Received: 3 June 2020 | Revised: 13 May 2021 | Accepted: 5 July 2021

DOI: 10.1111/lnc3.12433

ARTICLE

WILEY

Natural language processing as a technique for conducting text-based research

Laura K. Allen | **Sarah D. Creer** | **Mary Cati Poulos**

Department of Psychology, University of New Hampshire, Durham, New Hampshire, USA

Correspondence
Laura K. Allen, Department of Psychology, University of New Hampshire, Durham, NH 03824, USA.
Email: laura.allen@unh.edu

Funding information
Office of Naval Research, Grant/Award Numbers: N00014-19-1-2424, N00014-20-1-2627; Institute of Education Sciences, Grant/Award Numbers: R305A180144, R305A180261, R305A190063

Abstract

Research in discourse processing has provided us with a strong foundation for understanding the characteristics of text and discourse, as well as their influence on our processing and representation of texts. However, recent advances in computational techniques have allowed researchers to examine discourse processes in new ways. The purpose of the current paper is to build on prior work in this domain and describe how new methodologies that consider the multi-dimensional nature of texts can serve as a complement to the existing literature. We focus on natural language processing (NLP) methodologies, in which computers calculate information about the linguistic and semantic properties of language data. We first provide a context for the origins of computational discourse analysis through the integration of research across computer science and psychology. We then provide an overview of different NLP methodologies and describe prior work that has leveraged these techniques to advance theoretical perspectives of discourse comprehension and production. Finally, we propose new areas of research that integrate these advances with traditional research methodologies in the field.

1 | INTRODUCTION

Despite being commonplace in daily life, both reading and writing are complex processes. A wealth of research has investigated these processes, from lower level processes that enable readers to perceive, decode and recognise words on the page to higher level processes that assist readers in constructing meaning for the text as a whole and ensuring the text makes sense. From this literature, it is clear that reading and writing emerge from a host of sub-processes, are influenced by contextual and task information and rely on varied sets of knowledge and skills. Accordingly, a number of models of discourse comprehension describe comprehension processes in a variety of contexts (e.g., Gernsbacher, 1997; Graesser et al., 1994; Kintsch, 1998; Kintsch & van Dijk, 1978; O'Brien & Cook, 2016a, 2016b; van den Broek et al., 2005; Zwaan et al., 1995).

This field of discourse comprehension has a strong foundation in experimental psychology methodologies, which has led to the rigorous testing of comprehension theories and has established our awareness of the representations, processes and characteristics of language that drive our understanding of discourse. As a result, our knowledge about discourse processes derives predominately from research designed with tight control over materials and methodologies that isolate and examine highly specific, uni-dimensional elements of comprehension. Accordingly, researchers often manipulate single characteristics of the text or task (e.g., text cohesion, reader goals) in order to isolate and examine the specific sub-processes contributing to comprehension. In such studies, online measures (e.g., lexical decision tasks, reading times for target sentences) aim to capture nuances associated with text processing such as the activation of semantically congruent (or incongruent) information from memory. Offline measures (e.g., free recall, essay writing) focus on developing an understanding of individuals' memory and comprehension of the text material. These methodologies have gone a long way in the discovery of important information about the comprehension process. As such, researchers have conducted decades of research manipulating singular features of texts and reading tasks to better understand the many processes that impact reading.

Recent advances in computational power, machine learning methodologies and availability of large amounts of language data (e.g., discussions in online courses; news articles from different sources) have offered the means to examine discourse processes in new ways. These sophisticated analysis techniques enable researchers to explore reading comprehension through a more ecologically valid lens, with increased complexity of materials and data sets. In this new era, it is no longer necessary to analyse texts uni-dimensionally—we can now consider discourse at multiple dimensions (e.g., we can look at words, sentence, discourse in *parallel*) and examine how these dimensions interact to influence cognitive and social processes in naturalistic discourse contexts. Additionally, increases in computational power have made it possible to rapidly calculate and model multiple characteristics of texts, such as the complexity of the syntactic constructions and the cohesion of the text content simultaneously. Thus, we can now consider texts at multiple dimensions (e.g., word, sentence, discourse) of the text as well as across varied contexts (e.g., narratives, essays, social media). For example, researchers can now collect large corpora from the Internet that span a variety of genres, such as literary novels, news articles and dialogue and examine how these genres may differ at multiple levels of analysis.

In this paper, we build on the substantial work that has been conducted on discourse processes and explore new methodologies that consider the multi-dimensional nature of texts as a complement to the existing literature. Specifically, we focus on natural language processing (NLP) methodologies, which have been developed to automatically calculate information about the linguistic and semantic properties of text and discourse. We first provide a brief history of

linguistic analyses of discourse, focusing on work conducted in both psychology and computer science. We then advocate for the inclusion of computational linguistic analyses in discourse research, focusing on how this work can be integrated to provide new insights into discourse that leverage the unique strengths and weaknesses of each field. We then describe more recent work that has leveraged NLP techniques to advance theoretical perspectives of discourse comprehension and production. Finally, we provide recommendations for researchers, including proposing new areas of research that integrate these advances with traditional research methodologies in the field and highlighting limitations for those beginning to implement these methodologies.

2 | A BRIEF HISTORY OF LINGUISTIC ANALYSIS IN DISCOURSE

Before discussing how NLP can be applied to discourse processing, it is important to understand the context from which these methodologies originate. There are two primary lines of research that have been developed to inform contemporary approaches to computational discourse analysis. One of these lines stems from work in psychology and aims to identify the cognitive processes underlying comprehension as well as the text- and reader-based features that elicit differences in these processes. The second line comes from computer science and aims to leverage naturally occurring linguistic data for the aim of developing computational models and algorithms that drive personalisation in applied domains. Although these lines have evolved separately, contemporary discourse analysis benefits from how each approach complements the other. Thus, it is important to understand each of their origins to gain insight into the strengths and limitations of each approach, as well as best practices for integrating them.

2.1 | Psychological approaches

There is a long history in psychological research of examining the content of discourse across a variety of dimensions (e.g., lexical, syntactic, semantic) and contexts (e.g., reflective writing, narrative texts, science textbooks, spoken dialogues) (Chi et al., 1994; Garrod & Anderson, 1987; Gerrig, 1993; McKoon et al., 1994; O'Brien et al., 1998; Singer et al., 1997; Zwaan, 1994). This research has allowed researchers to identify a number of text and discourse characteristics that relate to learning and performance, such as which features contribute to the overall *difficulty* or *quality* of varied texts (see McNamara et al., 2014 for a review). Substantial research has been devoted to identifying the linguistic features that affect individuals' processing and ultimate comprehension of texts (Bruner, 1986; Graesser et al., 2006, 2011; Haberlandt & Graesser, 1985; Kincaid et al., 1975). For example, causal information plays an integral role in comprehension, increasing readability and memory performance (Keenan et al., 1984; O'Brien & Myers, 1987; Trabasso & Sperry, 1985). Additionally, the degree to which a text is narrative or expository has been linked to its difficulty, with more narrative texts typically being easier to process (see Mar et al., 2021 for a review).

Research in this area has provided crucial information about the multiple linguistic dimensions that relate to the processing and production of discourse. However, human coding of every linguistic text feature that may impact comprehension processes can be an extremely time-consuming process and relies heavily on subjective ratings of text features. The laborious nature of this methodology has resulted in researchers examining individual features of text in isolation, rather than modelling discourse across multiple scales. This awareness of the impact of various

linguistic features on comprehension has led researchers to implement tight *control* over the structure and content of the materials presented to participants.

The extent of this control varies across studies, from researchers using adapted versions of published narratives or news articles to texts constructed specifically for the purpose of the experiment. Often, researchers will develop a set of materials that control for specific linguistic variables. Researchers may then manipulate specific linguistic features to determine their impact on comprehension processes or outcomes. For example, Kendeou et al. (2013) examined the impact of causal information on knowledge revision processes by systematically changing the amount of causal information provided in a text set over the course of several studies. They found that causal explanations can be used to overcome the interference of inaccurate information during knowledge revision. Similarly, Kleijn et al. (2019) examined the role of connectives on comprehension. In particular, they manipulated the types of connectives (e.g., additive, causal, temporal) in texts across varied reading contexts. They found that connectives were related to comprehension at local levels but not at global levels; however, further analyses revealed that these relations varied based on the type of connective and the difficulty of the text.

Overall, these approaches in the psychological domain have provided substantial insights into multiple dimensions impacting the processing and comprehension of text. However, there has been less of an emphasis on examining how multiple features of text may *interact* during comprehension, which has led to the use of texts that are not ecologically valid. For instance, to control for the varied properties of text, researchers will often simplify texts to such a degree that they may not vary in naturalistic ways. More recent approaches aim to integrate insights that allow for natural complexity that is inherent in ecologically valid texts.

2.2 | Computer science approaches

There is substantive precedence for applying computer science methodologies to attempt to answer common problems in psychological research. Perhaps most notably, a subset of researchers in cognitive science have relied on the use of artificial neural networks to model differential influences of top-down and bottom-up information on cognitive processes (e.g., Rosenblatt, 1958; Rumelhart et al., 1986). However, this work has often occurred independently from experimental work in discourse and thus does not typically focus on the same issues that are of high importance to discourse research.

The primary computational methodology for analysing language is *natural language processing (NLP),* which simply refers to the use of computers to process large amounts of human language data (Bird et al., 2009; Manning & Schütze, 1999). NLP techniques date back to Turing (1950) but have become a more central focus of computer science in the past couple of decades (Jurafsky & Martin, 2008). At the most basic level, NLP techniques can reveal important information about the word-level properties of language and discourse, which can be powerful predictors of language processes and abilities in a variety of contexts. However, texts can also be analysed at the level of the sentences and the overall document. For example, researchers can consider the concreteness of individual words in a text, the complexity or diversity of its sentence structures, or its overall coherence.

In the mid-1990s, research began to focus on sub-symbolic approaches that could be used to model the *semantics* (i.e., meaning) of individual words and texts. To this end, researchers have relied on vector-space models along with neural network machine learning models (e.g., recurrent networks, transformer networks) to represent text and language (i.e., word embeddings).

Here, researchers aimed to move away from analysing individual words and examine whether it was possible to model the broader meaning of a word, sentence or text (for a review of these earlier approaches, see Landauer et al., 2007). The general approach was to efficiently model the meaning of words by examining their patterns of use in naturalistic texts. The original vector-space models focused on representing meaning in high-dimensional spaces where words with similar meanings but no similar root words (e.g., dog and cat) would be located in similar regions of semantic space.

One of the first examples of these semantic models was *Latent Semantic Analysis*, which defined word meaning by calculating word co-occurrence in a word-document matrix and reducing the dimensionality such that each word is represented by its own semantic vector (Landauer et al., 2007). These models have been successfully applied to a number of machine learning tasks, such as the modelling of language acquisition (Landauer & Dumais, 1997) and grading of student essays (Miller, 2003). More recent attempts have refined these models to rely on less training data and incorporate more sophisticated and nuanced features of language, such as word order and other contextual factors (Devlin et al., 2019; Jones & Mewhort, 2007; Pennington et al., 2014). For example, BERT (Bidirectional Encoder Representations from Transformers) is a recent deep learning model that is designed to improve upon old word embedding models by pretraining data, fine tuning the model based on data that is unlabelled (i.e., it has not been scored by humans) and finally training the model on labelled data (Devlin et al., 2019). This approach enables researchers to use substantially less training data when developing their language models, provides better results in a number of classification tasks and allows for the same pre-trained model to be applied to a broader range of tasks. These sub-symbolic language models have been applied to many real-world applications, such as the prediction of mild cognitive impairments (Johns & Jamieson, 2018) and the prediction of personality profiles (Kwantes et al., 2016).

A computational approach to studying discourse comprehension through the use of NLP methodologies can be advantageous, particularly due to the ability to model discourse across these multiple dimensions. NLP analyses can be performed on multiple types of discourse, such as experimental materials, open-ended responses to text (e.g., think-alouds) and essays. Additionally, NLP can process language across multiple levels (e.g., words, sentences) and contexts (e.g., essay scoring, spam detection). Therefore, NLP can be vastly informative to researchers in a number of ways, such as assisting with material development, providing insight into a reader's mental representation of a text and contributing to understanding individuals' ability to produce coherent discourse in varied contexts.

An additional benefit is that NLP can remove the reliance on human coding. As a result, these techniques provide a more objective measure of specific indices, reducing human intuition and subjectivity during coding. NLP tools are also efficient, completing analyses of large data sets in a few seconds or minutes that would otherwise take human coders weeks or months. Researchers interested in employing NLP methodologies now have a much broader number of options, libraries that have been developed for popular programming languages (e.g., spaCy in Python and R; NLTK in Python; tidytext in R), as well as through freely available NLP tools for which programming knowledge is not necessary (Crossley et al., 2019; Kyle et al., 2018; McNamara et al., 2014).

Overall, there have been substantial improvements in the methodologies afforded to researchers for automatically analysing language and discourse. Despite the technological advances, there remains significant areas for improvement. Importantly, NLP is a theory-neutral methodology; thus, individuals across a number of disciplines and theoretical perspectives can benefit from these techniques. Therefore, while NLP techniques can analyse multiple dimensions across multiple levels, less research in this area has systematically examined why and how

these features of language may behave differently. For example, how might word concreteness influence text processing differently compared to text cohesion or affect? More recent work has taken an integrative approach by drawing from theoretical work in psychology and cognitive science more specifically. This integration has allowed researchers to consider language from a multi-dimensional perspective, wherein discourse emerges across multiple scales.

3 | NLP AND DISCOURSE PROCESSING RESEARCH

In recent years, interest in understanding language from both the psychological and computer science perspectives have converged. Implementing computational analyses of language have afforded researchers the ability to analyse texts along a variety of dimensions that can then be analysed in parallel. Discourse research has benefitted from the use of NLP to provide more nuanced information about the discourse comprehension and production processes (Crossley et al., 2014; Graesser et al., 2014; Kim et al., 2018; Tausczik & Pennebaker, 2010). For example, NLP techniques provide the opportunity to model complex interactions among text features and relate these to varying factors of the discourse context, such as performance and individual differences. Considerable work has built on this foundation and identified differential benefits of automatic analyses at multiple levels, from the sentence- and document-levels of texts (Crossley, Allen, et al., 2014; Crossley, Roscoe, et al., 2014; Nguyen & Rosé , 2011; Vytasek et al. 2019). This research has provided a strong foundation for identifying the features that drive and interact with discourse processes and led to work in a number of related areas, such as artificial intelligence, essay scoring and social network modelling (Dowell et al., 2019; Elfenbein, 2011; McNamara et al., 2018; Yan et al., 2020).

One of the first large-scale applications of *theoretically motivated* NLP to discourse was the analysis of text cohesion and readability (Graesser et al., 2004; McNamara et al., 2014). The purpose of these analyses was to provide more nuanced models of text readability that specifically focused on the aspects of texts that made them more or less difficult to comprehend for different types of learners. This research led to the identification of a number of text features that influence an individual's ability to effectively process and comprehend texts. For instance, *word frequency* has been cited as one of the key features of texts that can be manipulated to increase or decrease the difficulty of a text. The frequency of a word relates to how commonly it occurs in a given language (e.g., *car* is a high-frequency word in English, whereas *mitosis* is a low-frequency word). This metric has been linked to the speed at which individuals process texts, as well as how strongly the word is linked to rich networks of prior knowledge (Beck et al., 2002; Haberlandt & Graesser, 1985; Perfetti, 2007). Thus, texts with high-frequency words are often considered much easier to read than those with low-frequency words.

Beyond the word level, there are numerous aspects of a text that can influence its readability, such as its degree of narrativity or its syntactic complexity (Graesser & McNamara, 2011). Many of these features have been shown to affect the ease at which the text can be understood and recalled, even after controlling for readers' familiarity with the topic and the frequency of the words (Haberlandt & Graesser, 1985). One of the most commonly studied of these features is *cohesion*. A text's cohesion refers to the degree to which it contains explicit cues that signal its readers to make connections amongst ideas (Halliday & Hasan, 1976). For instance, connectives can specify relationships between ideas in a text and provide information about the types of relationships they signify (Longo, 1994). Research has shown that these features of texts can be manipulated to help scaffold less-skilled readers through the comprehension process (Graesser et al., 2011;

McNamara & Kintsch, 1996). For example, McNamara and Kintsch (1996) found that increased levels of text cohesion helped low prior knowledge readers better comprehend texts, but that decreased levels were beneficial for high prior knowledge readers.

Work in this area culminated in the development of Coh-Metrix, which was developed to provide automated analyses of the readability of texts at multiple dimensions (McNamara et al., 2014). Coh-Metrix was unique in its approach because it focused on the development of *theoretically derived* linguistic indices, including word information (e.g., frequency, concreteness, diversity), syntactic complexity, cohesion and genre (e.g., narrative, informational). Graesser et al. (2011) reduced these dimensions into a set of component scores that were representative of text difficulty along five primary dimensions: narrativity, syntactic complexity, word concreteness, referential cohesion and situation model cohesion. These components can now be readily applied to new texts to provide educators with specific recommendations on which texts may present challenges to different readers based on their varied sets of strengths and weaknesses. For instance, a student with high domain knowledge but lower language knowledge may be better scaffolded with specific text profiles compared to a peer with lower domain knowledge but higher language skills.

The development of Coh-Metrix represented one of the first attempts to integrate the computer science and psychological approaches to automated discourse analysis. Below, we provide examples of how these perspectives have been combined and extended in more recent work. In particular, we focus on the automated scoring of natural language, the modelling of individual differences in the producers of discourse, and finally, the inference of psychological processes from natural language.

3.1 | Evaluating writing

One of the most prevalent applications of NLP to discourse has been in the automated scoring of written language (see Yan et al. 2020 for a more substantive review). Automated essay scoring (AES) systems are commonly trained on a corpus of essays that have been rated by expert human raters according to a scoring rubric. This corpus is then divided into two sets of essays: a training set (used to train a model) and a testing set (used to examine the extent to which the model generalises to a separate set of essays). Machine learning algorithms are then applied to optimally fit the essays in the training set. The developed model is then applied to the essays in the testing set and these scores are compared to the scores of the human raters. An AES model is considered successful if the scores between the computer and humans are similarly aligned to the scores between humans. Indeed, AES has now reached a level of accuracy that the scoring of many classes of written essays is as accurate as expert human raters (Attali & Burstein, 2006; McNamara et al., 2015; Shermis et al., 2010; Taghipour & Ng, 2016).

Various AES systems have been developed to assess multiple genres of writing. e-Rater (Attali & Burstein, 2006) was one of the first and relies on a wide range of features that measure grammar, usage, mechanics, style, organisation, development, lexical complexity and prompt-specific vocabulary usage. The initial version of e-Rater offered users a method of manually combining these features using weighted averages in an intuitive and explainable system. More recent approaches rely on neural networks to score student writing. For example, SkipFlow (Tay et al., 2018) uses a mechanism for modelling relationships between hidden representation snapshots generated by Long Short-Term Memory Networks (LSTMs; Hochreiter & Schmidhuber, 1997). Alikaniotis et al. (2016) built on this work with the development of fully automated

framework based on LSTMs trained on the same dataset as SkipFlow, with a reported Spearman rank correlation coefficient of 0.91.

Similarly, Jin et al. (2018) introduced a two-stage neural network that aims to increase the performance of AES models in prompt-independent contexts. Their network was trained on human-rated essays with different prompts to detect essays with a level of quality that has high deviation from the average; then, these essays were used as pseudo-training data in the second stage. Finally, Cummins et al. (2016) proposed a constrained multi-task pairwise-preference model that is capable of successfully combining data from multiple tasks to improve the model's ability to generalise. This helped to address the need for AES models to be trained on large amounts of task-specific data through the sharing of feature encodings across different tasks.

Work in this area initially started from a desire to increase the efficiency of standardised test scoring. However, this work has subsequently led to an increase in theoretical research examining the ways in which we can define writing quality along multiple linguistic dimensions and contexts (Crossley, 2020). Text cohesion, for instance, has been identified as an important element of writing quality amongst young L1 writers (Struthers et al., 2013). High quality essays written by younger students tend to contain more causal, adversative, additive and manner adverbials than lower-quality essays (Myhill, 2008). This trend does not continue once writers have developed, however. Examinations of writing produced by adults have shown little to no relation between cohesive devices and writing quality. In fact, there is some evidence to suggest that writers' *flexible* use of cohesive devices across varied contexts is more strongly related to their writing skill (Allen et al., 2016).

Recently, research in NLP has begun to combine multiple approaches to provide more holistic, multi-dimensional accounts of written language. Wilson et al. (2017) identified latent factors for lexical complexity, syntactic complexity and referential cohesion in middle school students' written essays. Additionally, they found that measures of writing ability were predicted by the sentence and discourse factors. More recently, MacArthur et al. (2019) examined which linguistic features were predictive of writing quality and how those features changed over the course of instruction. They found that referential cohesion and lexical complexity were positively related to the quality of students' posttest essays, whereas syntactic complexity was negatively related to essay quality. Additionally, they found that students in the treatment condition produced posttest essays with greater lexical complexity, referential cohesion and connectives than a control group. These two studies serve as a proof of concept for computational linguistic analyses that consider the importance of written language and discourse across multiple dimensions.

3.2 | Psychological processes and NLP

NLP has also been used to infer psychological processes, most notably through the analyses of students' constructed responses generated during reading or other complex tasks (Magliano & Graesser, 2012). Students' responses to text can vary widely in format, from think-aloud statements collected during reading, to short answers to comprehension questions, or to essays written in response to a set of documents. Analyses of these open-ended responses with NLP techniques can provide more nuanced information about student learning than typical multiple-choice assessments.

The majority of advancements in this domain have been in the area of computer-based assessments, which are developed to provide students with summative and formative feedback on their reading and writing (e.g., Landauer, 2003; Litman et al., 2006; Magliano et al., 2011). The

development of these tools has provided important information about the processes underlying skilled text comprehension. For instance, Magliano and Millis (2003) and Magliano et al. (2011) have developed the Reading Strategy Assessment Tool (RSAT). In RSAT, students are prompted to provide open-ended responses (e.g., think-alouds, answers to questions) while reading various texts (Magliano et al. 2011). The system is then able to analyse the linguistic and semantic properties of these responses to automatically identify the specific comprehension strategies individuals use. Prior assessments of the effectiveness of RSAT (Magliano et al., 2011) have revealed that it is relatively accurate at predicting readers' comprehension of the texts as well as at discriminating the comprehension strategies they are employing.

Implications of this work have been twofold. First, researchers have been able to identify comprehension strategies that are most effective in varied contexts. For example, research with RSAT has demonstrated skilled readers tend to bridge statements in the texts they are reading and elaborate the material with inferences, whereas less-skilled readers engage more frequently in simple paraphrasing strategies. Second, this work has led to the development of a system that can be used to *enhance* comprehension, as the algorithms developed can now be used to provide readers with automated feedback during the reading process and encourage readers to engage in more effective comprehension strategies. Findings such as these reveal important information to researchers and educators about the processes that contribute to successful text comprehension, providing insight for how to support struggling readers.

3.3 | Individual differences in discourse processes

Finally, NLP can be used to provide information about the role of individual differences and other contextual factors in discourse processes (Bell et al., 2012; Pennebaker & King, 1999). For example, one of the most commonly used word-based NLP tools—the *Linguistic Inquiry and Word Count* (LIWC) software—has been used in a wealth of research to model the social and psychological characteristics of natural language. LIWC was developed by Pennebaker et al. (2007) to provide word counts based on psychologically relevant categories such as *positive emotion* words or *social* words (Tausczik & Pennebaker, 2009). Robinson et al. (2013) used LIWC to predict final course performance based on the words in students' self-introductions at the beginning of the semester. Research using LIWC and other similar word-based tools have confirmed the predictive power gleaned from simple analyses of the words in text and discourse.

Notably, NLP provides an avenue to examine *how* individual differences influence writing. To date, research has provided a wealth of information about *which* individual differences contribute to reading and writing performance (Cain et al., 2004; Jenkins et al., 2003; Santangelo et al., 2016; Smith & O'Brien, 2016; Swanson & Berninger, 1996; Turner & Engle, 1989). These studies tend to involve correlations between a score on an individual difference test (e.g., vocabulary knowledge, working memory) and a discourse task (e.g., science-based text comprehension, argumentative essay writing). Significant correlations between these scores are taken as evidence of a centralised role of individual differences in these processes. Incorporating NLP into this domain can extend our understanding of how individual differences in processing influence comprehension and writing performance. For instance, do limitations in working memory capacity influence writing performance at all levels (e.g., sentence production, argument development) or only at specific dimensions?

In this vein, research has been conducted to more closely examine how individual differences manifest in the properties of discourse that individuals produce (Crossley et al., 2011). Allen

et al. (2015) for instance, examined the extent to which reading skills could be modelled from the linguistic features of students' constructed responses to science texts. They found that linguistic features accounted for unique variance in students' reading comprehension, over and above simple quality metrics. In particular, more skilled readers generated constructed responses that were more cohesive and lexically sophisticated. Thus, these students may have relied more on specific academic language and engaged in establishing more connections amongst their responses across the text reading process. This example provides a demonstration of how these techniques could be used to provide more specific information about how individual differences manifest in individuals' processing and production of discourse.

4 | RECOMMENDATIONS FOR DISCOURSE RESEARCHERS

In this paper, we have provided a brief overview of the ways in which the fields of discourse processing and computer science complement each other and how NLP techniques may be used to inform and extend research on reading and writing. This is a new and emerging field with many open areas for future research. To conclude, we highlight two areas that may be particularly interesting for researchers in the discourse domain: the development of experimental materials and the examination of group-level discourse, which involves more than one or two individuals. We conclude with a cautionary note describing the current limitations and the potential for misuse of NLP techniques.

4.1 | Material development

As previously mentioned, the multi-dimensional nature of language provides challenges for researchers. Because naturalistic texts can vary among many dimensions including genre, reading level and cohesion, some researchers develop a set of experimental materials they control and manipulate across a series of studies to address various questions. Other researchers have noted that these tightly controlled texts are problematic because they are not naturalistic and argued for the use of longer multi-paragraph narratives and descriptions to increase ecological validity in reflecting how individuals commonly read (Graesser et al., 1994).

The inherent limitations of both approaches are reflected in Herb Clark's Language-as-a-Fixed-Effect Fallacy, which argues that findings from discourse research fail to identify the conditions under which the effects would generalise to new contexts (e.g., narrative vs. expository text). In addressing this, most researchers have turned to statistical analyses that can better account for variability in items (Baayen et al., 2008; Raaijmakers et al., 1999). Now, advances in computational modelling and NLP provide a means by which to explore the text materials implemented in discourse processing research to enhance ecological validity and determine generalisability.

Analysing experimental materials with NLP tools can help determine the dimensions across which the texts remain the same and those across which they might differ. When choosing ecological materials, researchers can use NLP analyses to ensure material sets remain consistent across critical dimensions. For example, NLP tools can be used to ensure experimental texts have similar levels of readability. Similarly, when researchers develop and manipulate texts, NLP can be helpful to determine if intentional changes between conditions were successful. For example, in experiments comparing causal explanations to non-causal conditions, NLP indices can detect the differences in causal cohesion but should remain the same on other dimensions.

Additionally, text manipulation can lead to unintended changes; for example, increasing cohesion tends to increase sentence length (Graesser & McNamara, 2011). Implementing NLP analyses during material development could help identify if any manipulations caused the texts to systematically differ in inadvertent ways. Research can benefit from including NLP methodologies during material development by informing researchers about the precise nature of the changes made during text manipulation, ensuring satisfactory manipulations and avoiding any unintentional changes that could also influence comprehension and confound the results.

Finally, NLP also provides a means by which to compare experimental materials to corpuses of ecological texts, thereby quantifying how experimental materials may differ. For example, the degree to which a text is narrative or expository has been commonly cited as an important aspect of its readability, with more narrative texts typically being easier to read (Bruner, 1986; Haberlandt & Graesser, 1985). Thus, NLP could provide a useful tool for researchers in choosing, writing, editing, and implementing text materials for studies of discourse processing.

4.2 | Social interactions

The application of NLP methodologies has also been extended to examine the cohesion of entire group interactions. For example, Dowell et al. (2019) have proposed a novel approach—Group Communication Analysis—that combines NLP techniques and social network analysis to examine large-scale group interactions. In particular, the authors analyse the features of individual's contributions in online communication contexts (e.g., online classes) using NLP and temporal interactions to identify roles with distinct behavioural patterns. This approach affords researchers the ability to examine how different types of interpersonal interactions may interact to influence collaboration and learning outcomes in online group discourse contexts. For instance, Dowell et al. (2019) found that roles related to learning outcomes. Specifically, roles with greater quantity of participation (i.e., "Drivers") performed better than roles with less participation (i.e., "Lurkers"). However, it was not just quantity of participation predicted learning, but also quality of participation as some in less engaged roles also performed well (i.e., "Socially Detached Learners").

Similarly, others have combined NLP and dynamic modelling techniques to examine how individuals shift their language style in different conversational contexts. For example, Müller-Frommeyer et al. (2020) examined how different conversational contexts (e.g., monologues vs. dialogues and conflict vs. friendly discussions) influenced individuals' dynamic patterns of language use. Specifically, they examined individuals' language *style,* which Pennebaker has described as their use of function words, such as pronouns and articles. They then applied Recurrence Quantification Analysis, which is a dynamic technique that examines nonlinear patterns in categorical or continuous data. Results of this study found that language style patterns were different across monologues and dialogues; additionally, exchanges with conflict and friendly discussions also differ in patterns of function word usage. Overall, these studies demonstrated that NLP techniques can be combined with other modelling approaches to provide more nuanced information about complex discourse contexts, such as group-level social interactions.

4.3 | A cautionary note

Despite the many avenues for future work with regard to NLP, we would be remiss if we didn't describe the limitations, caveats and potential for misuse with these methodologies. It can be

tempting to take these techniques and apply them unilaterally to all language data; however, it is important to be intentional about how and why NLP techniques are used and when and where they are implemented. We broadly categorised these limitations into two domains: theoretical and applied.

When using NLP, it can be tempting to examine all possible indices you are provided with—which can be many—rather than identifying a priori indices that would address specific theoretical questions. If not addressed, this can lead to inflating the risk of Type I error. Thus, researchers should be cautious when implementing NLP, especially with small datasets. Researchers should be sure to maintain best practices for statistical analyses. In practice, this may mean that when designing studies researchers should identify indices as part of the methodological design process, rather than only at the analyses phase. If researchers don't have a priori hypotheses, they should ensure exploratory post-hoc findings can be generalised to a reserved subset of the data or be replicated on a completely new dataset.

Additionally, just because NLP can provide an objective measure does not mean it is free from bias. This realisation is critical to consider in applied settings. As with many big data problems, such as training facial recognition technology, developing predictive models based on language data can be problematic when the training set is not representative (e.g., Amorim et al., 2018; Mehrabi et al., 2021). For instance, when NLP is used in the context of automated essay scoring it can lead to bias if not trained on a dataset that is representative of the entire community being assessed. The training set defines the goal standard, as a result any differences in features, such as dialect, that are not included in the training phase could be penalised for inherent language differences rather than being fairly evaluated on other elements of their writing. Therefore, when using NLP for applied purposes, it is important that researchers consider the language being evaluated and ensure their models generalise appropriately for their intended context.

5 | CONCLUSION

In this chapter, we provided a brief overview of how NLP techniques can be used to move the discourse field forward. There are numerous examples of the use of NLP to study language beyond those cited in this paper (see Chung & Pennebaker, 2019; McNamara et al., 2018 for reviews). Further, and perhaps more importantly, there remain a broad range of unanswered questions that can be addressed with these novel techniques. It is our aim that this introduction will provide readers with ideas for using NLP that they can develop to align with their own research ideas, leveraging NLP to provide insights into discourse that would otherwise be nearly impossible to achieve through traditional controlled studies. In particular, we view NLP to be a strong complement to the many methodologies that have already been developed in this area.

ACKNOWLEDGEMENTS

This research was supported in part by IES Grants R305A180261, R305A180144 and R305A190063 as well as the Office of Naval Research (grant no.: N00014-19-1-2424 and N00014-20-1-2627). Opinions, conclusions or recommendations do not necessarily reflect the view of the Department of Education, IES or the Office of Naval Research.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Laura K. Allen https://orcid.org/0000-0001-5582-1402

REFERENCES

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). *Automatic text scoring using neural networks*. arXiv:1606.04289.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with nat ural language processing techniques. *Proceedings of the 5th International Learning Analytics & Knowledge Conference (LAK'15)* (pp. 246–254). ACM.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, *108*(7), 911–924.

Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, pp. 229–237.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with E-rater R V.2. *Journal of Technology, Learning and Assessment*, *4*, 1–30.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Bell, C. M., McCarthy, P. M., & McNamara, D. S. (2012). Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles. In P. M. McCarthy, & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 545–556). http://doi.org/10.4018/978-1-60960-741-8.ch032

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bruner, J. S. (1986). *Actual minds, possible worlds*. Harvard University Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31–42. https://doi.org/10.1037/0022-0663.96.1.31

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Chung, C. K., & Pennebaker, J. W. (2019). Textual analysis. In H. Blanton, J. M. LaCroix, & G. D. Webster (Eds.), *Measurement in social psychology* (pp. 153–173). Routledge.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, *51*(5–6), 511–534. https://doi.org/10.1080/0163853X.2014.910723

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27.

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, *31*(2), 184–214. https://doi.org/10.1177/0741088314526354

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. Language Testing, 28(4), 561–580. https://doi.org/10.1177/0265532210378031

Cummins, R., Zhang, M., & Briscoe, E. (2016). Constrained multi-task learning for automated essay scoring. *Proceedings of the Association for Computational Linguistics*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1, 4171–4186.

Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*, *51*(3), 1007–1104. https://doi.org/10.3758/s13428-018-1102-z

Elfenbein, A. (2011). Research in text and the uses of Coh-Metrix. *Educational Researcher*, *40*(5), 246–248.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes*, *23*(3), 265–304. https://doi.org/10.1080/01638539709544994

Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly*, *70*(1), 3–22. https://doi.org/10.1093/poq/nfj012

Graesser, A. C., Magliano, J. P., & Haberlandt, K. (1994). Psychological studies of naturalistic text. *Advances in Discourse Processes*, *53*, 9–34.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh- Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text cohesion, on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395. https://doi.org/10.1037/0033-295X.101.3.371

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, *114*(3), 357–374.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*(4), 719–729. https://doi.org/10.1037/0022-0663.95.4.719

Jin, C., He, B., Hui, K., & Sun, L. (2018). TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 1088–1097.

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science*, *42*(4), 1360–1374.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing*. Prentice Hall.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 115–126.

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 854–865.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy-enlisted personnel. *Research Branch Report 8-75*. Naval Technical Training.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W., & van Djik, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kleijn, S., Pander Maat, H. L., & Sanders, T. J. (2019). Comprehension effects of connectives across texts, readers, and coherence relations. *Discourse Processes*, *56*(5–6), 447–464.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the big five personality traits with latent semantic analysis. *Personality and Individual Differences*, *102*, 229–233.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046.

Landauer T. K., McNamara D. S., Dennis S., & Kintsch W. (Eds.). (2007). *Handbook of latent semantic analysis*. Erlbaum.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308. https://doi.org/10.1080/0969594032000148154

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.

Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, *16*(2), 145–170.

Longo, B. (1994). The role of metadiscourse in persuasion. *Technical Communication*, *41*(2), 348–352. www.jstor.org/stable/43090348

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, *32*(6), 1553–1574.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student constructed responses. *Behavior Research Methods*, *44*, 608–621.

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. https://doi.org/10.1207/S1532690XCI2103_02

Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacognition and Learning*, *6*(2), 131–154. https://doi.org/10.1007/s11409-010-9064-2

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

Mar, R. A., Li, J., Nguyen, A. T., & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, *28*, 732–749.

McKoon, G., Ratcliff, R., & Ward, G. (1994). Testing theories of language processing: An empirical investigation of the on-line lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1219–1228. https://doi.org/10.1037/0278-7393.20.5.1219

McNamara, D. S., Allen, L. K., McCarthy, S. & Balyan, R. (2018). NLP: Getting computers to understand discourse. In K. Millis, Long, D., Magliano, J. & Wiemer, K. (Eds.), *Deep learning: Multi-disciplinary approaches*. Routledge.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). Hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59.

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

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, *29*(4), 495–512.

Müller-Frommeyer, L. C., Kauffeld, S., & Paxton, A. (2020). Beyond consistency: Contextual dependency of language style in monolog and conversation. *Cognitive Science*, *44*(4), e12834.

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. *Language and Education*, *22*(5), 271–288. https://doi.org/10.1080/09500780802152655

Nguyen, D., & Rosé, C. P. (2011). Language use as a reflection of socialization in online communities. *Proceedings of the Workshop on Languages in Social Media*, 76–85.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1200–1210. https://doi.org/10.1037/0278-7393.24.5.1200

O'Brien, E. J., & Cook, A. E. (2016a). Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes*, *53*, 326–338.

O'Brien, E. J., & Cook, A. E. (2016b). Separating the activation, integration, and validation components of reading. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 249–276). Elsevier Academic Press.

O'Brien, E. J., & Myers, J. L. (1987). The role of causal connections in the retrieval of text. *Memory & Cognition*, *15*(5), 419–427.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Operator's manual: Linguistic inquiry and word count: LIWC2007*. Lawrence Erlbaum Associates.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416–426.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, *32*(4), 469–479. https://doi.org/10.1177/0261927X13476869

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L., & The PDP Research Group. (1986). A general framework for parallel distributed processes. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (1, pp. 45–77). MIT Press.

Santangelo, T., Harris, K., & Graham, S. (2016). Self-regulation and writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 174–193). The Guilford Press.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed.). Elsevier.

Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes*, *24*(2–3), 199–228. https://doi.org/10.1080/01638539709545013

Smith, E. R., & O'Brien, E. J. (2016). Enhancing memory access for less skilled readers. *Scientific Studies of Reading*, *20*(6), 421–435. https://doi.org/10.1080/10888438.2016.1214590

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing*, *18*(3), 187–201. https://doi.org/10.1016/j.asw.2013.05.001

Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology*, *63*(2), 358–385. https://doi.org/10.1006/jecp.1996.0054

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891.

Tausczik, Y. R., & Pennebaker, J. W. (2009). Leadership in informal groups: A linguistic investigation of Wikipedia. *Annual Meeting of Group Processes*.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C. (2018). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 1.

Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language*, *24*(5), 595–611.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist approaches in accounts of reading comprehension. *Discourse Processes*, *39*(2–3), 299–316. https://doi.org/10.1080/0163853X.2005.9651685

Vytasek, J. M., Patzak, A., & Winne, P. H. (2019). Topic development to support revision feedback, feedback. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 220–224. https://doi.org/10.1145/3303772.3303816

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, *34*, 16–36. https://doi.org/10.1016/j.asw.2017.08.002

Yan D., Rupp A. C., & Foltz P. W. (Eds.). (2020). *Handbook of automated assessment: Theory into practice*. Taylor & Francis.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 920–933. https://doi.org/10.1037/0278-7393.20.4.920

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x

AUTHOR BIOGRAPHIES

Laura K. Allen is an Assistant Professor in the Psychology Department at the University of New Hampshire. The overarching aim of her research is to better understand the cognitive processes involved in language comprehension, writing, knowledge acquisition and conceptual change, and to apply that understanding to educational practice by developing and testing educational technologies.

Sarah D. Creer is a postdoctoral research associate in the Department of Psychology at the University of New Hampshire. Her research has examined the cognitive processes involved in the activation, integration and validation of information during reading. Specifically, she is interested in the impact of reader knowledge, contextual information and reader strategies on reading comprehension.

Mary Cati Poulos is a first-year PhD student in the Department of Psychology at the University of New Hampshire. Her research examines the multi-dimensional properties of mind wandering. She is particularly interested in the potential effects of different types of mind wandering on comprehension and the neural correlates underlying mind wandering.

How to cite this article: Allen, L. K., Creer, S. D., & Poulos, M. C. (2021). Natural language processing as a technique for conducting text-based research. *Language and Linguistics Compass*, e12433. https://doi.org/10.1111/lnc3.12433