Chapter 9: Discourse Analytics

Carolyn Penstein Rosé

Language Technologies Institute and Human–Computer Interaction Institute, Carnegie Mellon University, USA

DOI: 10.18608/hla17.009

ABSTRACT

This chapter introduces the area of discourse analytics (DA). Discourse analytics has its impact in multiple areas, including offering analytic lenses to support research, enabling formative and summative assessment, enabling of dynamic and context sensitive triggering of interventions to improve the effectiveness of learning activities, and provision of reflection tools such as reports and feedback after learning activities in support of both learning and instruction. The purpose of this chapter is to encourage both an appropriate level of hope and an appropriate level of skepticism for what is possible while also exposing the reader to the breadth of expertise needed to do meaningful work in this area. It is not the goal to impart the needed expertise. Instead, the goal is for the reader to find his or her place within this scope to discern what kinds of collaborators to seek in order to form a team that encompasses sufficient breadth. We begin with a definition of the field, casting a broad net both theoretically and methodologically, explore both representational and algorithmic dimensions, and conclude with suggestions for next steps for readers who are interested in delving deeper.

Keywords: Discourse analytics, collaborative learning, machine learning, analysis tools

Discourse analytics (DA) is one area within the field of learning analytics (LA; Buckingham Shum, 2013; Buckingham Shum, de Laat, de Liddo, Ferguson, & Whitelock, 2014). It includes processing of open response questions in educational contexts, and a large proportion of research in the area focuses on assessment of writing, but it encompasses more than that, including analysis of discussions occurring in discussion forums, chat rooms, microblogs, blogs, and even wikis. We consider LA broadly as learning about learning by listening to learners learn, with our listening normally assisted by data mining and machine learning technologies, though the published work in the area may precede but not yet include automation in all cases (Knight & Littleton, 2015; Milligan, 2015). Furthermore, we consider that what makes this area distinct is that the listening focuses on natural language data in all of the streams in which that data is produced.

This chapter offers a very brief introduction to this area situated within the field of LA broadly. DA is an area that has alternately suffered from two dangerous misconceptions. The first is an extreme over-expectation fuelled by the desire of many to have an off-the-shelf solution that will do their analysis work for them at the click of a button. Those falling prey to this misconception are almost certainly doomed to disappointment. Making effective use of either the most simple or the most powerful modelling technologies requires a lot of preparation, effort, and expertise. The second misconception is an extreme skepticism, sometimes resulting from disappointments arising from starting with the first misconception, or other times coming from a deep enough understanding of the complexities of discourse that it is difficult to get past the understanding that no computer could ever fully grasp the nuances that are there. While it is true that discourse is incredibly complex, it is still true that there are meaningful patterns that state-of-the-art modelling approaches are able to identify. Much published work from recent Learning Analytics and Knowledge and related conferences that illustrate the state-of-the-art are cited throughout this chapter. A recent survey on computational sociolinguistics tells the story from the perspective of the field of language technologies (Nguyen, Dogruöz, Rosé, & de Jong, in press), and might be of interest to dedicated readers.

The hope of this chapter is that it provides helpful pointers to readers who want to dig a little further. Two previous workshops on the topic of DA survey the foundational work within the LA community (Buckingham Shum, 2013; Buckingham Shum et al., 2014). An extensive overview of issues and methods situated more narrowly within the field of computer-supported collaborative learning can be found in three earlier published journal articles (Rosé et al., 2008; Mu, Stegman, Mayfield, Rosé, & Fischer, 2012; Gweon, Jain, McDonough, Raj, & Rosé, 2013). A short course in the area can be found in the text-mining unit of the Fall 2014 Data, Analytics, and Learning[1] MOOC offered on the edX platform. Other resources will be presented at the end of this chapter.

In this chapter, we are interested in the natural language uttered during episodes of learning. We seek to be theoretically and methodologically inclusive. Much of the existing work on discourse analytics views learning and its connection with language from a cognitive lens, in other words, seeking categories of language behaviour whose presence in a discourse makes predictions about learning gains because of the connection between the associated discourse processes and cognitive processes associated with learning. In this chapter, we seek to view learning and its connection with language through a social lens in order to leverage the important interplay between the cognitive and social factors in learning (Hmelo-Silver, Chinn, Chan, & O'Donnell, 2013; O'Donnell & King, 1999). For example, we seek to identify discourse processes that reveal underlying dispositions, attitudes, and relationships that play a supporting (or sometimes interfering) role in the learning interactions. Regardless of the situation in which it is uttered, natural language is deeply personal and deeply cultural. Embedded within it are artifacts of our personal experiences and those of generations that came before us. The details of language choices provide clues about the identities we purposefully project as well as sometimes those we seek to hide or even those of which we are not consciously aware. They project assumptions about and attitudes towards our audience and our positioning with respect to our audience, or sometimes just assumptions we want our audience to think we are making. We use these choices as currency in an economy of relationships in which we seek to achieve goals that we have adopted (Ribeiro, 2006).

With this understanding, as we use computation as a lens to aid in our listening to learners, we must acknowledge that we are always abdicating some of the responsibility for interpretation to the technologies that sit between us and the learning process, including whatever was lost or transformed in the recording into some digital form, and the further reduction and transformation that occurred during the application of the analytic technology (Morrow & Brown, 1994). With that caveat in mind, in this chapter we will focus heavily on questions of model interpretation and assessment of validity.

SCOPE AND FOCUS OF THIS CHAPTER

When one initially thinks about analytics, algorithms immediately pop to mind (Witten, Frank, & Hall, 2011). However, it is important to take a lesson from applied statistics and instead think about representation first. At the heart of DA work is a focus on representation of the data. Machine learning models cannot be applied directly to texts. Rather, the predictor features must be extracted from the text. These predictor features can be conceived of as questions: "Is __ found in the text?" or "How many times is __ found in the text?" If each feature is one of these questions, then for each instance, the feature value is the answer to the question. Interested readers can get a good feel for the breadth of simple features that can readily be extracted from text and what impact they have on predictive accuracy of classification models by experimenting with the publically available LightSIDE tool bench[2] (Mayfield & Rosé, 2013; Gianfortoni, Adamson, & Rosé, 2011), a freely available, off-the-shelf workbench with an extensive user's manual, example data sets, instructions about process, and contact information for researchers who are willing to offer help.

The key to success with modelling technologies applied to text is to ask the right questions, which produce meaningful clues. Thinking about this question begins by considering how language is structured. Though on the surface language may appear to the naked eye as a monolithic, unstructured whole, the fact is that it is composed of multiple layers of structure, each described within a separate area of linguistics. An introductory survey of a linguistics textbook (O'Grady, Archibald, Aronoff, & Rees-Miller, 2009) would be a valuable resource for researchers desiring to get into this area of LA. At the finest grain is the sound structure level, referred to as phonology and phonetics. Here the basic sound units of a language and how they fit together into the syllabic structure of a language are described. A basic alphabet of sounds comprise the set of phonemes, but within dialects these may be

[1] https://www.edx.org/course/data-analytics-learning-utarlington-link5-10x

[2] http://lightsidelabs.com/research/

pronounced in particular ways, which carry social significance because of their association with a host of socially relevant variables such as ethnicity, socioeconomic status, and region. Just above that level, the inner structure of words is described in a layer referred to as morphology. This is where systems of affixes we learn in our grammar classes come into the picture, which change the tenses on verbs or number on nouns, among other things. Above that is the level of syntax, where the grammatical structure of whole sentences is described. Also at the level of a sentence is the area of semantics, which describes how meaning is composed through fixed expressions, by convention, or by composing smaller units, guided through syntax, and referencing low level semantic units at the level of lexical semantics. Above the sentence level is the level of discourse, where we find rhetorical strategies among other aspects of structure. While these technical terms might be unfamiliar to many readers, they may provide useful search terms for readers who desire to find relevant resources for further reading.

If one traces the history of several areas in which natural language data has been the target of automated analysis, we hear the same refrain, namely the key to valid modelling is design of meaningful representations. The hope in including this example in this chapter is that readers can be spared from learning the same lesson the hard way. Taking one of the earliest cases where this lesson about DA was well learned was that of automated essay scoring (Page, 1966; Shermis & Hammer, 2012). The earliest approaches used simple models, like regression, and simple features, such as counting average sentence length, number of long words, and length of essay. These approaches were highly successful in terms of reliability of assignment of numeric scores (Shermis & Burstein, 2013); however, they were criticized for lack of validity in their usage of evidence for assessment. In later work, the focus shifted to identification of features more like what instructors included in their own rubrics for scoring writing. This investigation led to inclusion of content focused features, including techniques akin to factor analysis such as latent semantic analysis (LSA: Foltz, 1996) or latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004) to aid in content based assessments, though these still fall prey to problems with unigram features since they are also usually grounded in a unigram language representation. Other factor analytic language analysis approaches such as CohMetrix (McNamara & Graesser, 2012) have recently been used for assessment of student writing along multiple dimensions, including such factors as cognitive complexity. In highly causal domains that build in some level of syntactic structural analysis, CohMetrix has shown benefits (Rosé & VanLehn, 2005). In science education, success with assessment of open-ended responses has been achieved with LightSIDE (Nehm, Ha, & Mayfield, 2012; Mayfield & Rosé, 2013).

At this point, it is useful to return to the tension between the over- and under-expectation of DA. If we think about the challenges in identifying appropriate, meaningful features, we must come to terms with the limitations of the lenses we construct through modelling tools. The analytic technologies applied in DA may serve as a lens in the hands of researchers or practitioners that sits between them and the episodes of learning that occur within the world, or they may be a filter that mediates the interaction between learners and instructors, between learners, or between learners and learning technologies. Lenses are useful precisely because they do not simply transfer the exact details of the world viewed through them. Instead they accentuate aspects of those images that would not as effectively been seen without them. That is what we need them to do. At the same time, they obscure other details that are deemed less interesting by design. Lenses always distort. But in order to use them in a valid way, we must understand what each accentuates and obscures so that we can select an appropriate lens, and so we can interpret what we see in a valid way, always questioning how the picture would be different without it or with a different lens. Thus, from the beginning, we would caution those who consume the research in this area, develop these lenses, or actively apply them in research or practice, to be wary of what is inevitably lost or transformed in the process of application. Now this chapter will turn its attention to specific areas within the scope of DA.

REPRESENTATION OF TEXT

Key decisions that strongly influence how the data will appear through the analytic lens are made at the representation stage. At this stage, text is transformed from a seemingly monolithic whole to a set of features that are said to be extracted from it. Each feature extractor asks a question of the text, and the answer that the text gives is the value of the corresponding feature within the representation. Imagine that all you knew about a person was the set of answers to questions posed during a game of twenty questions, and now your task is to classify that person into a number of social categories of interest. If the questions are carefully constructed, you may be able to make an accurate prediction; nevertheless, you must acknowledge that much information and insight into that person as an individual will have been lost in the process. Once information is lost at this important stage in the process, it cannot be recovered through application of an algorithm, no matter how advanced and generally effective that algorithm is. Thus, we emphasize throughout this chapter the importance of careful decision making about representation, careful reflection about interpretation, and careful questioning of the validity of inferences made. While readers new to this area may find these caveats somewhat illusive, they will become clearer with experience.

Overview

Unigram features are the most typical feature extractors used in text mining problems. In the case of a unigram feature space, for each word appearing within the set of texts in the training data, there will be a corresponding feature that asks about the presence of that word within each text. While unigram feature spaces frequently achieve reasonably high performance, the models often fail to generalize beyond data collected under very similar circumstances to that of the training data. The reason for the lack of generalization is that these unigram models essentially memorize for each class value label in a superficial fashion what kinds of things people talk about in the set of instances associated with that label in the training data. If there is some consistency in that, then it can be learned by these models, but that consistency rarely generalizes very far. Generalization comes when the features extracted come from a relevant layer of structure.

The purpose of the feature-based representation of text is frequently to enable predictive modelling for classification or numerical assessment, where the objective is to achieve this predictive modelling with the highest possible accuracy (Rosé et al., 2008; McLaren et al., 2007; Allen, Snow, McNamera, 2015). This orientation will be the focus of this section. However, it is important to note that in some work within the broad area of DA, the representation work is the focus, and meaning is made of the identified predictive features, and thus the predictive modelling, if any, serves mainly as a validation of the meaningfulness of the identified features (Simsek, Sandor, & Buckingham Shum, 2015; Dascalu, Dessus, McNamera, 2015; Snow, Allen, Jacovina, Perret, McNamera, 2015).

With respect to predictive modelling for classification, in this vector-based comparison, the chosen features should make instances that are of different categories look far apart within the vector space, and instances that are of the same category look close within the vector space. This principle can also be used to troubleshoot a text representation. Features that either make instances that should be classified the same way look different or make instances that should be classified differently look similar are very likely to cause confusion in the classifications made by models trained using representations that include those features. The problem is often either ambiguous features (i.e., features that mean different things in different contexts, but the representation does not enable leveraging that context in order to disambiguate) or fragmentation (i.e., the same abstract feature is being represented by several more specific features, some of which are missing or too sparse in your data). It may also be that the most meaningful features are simply missing from your feature space, and other features, which may correlate with the meaningful ones within the specific data used as training data, will often "steal the weight," which ends up being counter-productive when the model is applied to new data where the spurious correlations between the meaningful features and less meaningful features may not exist or may be different.

Case Study

In order to illustrate the thinking that goes into representation of text for DA, we will start with a common example, namely analysis of affect in text, otherwise known as *sentiment analysis* (Pang & Lee, 2008). It is one of the most heavily marketed applications of text mining, and it is frequently the first thing researchers think to apply to their text data when faced with analyzing it. We will begin by introducing some issues in this area of text analytics and conclude with an investigation of what these analytics do or do not offer in terms of explaining patterns of attrition in MOOCs, where one might reasonably expect to see more expressions of negative affect from students who are struggling and ultimately drop out. We will see that the picture is far more complex than that (Wen, Yang, & Rosé, 2014a). In leading the reader through this case study, the hope is that the reader will see how one might progress through cycles of data analysis from pre-conceptions that start out overly simplistic, but become more informed through iteration. The most interesting work in the area of DA, or any area of analytics applied to rich, relatively unstructured data, will follow a similar storyline.

Simplistic treatments of sentiment identify texts as exhibiting either a positive or negative sentiment, and rely on an association between words and this affective judgment. Thus, much work has gone into the construction of sentiment lexicons, which associate words with a positivity or negativity score. The area of sentiment analysis is well developed, gaining substantial representation in industry, providing services to businesses related to marketing issues. Nevertheless, the limitations of the technology are clear. Furthermore, what is learned from examination of the linguistic literature is that much about attitude is not conveyed in text through words that are specifically positive or negative (Martin & White, 2005). This can be illustrated with the following example related to the weather. A statement such as "The weather is beautiful today" contains the required positive word; however, "The sun is shining" is only obviously positive if one knows that typically sunny days are preferred over rainy days. "It's a great day for staying indoors," indicates that the weather is not so good, despite the presence of a positive word. "My rain boots are feeling neglected," could easily be taken as a positive comment about the weather despite the presence of a negative word.

Now we will investigate situations more close to home where the approach may fall short. Because sentiment analysis is one of the most widely known and widely used language technologies by researchers and practitioners in other fields who are interested in text, it is not surprising that analysis of forum data from MOOCs is one area where we find applications of this technology, and thus that work will be a convenient case study. The rationale for its application was that discussion forum data may be useful for understanding better how, why, and when students drop out of MOOCs, with the idea that students may drop out because they are dissatisfied with a course, and that dissatisfaction should be visible using sentiment analysis as a lens. In an early such investigation, however, Ramesh, Goldwasser, Huang, Daumé, and Getoor (2013) found no relation between overall sentiment expressed by students (as assessed using a completely automated method) and their associated probability of course completion. Adamopoulos (2013) developed a sentiment related assessment method to measure sentiment associated with different course affordances in order to understand what students express their attitudes about in course discussion forums. They used a combination of automatically identified sentiment expressions paired with a grounded theory approach to identify themes in the course aspects mentioned in connection with attitudes. With this more detailed view, they were able to identify that not attitude in general, but attitude towards the professor, the assignments, and other course materials had the strongest association with dropout. In more recent work (Wen et al., 2014a), we pushed the automated analysis further, increasing the accuracy of sentiment measurement, and contrasting sentiment expressed by a student versus sentiment they were exposed to as well as contrasting sentiment at the student level with sentiment at the course level. In this work, the exact connection between sentiment-related variables and dropout depended upon the nature of the course.

With more probing, it became clear that a far more nuanced way of characterizing affect in posts was needed. For example, negative affect expressed in purely social exchanges might be disclosure, leading to enhanced emotional connection. Problem talk in a problem-solving course might just indicate engagement with the material. Negative affect words, expressions, and images may come up in a literature course where stories about unfortunate or stressful events are discussed, and yet that expressed sentiment might have nothing to do with a student's feeling about the experience of reading that material or even discussing that material. We conclude that sentiment analysis is not as simple as counting positive and negative words. Individual words are not enough evidence of attitude, context matters. Some rhetorical strategies combine negative and positive comments in the same review, and sometimes sentiment is expressed indirectly. Nuances like this observed through qualitative analysis must be taken into account when representing your data.

UNSUPERVISED METHODS

A variety of factor analytic (Garson, 2013; Loehlin, 2004) and latent variable analysis techniques (Skrondal & Rabe-Hesketh, 2004; Collins & Lanza, 2010) have been popular in the area. These may be unsupervised (i.e., not requiring pre-assigned labels), supervised (i.e., requiring examples to have pre-defined labels), or lightly supervised (i.e., requiring some external guidance to learning algorithms, but not requiring a pre-assigned label for every example). In this section, we focus on unsupervised methods. The most popular such techniques in the education space include factor analytics approaches like latent semantic analysis (LSA: Foltz, 1996) or structured latent variable models like latent Dirichlet allocation or LDA (Blei et al., 2003) mentioned briefly above. Thus, here we delve slightly deeper into the details and discuss strengths and limitations. In recent work in LA, unsupervised approaches have been used for exploratory data analysis (Joksimović et al., 2015; Sekiya, Marsuda, & Yamaguchi, 2015; Chen, Chen, & Xing, 2015), sometimes paired with visualization techniques (Hsiao & Awasthi, 2015), or alternating with or building on hand analysis (Molenaar & Chiu, 2015; Ezen-Can, Boyer, Kellog, & Booth, 2015). These modelling technologies have widely been used because researchers think of them as approximating an analysis of textual meaning. The reality is that they are much less apt at doing so than the prevailing view would have one believe. These tools do indeed have their place in the arsenal of DA tools. However, the hope of this chapter is to raise the curiosity of the reader to dig a little deeper in order to foster an appropriate scepticism, as described above.

Topic modelling approaches have become very popular for modelling a variety of characteristics of unlabelled data. A well-known and widely used approach is LDA (Blei et al., 2003), which is a generative model effective for uncovering the thematic structure of a document collection. Hidden Markov modelling (HMM) and other sequence modelling approaches are becoming popular for capturing progressions in student experiences (Molenaar & Chiu, 2015). Sometimes these approaches are combined in order to identify how language expression changes in predictable ways over time in terms of the representations of thematic content (Jo & Rosé, 2015). Statistical approaches such as these are meant to capture regularities. They are most valuable as tools in methodologies that value data reduction and simplification. Because they dismiss as noise the unusual occurrences within the data, they are less valuable in methodologies that seek unusual happenings that challenge assumptions. Though one might adopt an anomaly detection approach to identify instances that violate assumptions as a way of identifying such examples, in practice the examples found are more likely to be unusual in ways that are not necessarily interesting from the standpoint of challenging assumptions of theoretical import.

LDA works by associating words together within a latent word class that frequently occur together within the same document. The learned structure is more complex than traditional latent class models, where the latent structure is a probabilistic assignment of each whole data point (which is a document) to a single latent class (Collins & Lanza, 2010). An additional layer of structure is included in an LDA model such that words within documents are probabilistically assigned to latent classes in such a way that data points can be viewed as mixtures of latent classes. This structure is important for topic analysis. By allowing the representation of documents as arbitrary mixtures of latent word classes, it is possible then to keep the number of latent classes down to a manageable size while still capturing the flexible way themes can be blended within individual documents. Each latent word class is represented as a distribution of words. The words that rank most highly in the distribution are those treated as most characteristic of the associated latent class, or topic.

Because LDA is an unsupervised language processing technique, it would not be reasonable to expect that the identified themes would exactly match human intuition about organization of topic themes, and yet as a technique that models word co-occurrence associations, it can be expected to identify some things that would be expected to be associated. At heart, LDA is a data reduction technique. Its strengths lie in identification of word associations that are very common in a corpus, which frequently correspond to common themes. However, the common themes do not necessarily have a one-to-one correspondence with the themes of interest. Unfortunately, that means within the resulting representation, there will not be a distinct representation for those themes of interest that are not common. Similarly, unusual phrasings of common ideas will also typically fail to map to an intuitive representation within the LDA space. Representation of the textual data is also an important consideration. Typically, LDA models are computed over feature spaces composed of individual word features. Thus, whatever is not captured by individual words will not be accessible to the model.

SUPERVISED METHODS

At the other end of the spectrum are supervised methods. Taking a somewhat overly simplistic view, supervised machine learning methods are typically algorithms that operate over sets of vectors that associate a collection of predictor features, often referred to as attributes, with an outcome feature, often referred to as a class value. Recently, applications of supervised machine learning have been applied to the problem of assessment of learning processes in discussion. This problem is referred to as automatic collaborative-learning process analysis. Automatic analysis of collaborative processes has value for real-time assessment during collaborative learning, for dynamically triggering supportive interventions in the midst of collaborative-learning sessions, and for facilitating efficient analysis of collaborative-learning processes at a grand scale. This dynamic approach has been demonstrated to be more effective than an otherwise equivalent static approach to support (Kumar, Rosé, Wang, Joshi, & Robinson, 2007). Early work in automated collaborative learning process analysis focused on text-based interactions and click stream data (Soller & Lesgold, 2007; Erkens & Janssen, 2008; Rosé et al., 2008; McLaren et al., 2007; Mu et al., 2012). Early work towards analysis of collaborative processes from speech has begun to emerge as well (Gweon et al., 2013; Gweon, Agarwal, Udani, Raj, & Rosé, 2011). A consistent finding is that representations motivated by theoretical frameworks from linguistics and psychology show particular promise (Rosé & Tovares, in press; Wen, Yang, & Rosé, 2014b; Gweon et al., 2013; Rosé & VanLehn, 2005). We have already mentioned the LightSIDE tool bench as a good place to start getting experience in this area.

MOVING AHEAD

Readers who are interested in getting more familiar with the area of DA would benefit from digging first into some foundational literature. It is grounded in the fields of linguistics (Levinson, 1983; O'Grady et al., 2009), discourse analysis (Martin & Rose, 2003; Martin & White, 2005; Biber & Conrad, 2011), and language technologies (Manning & Schuetze, 1999; Jurafsky & Martin, 2009; Jackson & Moulinier, 2007).

REFERENCES

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. *Proceedings of the 34th International Conference on Information Systems: Reshaping Society through Information Systems Design* (ICIS 2013), 15–18 December 2013, Milan, Italy. http://aisel.aisnet.org/icis2013/proceedings/BreakthroughIdeas/13/

Allen, L., Snow, E., & McNamera, D. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 246–254). New York: ACM.

Biber, D., & Conrad, S. (2011). *Register, Genre, and Style*. Cambridge, UK: Cambridge University Press.

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

Buckingham Shum, S. (2013). *Proceedings of the 1st International Workshop on Discourse-Centric Learning Analytics* (DCLA13), 8 April 2013, Leuven, Belgium.

Buckingham Shum, S., de Laat, M., de Liddo, A., Ferguson, R., & Whitelock, D. (2014). *Proceedings of the 2nd International Workshop on Discourse-Centric Learning Analytics* (DCLA14), 24 March 2014, Indianapolis, IN, USA.

Chen, B., Chen, X., & Xing, W. (2015). "Twitter archaeology" of Learning Analytics and Knowledge conferences. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 340–349). New York: ACM.

Collins, L., & Lanza, S. T. (2010). *Latent class and latent transition analysis with applications in the social, behavioral, and health sciences*. Wiley.

Dascalu, M., Dessus, P., & McNamera, D. (2015). Discourse cohesion: A signature of collaboration. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 350–354). New York: ACM.

Erkens, G., & Janssen, J. (2008). Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer-Supported Collaborative Learning*, 3, 447–470.

Ezen-Can, A., Boyer, K., Kellog, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 146–150). New York: ACM.

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

Garson, G. D. (2013). Factor Analysis. Asheboro, NC: Statistical Associates Publishing. http://www.statisticalassociates.com/factoranalysis.htm

Gianfortoni, P., Adamson, D., & Rosé, C. P. (2011). Modeling stylistic variation in social media with stretchy patterns. *Proceedings of the 1st Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties* (DIALECTS '11), 31 July 2011, Edinburgh, Scotland (pp. 49–59). Association for Computational Linguistics.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, 5228–5235.

Gweon, G., Jain, M., McDonough, J., Raj, B., & Rosé, C. P. (2013). Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer Supported Collaborative Learning*, 8(2), 245–265.

Gweon, G., Agarwal, P., Udani, M., Raj, B., & Rosé, C. P. (2011). The automatic assessment of knowledge integration processes in project teams. *Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning, Volume 1: Long Papers* (CSCL 2011), 4–8 July 2011, Hong Kong, China (pp. 462–469). International Society of the Learning Sciences.

Hmelo-Silver, C., Chinn, C., Chan, C., & O’Donnell, A. (2013). *The International Handbook of Collaborative Learning*. Routledge.

Hsiao, I., & Awasthi, P. (2015). Topic facet modeling: Semantic and visual analytics for online discussion forums. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK ’15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 231–235). New York: ACM.

Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: Text retrieval, extraction, and categorization. Amsterdam: John Benjamins Publishing Company.

Jo, Y., Loghmanpour, N., & Rosé, C. P. (2015). Time series analysis of nursing notes for mortality prediction via state transition topic models. *Proceedings of the 24th* ACM *International Conference on Information and Knowledge Management* (CIKM ’15), 19–23 October 2015, Melbourne, VIC, Australia (pp. 1171–1180). New York: ACM.

Joksimović, S., Kovanović, V., Jovanović, J., Zouaq, A., Gašević, D., & Hatala, M. (2015). What do cMOOC participants talk about in social media? A topic analysis of discourse in a cMOOC. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK ’15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 156–165). New York: ACM.

Jurafsky, D., & Martin, J. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson.

Knight, S., & Littleton, K. (2015). Developing a multiple-document-processing performance assessment for epistemic literacy. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK ’15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 241–245). New York: ACM.

Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning support. *Proceedings of the 13th International Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (AIED 2007), 9–13 July 2007, Los Angeles, CA, USA (pp. 383–390). IOS Press.

Levinson, S. (1983). Conversational structure. *Pragmatics* (pp. 284–286). Cambridge, UK: Cambridge University Press.

Loehlin, J. C. (2004). Latent variable models: An introduction to factor, path, and structural equation analysis. Routledge.

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

Martin, J., & Rose, D. (2003). Working with discourse: Meaning beyond the clause. Continuum.

Martin, J., & White, P. (2005). *The language of evaluation: Appraisal in English*. Palgrave.

Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text accessible to non-experts. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Grading: Current Applications and New Directions* (pp. 124–135). Routledge.

McLaren, B., Scheuer, O., De Laat, M., Hever, R., de Groot, R., & Rosé, C. P. (2007). Using machine learning techniques to analyze and support mediation of student E-discussions. *Proceedings of the 13th International Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (AIED 2007), 9–13 July 2007, Los Angeles, CA, USA (pp. 331–338). IOS Press.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied Natural Language Processing: Identification, Investigation, and Resolution* (pp. 188–205). Hershey, PA: IGI Global.

Milligan, S. (2015). Crowd-sourced learning in MOOCs: Learning analytics meets measurement theory. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK ’15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 151–155). New York: ACM.

Molenaar, I., & Chiu, M. (2015). Effects of sequences of socially regulated learning on group performance. *Proceedings of the 5th Internation Conference on Learning Analytics and Knowledge* (LAK ’15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 236–240). New York: ACM.

Morrow, R. A., & Brown, D. D. (1994). Deconstructing the conventional discourse of methodology: Quantitative versus qualitative methods. In R. A. Morrow & D. D. Brown (Eds.), *Critical theory and methodology: Contemporary social theory*, Vol. 3 (pp. 199–225). Thousand Oaks, CA: Sage.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C. P., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, 138, 285–305.

Nehm, R., Ha, M., & Mayfeld, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21, 183–196.

Nguyen, D., Dogruöz, A. S., Rosé, C. P., & de Jong, F. (in press). Computational sociolinguistics: A survey. *Computational Linguistics*.

O'Donnell, A., & King, A. (1999). *Cognitive perspectives on peer learning*. Routledge.

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2009). *Contemporary linguistics: An introduction*. Boston/New York: Bedford/St. Martins.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 48, 238–243.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS Workshop on Data Driven Education: Advances in Neural Information Processing Systems* (NIPS-DDE 2013), 9 December 2013, Lake Tahoe, NV, USA. https://www.umiacs.umd.edu/~hal/docs/daume13engagementmooc.pdf

Ribeiro, B. T. (2006). Footing, positioning, voice: Are we talking about the same things? In A. De Fina, D. Schiffrin, & M. Bamberg (Eds.), *Discourse and identity* (pp. 48–82). New York: Cambridge University Press.

Rosé, C. P., & Tovares, A. (2015). What sociolinguistics and machine learning have to say to one another about interaction analysis. In L. Resnick, C. Asterhan, & S. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association.

Rosé, C. P., & VanLehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *International Journal of Artificial Intelligence in Education*, 15, 325–355.

Rosé, C. P., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F., (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer Supported Collaborative Learning, 3(3), 237–271.

Sekiya, T., Marsuda, Y., & Yamaguchi, K. (2015). Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 330–339). New York: ACM.

Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge.

Shermis, M., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Annual National Council on Measurement in Education Meeting, 14–16.

Simsek, D., Sandor, A., & Buckingham Shum, S. (2015). Correlations between automated rhetorical analysis and tutor's grades on student essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 355–359). New York: ACM.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multi-level, longitudinal, and structural equation models*. Chapman & Hall/CRC.

Snow, E., Allen, L., Jacovina, M., Perret, C., & McNamera, D. (2015). You've got style: Writing flexibility across time. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 194–202). New York: ACM.

Soller, A., & Lesgold, A. (2007). Modeling the process of collaborative learning. In H. U. Hoppe, H. Ogata, & A. Soller (Eds.), *The role of technology in CSCL: Studies in technology enhanced collaborative learning* (pp 63–86). Springer. doi:10.1007/978-0-387-71136-2_5

Wen, M., Yang, D., & Rosé, C. P. (2014a). 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. International Educational Data Mining Society. https://www.researchgate.net/publication/264080975_Sentiment_analysis_in_MOOC_discussion_forums_What_does_it_tell_us

Wen, M., Yang, D., & Rosé, C. P. (2014b). Linguistic reflections of student engagement in massive open online courses. *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.cs.cmu.edu/~mwen/papers/icwsm2014-camera-ready.pdf

Witten, I. H., Frank, E., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques*, 3rd ed. San Francisco, CA: Elsevier.