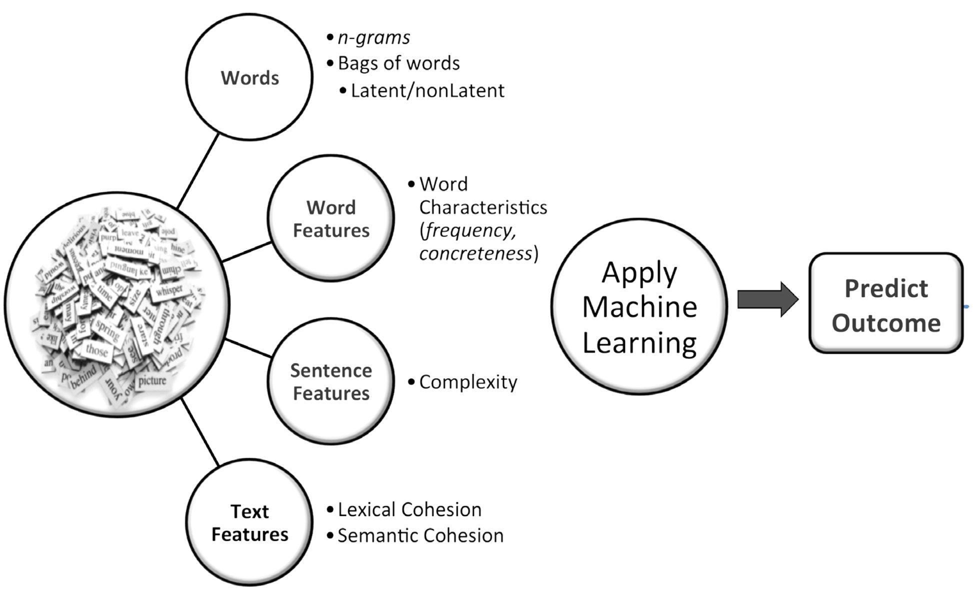

Figure 8.1. Developing algorithms using NLP requires machine-learning techniques applied to various sources of information on the text, including information from the words, sentences, and the entire text.

Tutoring Systems

Another use of NLP has been in the context of automated, intelligent tutoring technologies. NLP has been incorporated into a number of intelligent tutoring systems (ITSs), particularly those that interact with the student via dialogue (e.g., AutoTutor: Graesser et al., 2004) and 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). When a student enters natural language into a system and expects useful feedback or a reasonable response, NLP can be used to interpret that input and provide appropriate feedback (McNamara et al., 2013). For tutoring systems that accept natural language as input (e.g., verbal explanations of text, problems, or scientific processes), student responses can be open-ended and potentially ambiguous. For example, the student might be asked which phase of cell mitosis involves the lengthening of the microtubules. This type of question (e.g., *what* or *when* questions) can be answered using short answers or multiple-choice responses, requiring little to no NLP. By contrast, a question to describe the process of Anaphase would elicit answers likely to differ widely between students. Thus, automatically detecting the accuracy and quality of the student's answer requires the use of NLP.

Why not just use multiple-choice? Many tutorial systems do just that. However, students are more likely to construct a deep understanding of a construct or phenomenon by answering *how* and *why* questions (e.g., Johnson-Glenberg, 2007; McKeown, Beck, & Blake, 2009; Wong, 1985). Moreover, students' answers to these types of questions are more likely to unveil the depth of their understanding (Graesser & Person, 1994; Graesser, McNamara, & VanLehn, 2005; McNamara & Kintsch, 1996). AutoTutor is an ITS that focuses on providing instruction on challenging topics (e.g., physics, biology, computer programing) by prompting students to answer deep level *how* and *why* questions. AutoTutor engages the student via an animated agent in a dialogue that moves the student toward constructing the correct answers. It does so by using a variety of dialogue moves, such as hints, prompts, assertions, corrections, and answers to student questions. These moves are driven by a combination of NLP techniques. For example, AutoTutor uses frozen expressions to detect phrases that students are likely to produce in certain situations (e.g., *I don't know*; *I don't understand*) as well as key parts of the correct answer. AutoTutor also uses LSA to detect the similarity between the answer provided by the student and the ideal answer. The combination of frozen expressions, regular expressions or patterns, inverse-frequency weighted word overlaps between student verbal responses and expectations, and LSA, allows AutoTutor to simulate the understanding of the student's answer, and in turn, this simulated understanding drives an appropriate response to the student (Graesser, in press).

iSTART (Interactive Strategy Training for Active Reading and Thinking) is another ITS that relies on a combination of NLP techniques to respond to open-ended responses. iSTART was among the first automated systems to address the paraphrase problem in student's self-explanations, a difficult challenge in the both NLP and computational linguistics literature. iSTART enhances students' comprehension of challenging science texts by providing instruction and practice to use self-explanation (i.e., the process of explaining text to oneself) in combination with comprehension strategies such as generating bridging and elaborative inferences. During the practice phase of iSTART instruction, students generate self-explanations for challenging texts. Students' self-explanations in iSTART are scored using an algorithm that combines information from the words in the self-explanation and the text, using a combination of observable and latent semantic information about the words (McNamara, Boonthum, Levinstein, & Millis, 2007). The algorithm automatically assigns a score between 0 and 3 to each self-explanation. Higher scores are assigned to self-explanations that include information related to the text content (both the target sentence and previously read sentences), whereas lower scores are assigned to unrelated or short responses. The scoring algorithm is designed to reflect the extent to which students construct connections between the target sentence, prior text content, and world knowledge. The system successfully matches human scores of the explanations across a wide variety of texts (Jackson, Guess, & McNamara, 2010; McNamara et al., 2007).

Computer Supported Collaborative Learning (CSCL)

NLP techniques have also been applied to discourse generated in collaborative learning environments, and in particular Computer Supported Collaborative Learning (CSCL) systems (Stahl, 2006). A subset of these systems model CSCL conversations based on *dialogism*, a concept introduced by Bakhtin (1981) that later on emerged as a paradigm for CSCL (Koschmann, 1999). The most representative approaches are Dong's (2005) design of team communication, Polyphony (Trausan-Matu, Rebedea, Dragan, & Alexandru, 2007), the Knowledge Space Visualizer (Teplovs, 2008), and ReaderBench (Dascalu, Stavarache et al., 2015; Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). ReaderBench leverages the power of text mining techniques, advanced NLP, and social network analysis to achieve multiple objectives related to language comprehension as well as collaborative learning (Dascalu, 2014). ReaderBench models participation and collaboration from a Cohesion Network Analysis perspective in which the information communicated among participants is computed via semantic textual cohesion (Dascalu, Trausan-Matu, Dessus, & McNamara, 2015a). Moreover, ReaderBench has introduced an automated dialogic model for assessing collaboration based on the polyphonic model of discourse (Trausan-Matu, Stahl, & Sarmiento, 2007). Grounded in theories of dialogism (Bakhtin, 1981), the system automatically identifies *voices* or participant's points of view as semantic chains that include tightly cohesive or semantically related concepts spanning throughout the entire conversation (Dascalu, Trausan-Matu, Dessus, & McNamara, 2015b). Thus, collaboration emerges from the inter-animation of different participant voices, which is computationally captured in the co-occurrence patterns used to highlight the exchange of ideas between different participants.

Massive Open Online Courses (MOOCs)

Another use of NLP has been in the context of online courses, particularly massive open online courses (MOOCs). MOOCs use online platforms to make courses available to thousands of students without cost to the student. MOOCs are lauded for their potential to increase accessibility to distance and lifelong learners (Koller, Ng, Do, & Chen, 2013). These platforms can provide a tremendous amount of data via click-stream logs, assignments, course performance, as well as language generated by the students within discussion forums and emails. These data can be mined to examine student attitudes, completion, and learning (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014; Wen, Yang, & Rosé, 2014a, 2014b).

The most common NLP approach to analyzing student language in MOOCS has been through tools that analyze emotions. Sentiment analysis examines language for positive or negative emotion words or words related to motivation, agreement, cognitive mechanisms, or engagement (Chaturvedi, Goldwasser, & Daumé, 2014; Elouazizi, 2014; Moon, Potdar, & Martin, 2014; Wen et al., 2014a, 2014b). For example, Moon et al. (2014) used emotion terms and semantic similarity among participants to identify student leaders. Elouazizi (2014) showed that linguistic indices related to point of view (e.g., *think*, *believe*, *presumably*, *probably*) were correlated with low levels of engagement in the course. Wen and colleagues (2014a, 2014b) found that students' use of personal pronouns and words related to motivation within discussion forums was predictive of a lower risk of dropping out of the course.

Similarly, Crossley, McNamara et al. (2015) used multiple levels of linguistic features to examine students' language in a MOOC discussion forum within a course covering the topic of educational data mining (Baker et al., in press). Crossley, McNamara et al. (2015) successfully predicted the completion rates (with an accuracy of 70%) of 320 students who participated within the MOOC discussion forums (i.e., posted > 49 words). Students who were more likely to receive a certificate of completion in the course generally used more sophisticated language. For example, their posts were more concise and cohesive, used less frequent and specific words, and had greater overall writing quality. Interestingly, indices related to affect were not predictive of completion rates.

Collectively, this research provides promising evidence that NLP can be a powerful predictor of success in the context of MOOCs. Communication between the instructor and the students as well as between the students is crucial, particularly for distance courses. Further, this communication can then be used as forms of assessment of student performance. Therefore, it seems apparent that MOOCs should include discussion forums in order to better monitor student participation and potential success. The language that students use can also be utilized to identify students who are less likely to complete the course, and target those students for interventions such as sending emails, suggesting content, or recommending tutoring. Automating language understanding, and thereby providing information about the language and social interactions within these courses, will help to enhance both learning and engagement in MOOCs.

The Power of NLP

NLP is extremely powerful, primarily because language is ubiquitous and also because tools to analyze language automatically provide indices related to virtually any aspect of language (Crossley, 2013). NLP can detect the specific words used, groups of words, and the strength of the relations between words and between larger bodies of text. It can also detect the features of the text, such as the frequency, concreteness, or meaningfulness of the words, the complexity of the sentences, and various aspects of the text such as cohesion and genre. The words and their features serve as proxies to various constructs. For example, the frequency of the words in a text serves as a proxy to estimate the knowledge that might be required to understand the text. The cohesion of a text affords an estimate of the knowledge necessary to fill in the gaps in a text.

NLP has been used to identify a wide variety of other constructs. For example, Crossley and McNamara (2012) demonstrated that the linguistic features of second language (L2) writers' essays could predict the native language of those writers. Varner, Roscoe, and McNamara (2013) used indices provided by both Coh-Metrix and LIWC to examine differences in students' and teachers' ratings of essay quality. Louwerse, McCarthy, McNamara, and Graesser (2004) used NLP techniques to identify differences between spoken and written samples of English. McCarthy, Briner, Rus, and McNamara (2007) showed that Coh-Metrix could differentiate sections in typical science texts, such as *introductions*, *methods*, *results*, and *discussions*. Additionally, Crossley, Louwerse, McCarthy, and McNamara's (2007) investigations of second language learner texts, revealed a wide variety of structural and lexical differences between texts that were *adopted* (or authentic) versus *adapted* (or simplified) for second language learning purposes. Finally, NLP has also been used to detect deception. Duran, Hall, McCarthy, and McNamara (2010) examined the extent to which features of language discriminated between conversational dialogues in which a person was being deceptive and those in which the person was being truthful.

It is important to note that there are potential drawbacks to using NLP. For example, certain NLP techniques rely on simplified representations of dialogue that use word counts or "bag-of-words" approaches. The most notable and widely used NLP word representations, including LSA vector-spaces, latent Dirichlet allocation topic distributions (LDA; Blei et al., 2003), and word2vec models based on neural networks (Mikolov, Chen, Corrado, & Dean, 2013), are all subject to the "bag-of-words" assumption in which word order is disregarded. In addition, many NLP analyses ignore context, such as the intentions or pragmatic aspects of the speaker. Similarly, NLP analyses are often limited to particular corpora and situations, and fail to generalize to other contexts. Even with these (and other) caveats, NLP is extremely powerful. Because of the vast sources of information now available from NLP tools, and because the language we use can be an extension or externalization that represents thoughts and intentions, NLP can provide information about the individuals, their abilities, their emotions, their intentions, and social interactions. In the context of learning analytics, it is a means toward the automated understanding of learning processes and the learner.

The Big Picture

NLP provides techniques that automate the analysis of language, which allows researchers to establish a better understanding of language and of the roles that language potentially plays in various aspects of our lives. NLP informs feedback systems within tutoring systems that prompt the student to generate language within answers to questions, explanations, and essays. NLP provides a means of simulating intelligence within language-based tutoring systems. NLP is also informative in the context of online discussion forums. It provides information on student attitudes, motivation, and the quality of the language, which in turn is predictive of students' likelihood of performing well or completing the course.

One goal of learning analytics is to model the characteristics and skills of students in order to provide more effective instruction (Allen & McNamara, 2015). Specifically, we can use this data for various purposes: provide automated feedback on performance, intervene during learning, provide scaffolding or support, recommend tutoring, personalize learning, and so on, with the assumption that information gleaned from analytics will ultimately enhance learning. For this purpose, researchers are increasingly turning to large, complex data sources (i.e., big data) and using various combinations of data types and analytic techniques. NLP is crucial to this endeavour because the proposed techniques help to improve student learning through the prediction and assessment of comprehension

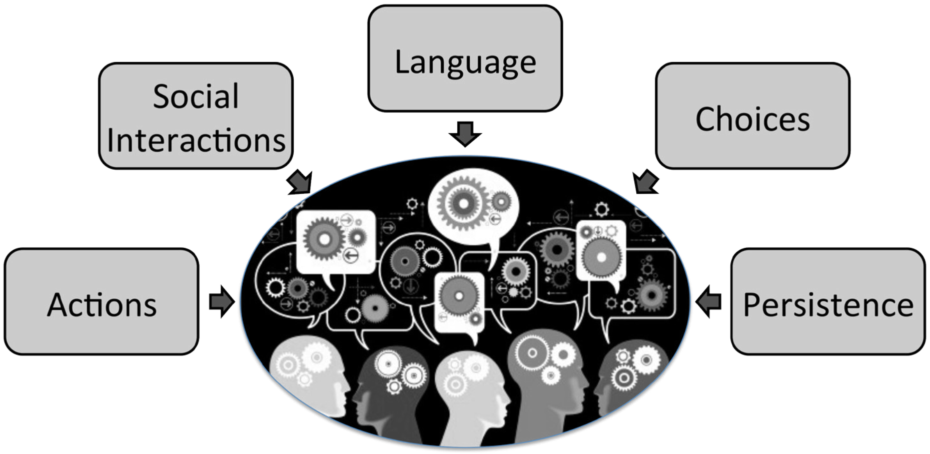

Figure 8.2. Predicting educational outcomes will require the integration of multiple sources of data.

across a variety of contexts. However, NLP is only one piece of the puzzle.

As depicted in Figure 8.2, developing a complete and highly predictive understanding of student outcomes requires multiple sources of information and a variety of approaches to data analysis. Learning is a complex process with multiple layers and multiple time scales. Relying on any single source or type of data to understand the learning process is myopic, particularly when so many automated sources of information are currently available. NLP is simply one source of data increasingly recognized as an integral piece of the big picture that ultimately we seek. Developing a complete understanding of learning will require an integration of multiple sources of data.

ACKNOWLEDGMENTS

The writing of this chapter was supported in part by the Institute for Educational Sciences (IES R305A120707, R305A130124), the National Science Foundation (NSF DRL-1319645, DRL-1418352, DRL-1418378, DRL-1417997), and the Office for Naval Research (ONR N000141410343). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES, NSF, or ONR. We are also grateful to the many students, postdoctoral researchers, and faculty who have contributed to our research on NLP over the years.

REFERENCES

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*, 2nd ed. (pp. 316–329). New York: The Guilford Press.

Allen, L. K., & McNamara, D. S. (2015). You are your words: Modeling students' vocabulary knowledge with natural language processing. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM 2015) 26–29 June 2015, Madrid, Spain (pp. 258–265). International Educational Data Mining Society.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(2). doi:10.1002/j.2333-8504.2004.tb01972.x

Baker, R., Wang, E., Paquette, L., Aleven, V., Popescu, O., Sewall, J., Rose, C., Tomar, G., Ferschke, O., Hollands, F., Zhang, J., Cennamo, M., Ogden, S., Condit, T., Diaz, J., Crossley, S., McNamara, D., Comer, D., Lynch, C., Brown, R., Barnes, T., & Bergner, Y. (in press). A MOOC on educational data mining. In. S. ElAtia, O. Zaïane, & D. Ipperciel (Eds.). *Data Mining and Learning Analytics in Educational Research.* Wiley & Blackwell.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin, TX: University of Texas Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5), 993–1022.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *Ai Magazine*, 25(3), 27.

Chaturvedi, S., Goldwasser, D., & Daumé III, H. (2014). Predicting instructor's intervention in MOOC forums. In D. Marcu, K. Toutanova, & H. W. Baidu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1501–1511). Baltimore, MD.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, 46(2), 256–271.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool (SiNLP). *Discourse Processes*, 51, 511–534.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment*, 8(1). http://www.journalofwritingassessment.org/article.php?article=80

Crossley, S. A. Kyle, K., & McNamara, D. S. (in press). Tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*.

Crossley, S. A., & Louwerse, M. (2007). Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics*, 12(4), 453–478.

Crossley, S. A., Louwerse, M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91, 15–30.

Crossley, S. A., & McNamara, D. S. (2012). Interlanguage talk: A computational analysis of non-native speakers' lexical production and exposure. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 425–437). Hershey, PA: IGI Global.

Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM 2015), 26–29 June 2015, Madrid, Spain (pp. 388–391). International Educational Data Mining Society.

Dascalu, M. (2014). Analyzing discourse and text complexity for learning and collaborating. *Studies in Computational Intelligence* (Vol. 534). Switzerland: Springer.

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015). ReaderBench: The learning companion. In A. Mitrovic, F. Verdejo, C. Conati, & N. Heffernan (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (AIED '15), 22–26 June 2015, Madrid, Spain (pp. 915–916). Springer.

Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015a). Discourse cohesion: A signature of collaboration. In P. Blikstein, A. Merceron, & G. Siemens (Eds.), *Proceedings of the 5th International Learning Analytics & Knowledge Conference* (LAK '15), 16–20 March, Poughkeepsie, NY, USA (pp. 350–354). New York: ACM.

Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015b). Dialogism: A framework for CSCL and a signature of collaboration. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Proceedings of the 11th International Conference on Computer-Supported Collaborative Learning* (CSCL 2015), 7–11 June 2015, Gothenburg, Sweden (pp. 86–93). International Society of the Learning Sciences.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., & Dessus, P. (2015). ReaderBench: Automated evaluation of collaboration based on cohesion and dialogism. *International Journal of Computer-Supported Collaborative Learning*, 10(4), 395–423.

Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2016). Age of exposure: A model of word learning. *Proceedings of the 30th Conference on Artificial Intelligence* (AAAI-16), 12–17 February 2016, Phoenix, Arizona, USA (pp. 2928–2934). Palo Alto, CA: AAAI Press.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1). http://files.eric.ed.gov/fulltext/EJ843855.pdf

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, 26(5), 445–461.

Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, 31(3), 439–462.

Elouazizi, N. (2014). Point-of-view mining and cognitive presence in MOOCs: A (computational) linguistics perspective. *EMNLP 2014*, 32. http://www.aclweb.org/anthology/W14-4105

Graesser, A. C. (in press). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education.*

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 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, 223–234.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point & Query, AutoTutor, and iSTART. *Educational Psychologist*, 40, 225–234.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104–137.

Jackson, G. T., Allen, L. K., & McNamara, D. S. (2016). Common Core TERA: Text Ease and Readability Assessor. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction*. New York: Taylor & Francis, Routledge.

Jackson, G. T., Guess, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science*, 2, 127–137.

Jarvis, S., Bestgen, Y., Crossley, S. A., Granger, S., Paquot, M., Thewissen, J., & McNamara, D. S. (2012). The comparative and combined contributions of n-grams, Coh-Metrix indices and error types in the L1 classification of learner texts. In S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 154–177). Bristol, UK: Multilingual Matters.

Johnson-Glenberg, M. C. (2007). Web-based reading comprehension instruction: Three studies of 3D-readers. In D. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 293–324). Mahwah, NJ: Lawrence Erlbaum Publishers.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause Review*, 48(3), 62–63.

Koschmann, T. (1999). Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. In C. M. Hoadley & J. Roschelle (Eds.), *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning* (CSCL '99), 12–15 December 1999, Palo Alto, California (pp. 308–313). International Society of the Learning Sciences.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading*, 15(1), 92–108.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (CogSci 2004), 4–7 August 2004, Chicago, IL, USA (pp. 843–848). Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, P. M., Briner, S. W., Rus, V., & McNamara, D. S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107–122). London: Springer-Verlag UK.

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, 44, 218–253.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, 2, 1–15.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227–241). Mahwah, NJ: Lawrence Erlbaum Associates.

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45, 499–515.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247–288.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers*, 36, 222–233.

McNamara, D. S., Ozuru, Y., Graesser, A. C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (CogSci 2006), 26–29 July 2006, Vancouver, British Columbia, Canada (pp. 573–578). Austin, TX: Cognitive Science Society.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S. A, Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). Hershey, PA: IGI Global.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representation in vector space. In *Workshop at ICLR*. Scottsdale, AZ. https://arxiv.org/abs/1301.3781

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235–244.

Moon, S., Potdar, S., & Martin, L. (2014). Identifying student leaders from MOOC discussion forums through language influence. *EMNLP 2014*, 15. http://www.aclweb.org/anthology/W14-4103

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX: liwc.net.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. UT Faculty/Researcher Works. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?sequence=3

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105, 1010–1025.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric' essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4). https://ejournals.bc.edu/ojs/index.php/jtla/article/download/1651/1493

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, 57(4), 58–65.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, 4, 20–26.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge* (pp. 451–473). Cambridge, MA: MIT Press.

Teplovs, C. (2008). The knowledge space visualizer: A tool for visualizing online discourse. In G. Kanselaar, V. Jonker, P. A. Kirschner, & F. Prins (Eds.), *Proceedings of the International Society of the Learning Sciences 2008: Cre8 a learning world.* Utrecht, NL: International Society of the Learning Sciences. http://chris.ikit.org/ksv2.pdf

Trausan-Matu, S., Rebedea, T., Dragan, A., & Alexandru, C. (2007). Visualisation of learners' contributions in chat conversations. In J. Fong & F. L. Wang (Eds.), *Blended learning* (pp. 217–226). Singapore: Pearson/Prentice Hall.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *Indiana University Press, E-service Journal*, 6(1), 58–74.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433–460. http://www.loebner.net/Prizef/TuringArticle.html

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1), 319–330.

Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research*, 5, 35–59.

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36–54). London: Routledge.

Wen, M., Yang, D., & Rosé, C. P. (2014a). Linguistic reflections of student engagement in massive open online courses. In E. Adar & P. Resnick (Eds.), *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (ICWSM '14), 1–4 June 2014, Ann Arbor, Michigan, USA. Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8057/8153

Wen, M., Yang, D., & Rosé, C. P. (2014b). Sentiment analysis in MOOC discussion forums: What does it tell us? In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK (pp. 185–192). International Educational Data Mining Society.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine-readable dictionary (Version 2). *Behavioral Research Methods, Instruments, and Computers*, 201, 6–11.

Wong, B. Y. L. (1985). Self-questioning instructional research: A review. *Review of Educational Research*, 55, 227–268.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300.