Big Data Approaches to Study Discourse Processes

Michael N. Jones & Melody W. Dye, *Indiana University*

Discourse science has a deep tradition of analyzing naturally occurring linguistic sources to explore higher-order cognitive phenomena. However, the study of discourse presents unique challenges for quantitative analyses due to its relatively large unit of interest. Stable estimates of letter or phoneme frequency can be obtained from even small text corpora, and stable estimates of word frequency from corpora the size of the classic Brown Word Corpus (at least in the higher range of the frequency spectrum). But as the unit of analysis increases in size, the corresponding increase in text required to obtain stable estimates of a target unit or phenomenon goes *way* up. The multi-sentence unit of analysis that interests discourse scientists has traditionally made quantitative analyses elusive, imprecise, or even impossible due to a poverty of data sources. However, the recent growth of big data resources has opened a whole new methodological toolbox for discourse researchers.

Over the past decade, a number of significant advances have been made both in available technologies and resources, which have important implications for discourse analysis. The first is the advent of massive digital archives of human speech and text. As human interaction has increasingly shifted online, enormous streams of data have been generated in the wake of this shift, on the order of 2.5 quintillion bytes daily (McAfee et al., 2012). The web has thus established itself as a vast repository of social engagement and communication (Gernsbacher, 2014), which can be selectively crawled and mined to create curated text corpora. To date, researchers have mined texts from a wide array of online interactive environments, including social networks, blogs, comment boards, community forums, and review sites, among others. In principle, these digital traces can be harvested from any site within the public domain for subsequent processing and analysis.

At the same time, offline resources are now, increasingly, being digitized. This has led to the creation of multimillion word subtitle corpora, which cover film, television, and radio; the digital archives of major national newspapers and periodicals; and the scanning of millions of books and articles. Today, one of the most promising ventures is the HathiTrust Digital Library, which represents a collaborative effort on the part of Google, the Internet Archive, and partner universities across the nation, to digitize and provide searchable access to the millions of volumes in America's research libraries. Liberman (2010) has likened the advent of such tools in this century to the invention of the telescope in the 17th. As he noted, these advances have begun to allow scholars to study discourse in a way that was never before thought possible: uncovering the intricacies of underlying patterns, revealing the idiosyncratic practices of individuals, collectives and cultures, and informing scientific questions about the mental processes.

New methodologies for quantifying discourse are very much intertwined with emerging big data resources. In some cases, the magnitude of text data has led to new insight techniques that did not previously exist, and in other cases old techniques have again become very relevant with the power that the massive scales of data provide. Hence we do not intend to separate data sources or methods in this chapter; they both go together to provide new tools for new insights in discourse processes. In this section, we will highlight two broad clusters of approaches that capitalize on these advances in especially fruitful and productive ways.

Surface Level NLP Tools and Databases

The practice of counting and analyzing verbal sequences may seem a distinctly modern pursuit. In fact, its origins are ancient. Yule (1944) cites the practice of the Masoretes, Jewish scholars and scribes of classical antiquity, who were tasked with preserving scripture after the fall of the Jewish state in 70 A.D. As guardians of the Hebrew Bible, they painstakingly recorded and tabulated the frequency of its words, phrases, and verses by hand, then clustered them by form and meaning, producing output not dissimilar to rudimentary count algorithms. The difference, of course, is that in modern times, scholars can apply such analytic techniques to any text (or collection of texts) of their choosing, while reaping the considerable benefits from automation and from significant advances in text modeling.

Progress in this domain should be of particular interest to discourse researchers, as rapid increases in the scale of corpora available over the past decade has finally yielded the data necessary to quantitatively evaluate theories that depend on multi-sentence units. But that increase in corpus scale has also produced a need for tools to mine these largely unstructured text sources. In addition, many of the early projects were funded by the private sector's need for machine translation and text-to-speech systems, which produced excellent data to study lexical and grammatical effects, but at too low a resolution for discourse analysis. For example, Google's announcement in 2006 that they were releasing a one trillion-word corpus of web text was a huge excitement to discourse researchers. But when released, the database consisted of only decontextualized n-gram counts from the corpus, with the largest unit being a 5-gram.

Nonetheless, with the growth of raw text corpora, simple surface-level frequency counts of structural elements have proven to be quite useful in empirically testing intuitions about patterns of register variation, narrative structure, and to make genre comparisons. The largest collection of corpora currently available for these analyses is the ambitious HathiTrust Digital Library (https://www.hathitrust.org/). The HathiTrust collection currently consists of 13.8 million volumes (618 terabytes) of digitized text sources from a wide range of registers, genres, and languages, and that number is currently growing by more than a thousand books a day. To put this massive scale in context, if all the books from HathiTrust were lined up on a shelf, the shelf would need to be 163 miles long, and the collection would weigh over 11,000 tons. Hathitrust has a variety of search tools and APIs that can be used by researchers interested in data-driven discourse analyses. In addition to HathiTrust's collection, which is largely composed of published book sources, Brian MacWhinney's group has compiled large sources of conversation, including child-directed speech, second-language tutoring, dementia transcripts, and conversation data linked with video (see http://talkbank.org for data and tools).

The Brigham Young Corpus Repository (http://corpus.byu.edu/), curated by Mark Davies, contains some of the most widely used large-scale corpora, including contemporary, historical, political speech, and book corpora, as well as other unconventional sources such as American soap operas. Each corpus is completely searchable with customizable tools. While the most common use of the site is determining simple frequencies of words, phrases, and collocates, the tools are well-suited for discourse researchers to explore variation in register, dialect, or genre, to study historical change, and to create experimental stimuli balanced on a variety of contextual factors. The BYU corpus site sees over 130,000 unique visitors each month.

In addition to these unstructured text sources, considerable effort over the past decade has been put into the construction and validation of annotated text corpora. The largest repository of such databases is the Linguistic Data Consortium (https://www.ldc.upenn.edu). Annotated corpora allow for supervised training of language comprehension models and algorithms, compared to the unsupervised learning and surface text mining methods that must be applied to untagged corpora. While the vast majority of annotations have focused on units that are too small for the discourse researcher's questions—such as part-of-speech tagging, co-reference disambiguation, or semantic role labeling—there is great promise for such databases as the annotations grow to a higher level. For example, Banarescu et al.'s (2013) AMRs (Abstract Meaning Representations) afford discourse researchers a statistical representation of sentence meaning that is both precise and robust, and can be aggregated across multiple sentences to compare the meaning of larger contexts or to import into construction-integration frameworks.

Data such as those in the HathiTrust or BYU collections are largely free text, with their own proprietary analytic tools. But there are also ad hoc tools that researchers can feed their own corpora to evaluate usage patterns. The simplest—and perhaps most widely used—of these tools are count algorithms, which scan texts for words in user-defined dictionaries, and return count lists, organized by dimensions of interest (such as 'personal pronouns' or 'emotion words'). In the case of the popular Linguistic Inquiry and Word Count (LIWC) program, verbal categories have been devised to be psychologically meaningful, and usage patterns have been reliably linked to a range of traits, such as extroversion and depression (Tausczik & Pennebaker, 2010). This has led to the development of automated text scoring systems to assess personality, sociality, and mental health. A similar approach can be taken to annotated corpora. Multi-dimensional factor analysis was one of the first methods for quantifying variation among written registers, using the surface characteristics of the text as input to statistical models (Biber, 1988).

Another widely used tool to apply to unstructured text is CohMetrix (McNamara, Graesser, McCarthy, & Cai, 2014). CohMetrix was specifically designed to evaluate semantic coherence within texts, but computes 108 indices including metrics on referential cohesion, syntactic complexity, and pattern densities, and measures based on situation models. Crossley, Kyle, and McNaramara (2015) have recently expanded on CohMetrix with their TAACO software (Tool for the Automatic Analysis of Text Cohesion), which allows for offline large-scale batch processing of texts, and expands to 150 indices that include both local and global cohesion measures. Crossley, Allen, Kyle, and McNamara (2014) have also produced a Python tool to provide simple access to a suite of NLP tools customized for discourse researchers.

Often, the curated text corpora and search tools are inadequate for a researcher's purposes, which may require exploring how language is used in the real world across very specific times or following the occurrence of specific social events. In cases such as these, a growing number of open-source tools exist to mine social media and discussion sites, allowing targeting of online conversations surrounding desired topics. For example, Dehghani et al.'s (2016) TACIT software (Text Analysis, Crawling, and Interpretation Tool; http://tacit.usc.edu) is a browser-based plugin that allows the user to target a range of real-time text sources (e.g., US Senate and Supreme Court speech transcriptions, Twitter, Reddit), and to automatically apply corpus preprocessing and count routines.

"Deep" Semantic Tools

The corpus mining methods discussed so far tend to focus on surface-level statistics of linguistic units, such as frequency of occurrence, incidence scores, or type-token ratios. While these surface level counts can be very useful to study the commonality of constructions and how they vary across contexts, registers, and genres, they do tend to be rather shallow and limiting for many research questions in discourse. In addition, such methodologies have been criticized for their reliance on hand coded lists, their insensitivity to surrounding verbal context, and their unsophisticated handling of lexical statistics. More recent "deep" semantic tools have grown from techniques in data science and cognitive modeling. These methods allow the researcher to go beyond surface level statistics and to explore topic, situational, and thematic shifts in larger units of text. It is beyond the scope of this chapter to describe all available deep semantic tools in detail; we instead focus on two that have been very prominent in the discourse literature (but see Jones, Willits, & Dennis, 2015, for a survey of these techniques).

The introduction of Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998) was a significant advance for quantitative discourse analyses. LSA is an unsupervised learning algorithm, closely related to factor analysis, that infers a semantic vector representation for words based on co-occurrence patterns over a large corpus of text. Although two exact synonyms may be orthographically distinct, their resulting LSA vectors will be identical, projecting them into the same point in a high-dimensional space. The benefit of this transformation is that two text units that have roughly the same meaning, albeit with very different words, will be seen as extremely similar to LSA. This affords the discourse researcher a tool to evaluate higher-level semantic similarity independent of word overlap. LSA can be used as an exploratory tool when trained on a text corpus that a researcher wants to evaluate the higher-order discourse structure of. But more commonly, the "meaning-infused" vectors from an LSA model trained on a large corpus of text are used to evaluate discourse transitions, similarity, or cohesion in a new set of texts.

LSA has been widely applied in discourse analytic tools ranging from automated essay scoring to conversation analysis, to automated tutoring. It is a component of many of the offline tools described in the previous section as surface level (Coh-Metrix, TAACO, TACIT, etc.); indeed, this is why we selected these tools—they all have both surface level and deep semantic indices. Despite its successes, a major issue that limits LSA's application to discourse analytics has to do with its fundamental geometric assumption that word meanings can be represented as points in a high-dimensional space. The meaning of a text is then simply the sum of the individual words occurring in that text. While the method has no doubt been fruitful, this simple vector summing produces saturation that makes it imprecise and insensitive when applied to discourse units either larger or smaller (because of the lack of syntax) than a few paragraphs.

More recently, there has been a great deal of interest surrounding probabilistic topic models for exploratory research in discourse. A topic model is an unsupervised machine-learning algorithm, similar to LSA, but it has several fundamental differences. When trained on a large text corpus, a topic model attempts to infer the main themes most likely to have generated the text. Each document in a text corpus is assumed to be generated by a mixture of these topics, and the algorithm then statistically infers the most likely set of topics given the text data at hand. The algorithm itself is a text generalization of Latent Dirichlet Allocation, which is used in a variety of data discovery fields including genetics, image analysis, and social networks (see Blei, 2012, for a

review). Topic models have proven to be very useful insight tools that allow us to evaluate topic flow, thematic change over time, and situation models, or to zoom in and zoom out semantic resolution while analyzing a text to discover more specific or broader themes.

Topic models have been used to study linguistic change in online discussion communities, and to explore how group language and reference categories shift over time with discourse (e.g., McFarland et al., 2013). Recently, Murdock, Allen, and DeDeo (2015) have applied topic models to Darwin's reading notebooks, offering new insights into how his theory of natural selection emerged from the semantic path of his readings and writings. In addition, they provide an online tool for exploring the HathiTrust repository using topic models. Topic modeling is also an option in the aforementioned TACIT tool.

The growth of big data has also led to methodological developments in networks and complex systems, which are starting to see increased use in discourse analysis. Social networks of characters, events, and narrative schemas can be efficiently built and evaluated. While older generation architectures relied on extensive hand-crafted rules and external knowledge input, the field is moving towards automatic named-entity recognition and template extraction (Chambers & Jurafsky, 2008), vastly simplifying the problem for researchers seeking to investigate the structure of texts on a large-scale. At the same time, cross-recurrence analysis is now being used to explore alignment and synchrony in interpersonal communications in the real world, for example, by applying it to massive numbers of Twitter feeds during the U.S. Presidential elections (Fusaroli et al., 2015).

In related work on online communities, complexity methods have been used to explore how social status influences linguistic alignment among members (Reitter, 2016; Vinson & Dale, 2016). Such investigations can also offer insight into the fine-grained temporal dynamics of these processes. For instance, when a local discourse adaptation spreads across the network, its cascading effects can be traced in time, revealing the reach and magnitude of change across both communities and time-steps (Danescu-Niculescu-Mizil et al., 2013). Similarly, systemic shifts in the language can be charted across historical corpora and related to changing social conditions. In a striking example of this line of research, Hills & Adelman (2015) report evidence that English-speakers have adopted increasingly concrete language over the past two hundred years, allowing them to communicate effectively in an ever more competitive information marketplace.

Discourse Relations

While for some analyses, treating discourse as a bag of words, sentences, or topics would suffice, in other analyses this approach fails to provide a suitable semantic representation. The interpretation of discourse not only depends on the pieces of information it contains but also on how they are organized. Discourse structure has been studied under the notions of cohesion and coherence for about five decades now (Asher & Lascarides, 2003; Grosz & Sidner, 1986; Halliday & Hasan, 1976; Webber & Joshi, 2012). Recent application-oriented research in machine translation, question answering, sentiment analysis, and text summarization has revealed that considering the local context of a sentence has a huge impact on a system's ability to comprehend and generate natural discourse.

A hot topic of research that benefits these applications as well as cognitive and computational theories of discourse processing is involved with identification of discourse relations, such as causal and temporal relations between neighboring sentences in a text. While surface cues such as discourse markers (*because, therefore, after*) are sometimes present, most semantic relations at this level are implicit in the way the deep meaning of two neighboring sentences are connected (*Mary was sick yesterday. She took a day off and visited a doctor*). Big data can help modeling word-knowledge that is required for understanding typical relations between familiar events (e.g., that being seek is a reason for taking a day off or visiting a doctor). Semi-supervised methods mine explicit discourse relations that occur in natural text with surface cues (*x therefore y*) and generalize the obtained relational knowledge to less familiar or unseen events (Hernault et al., 2011; Marcu & Echihabi, 2002; McKeown & Biran, 2013; Sporleder, 2008; Pitler et al., 2009; Zhou et al., 2010).

Challenges and Future Directions

In the tech industry, big data has already effected a sea of change, as computer and data scientists have struck away from verbal conceptual theories founded on cherry-picked examples, to data-driven discovery. Yet while research in the private sector is driven by profit margins rather than basic science, the methods being developed there are increasingly relevant to the social and cognitive sciences. For scholars not versed in large-scale data extraction and analysis, the prospect of employing these methodologies in their own research can appear daunting. However, efforts are currently being mounted to make such tools more widely accessible. Powerful open-source scientific computing and data mining packages are being authored in various languages, significantly lowering the barrier to entry for novice coders, and more and more graphical applications are being developed that eliminate the need for code entirely. Large-scale text repositories have been made possible, in part, by optical character recognition technology, which converts the written and printed word into machine-readable text. As speech-to-text technology continues to advance, it will become possible to create parallel record keeping of spoken communication and oral histories.

How might the advent of big data begin to inform how discourse processes are studied? In the past, when such resources have been scarce or non-existent, discourse researchers have tended to rely on careful, sustained reading, on tightly controlled experimental studies, and on inferences from experience or 'intuition'. Big data in no way obviates the need for this foundational body of knowledge, nor for the accumulated wisdom and insights of its experts, who are best positioned to ask productive questions of it. As Picasso was rumored to have opined of computers: "But they are useless. They can only give you answers."

Undoubtedly, the story here is rather more complex than Picasso had it. The prime advantages and opportunities are threefold: With scale, (1) old questions can be addressed with greater clarity or precision, reifying old knowledge, and with new technology, (2) old questions can be answered in new ways, and (3) new questions can be posed that would have previously been either unanswerable or inconceivable. In short, innovation is possible not only in our methods, but in our modes of thinking.

References

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Banarescu, L., et al (2013). Abstract Meaning Representation for Sembanking, In the *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Biber, D. (1998). *Variation across speech and writing*. Cambridge University Press.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM,* 55, 77-84.

Chambers, N. & Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL/HLT 2008*.

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*, 511-534.

Crossley, S., Kyle, K., & McNamara, D. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods.*

Dehghani, M., et al. (2016). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods.*

Fusaroli, R., Perlman, M., Mislove, A., Paxton, A., Matlock, T., & Dale, R. (2015). Timescales of Massive Human Entrainment. *PLOS-ONE.*

Gernsbacher, M. A. (2014). Internet-based communication. *Discourse Processes, 51*, 359-373.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175–204.

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

Hernault, H., Bollegala, D., & Ishizuka, M. (2011). Semi-supervised discourse relation classification with structural learning. *Computational Linguistics and Intelligent Text Processing*, 340–352.

Hills, T.T., & Adelman, J. (2015). Recent evolution in the learnability of American English from 1800 to 2000. *Cognition*, 143, 87-92.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 368–375).

McKeown, K., & Biran, O. (2013). Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 69–73).

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics, 41*, 607-625.

Murdock, J., Allen, C., & DeDeo, S. (2015). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. eprint arXiv:1509.07175

Pitler, E., Louis, A., & Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the joint conference of the 47th annual meeting of the ACL* (pp. 683– 691).

Sporleder, C. (2008). Lexical models to identify unmarked discourse relations: Does WordNet help? *Lexical-Semantic Resources in Automated Discourse Analysis*, 20.

Reitter, D. (2016). Alignment in Web-based Dialogue: Who Aligns, and how Automatic is it? In Jones, M. N. (Ed.). *Big Data in Cognitive Science: From Methods to Insights*. Taylor & Francis.

Vinson, D.W., & Dale, R. (2016). Social structure relates to linguistic information density. In Jones, M. N. (Ed.). *Big Data in Cognitive Science: From Methods to Insights.* Taylor & Francis.

Webber, B., & Joshi, A. (2012). Discourse structure and computation: past, present and future. In *Proceedings of the* acl-2012 special workshop on rediscovering 50 years of discoveries (pp. 42–54).

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zhou, Z., Xu, Y., Niu, Z., Lan, M., Su, J., & Tan, C. (2010). Predicting discourse connectives for implicit discourse relation recognition. *In Proceedings of the 23rd international conference on computational linguistics*: Posters (pp. 1507–1514).