ConText: Software for the Integrated Analysis of Text Data and Network Data

Jana Diesner
Graduate School of Library and Information Science, University of Illinois Urbana Champaign

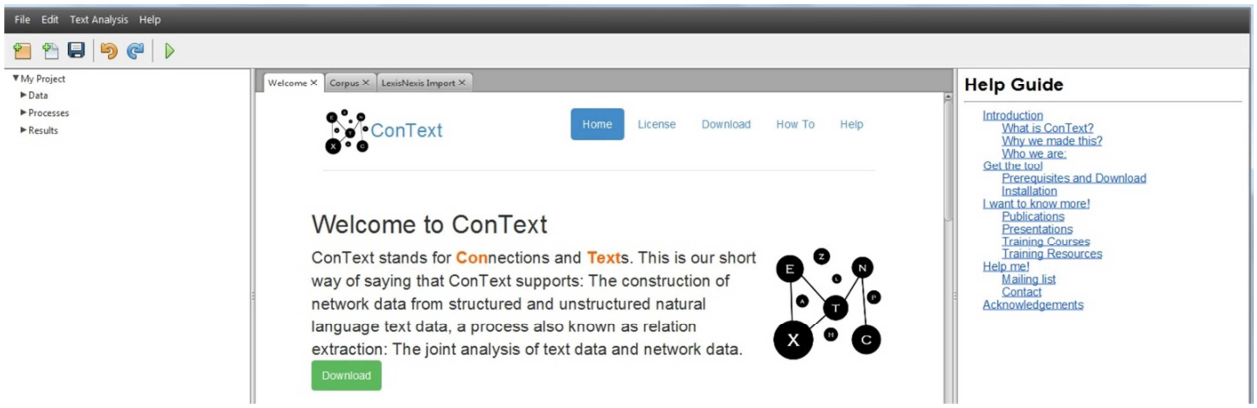
We are demoing ConText (http://context.lis.illinois.edu/), a software tool that we built to support a) the construction of different types of network data based on unstructured, semi-structured and structured natural language text data and b) the joint analysis of any such text data and network data. We have designed ConText as a general applicability tool for conducting text analysis and network analysis in an integrated, systematic and automated fashion, especially for researchers and practitioners from the digital humanities, computational social sciences and real-world application domains.

The different approaches to a) relation extraction) and b) combining explicit social network data with network data extracted from text data are the core of ConText (for on overview on these methods see (J. Diesner, 2012)). Overall, the research and development that are going into ConText are aiming to advance the rigorous integration of text analysis and network analysis. Work at the nexus of these two well-developed yet quickly evolving fields is still lacking behind in theory, methods and substantive understanding gained about the interplay and co-evolution of language use and social networks (Chang, Boyd-Graber, & Blei, 2009; J. Diesner, 2013; Gruzd, 2009; Kirchner & Mohr, 2010; McCallum, Wang, & Mohanty, 2007; Mihalcea & Radev, 2011; Roth & Cointet, 2010).

Starting from the most efficient level, ConText supports the construction of meta-data networks. Meta-data are concise descriptors of the content of a body of text data, such as keywords and index terms. These routines do not consider the actual content of text bodies. Diving deeper into the substance of text data, we support the extraction of association networks (Danowski, 1993), semantic networks (J. Diesner & Carley, 2011a), syntax-based networks and multi-mode networks (J. Diesner & Carley, 2011b) from text bodies. The resulting networks are sometimes referred to as socio-semantic networks. At the lowest common denominator, text-based networks consist of concepts (nodes) and the connections between them. Concepts are explicitly or implicitly encoded by text terms, and can consist of unigrams and meaningful n-grams. Links encode different types of associations or relationships between concepts. The outlined types of networks extracted from meta-data and/ or the substance of text data can be combined with social network data that represent the agents who have produced, shared or perceived the underlying pieces of information. Networks extracted from text data can be used for a variety of purposes. Traditionally, semantic networks were intended to be used for reasoning and inference, for identifying the meaning of concepts or texts in terms of emerging clusters of nodes in the neighborhood of a focal concept, and for various summarization and prediction tasks related to text data (Collins & Loftus, 1975; Griffiths, Steyvers, & Tenenbaum, 2007; Sowa, 1992). Advances in methods and computational power have enabled a wider spectrum of applications (Mihalcea & Radev, 2011).

ConText is currently provided for free as a client-side executable. Being built in Java, the technology runs on PCs and Macs. The tool integrates a variety of open source libraries as well as functionalities that we developed from scratch. The graphical user interface (GUI) is designed to allow for intuitive use and full user control over sequential and complex analysis processes. Figure 1 shows the GUI of ConText. Individual data analysis projects are organized into data, processes and results (left hand side panel). We also provide a handbook, training material and a mailing list; all accessible from the tool's webpage (http://context.lis.illinois.edu/).

Figure 1: Graphical User Interface of GUI



The corpus view as shown in Figure 2 displays the text data currently being processed. Showing the actual text data and allowing users to browse through each document at its different stages of processing supports qualitative analysis, the deep screening of small samples and close readings (Abello, Broadwell, & Tangherlini, 2012). We designed ConText to combine these paradigms with big data analytics, scalability and distant readings by implementing efficient algorithms for supporting a variety of statistical analysis, information extraction and dimensionality reduction tasks.

Figure 2: Corpus View

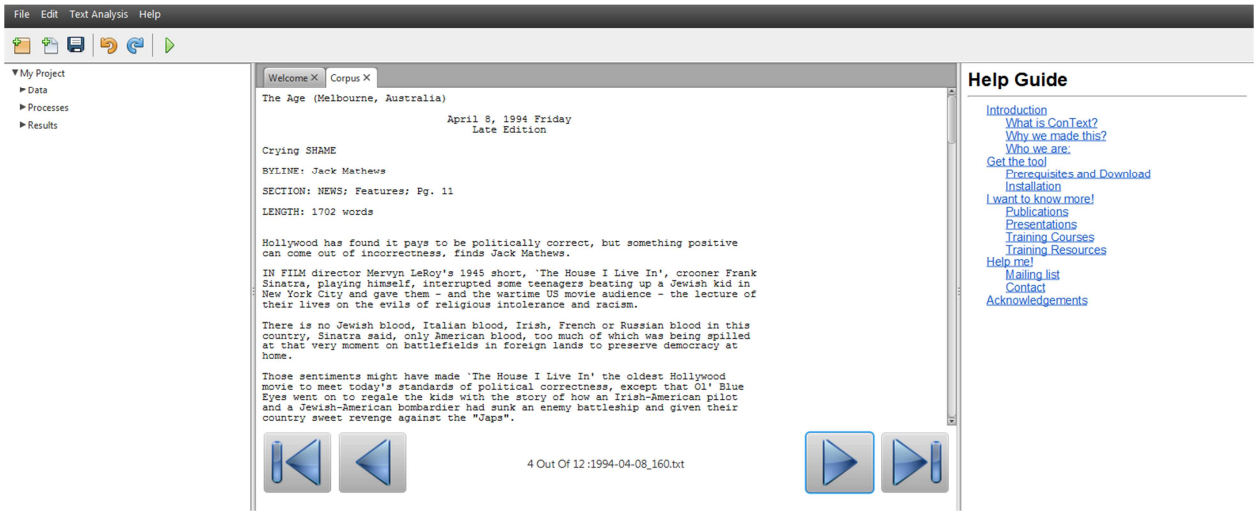
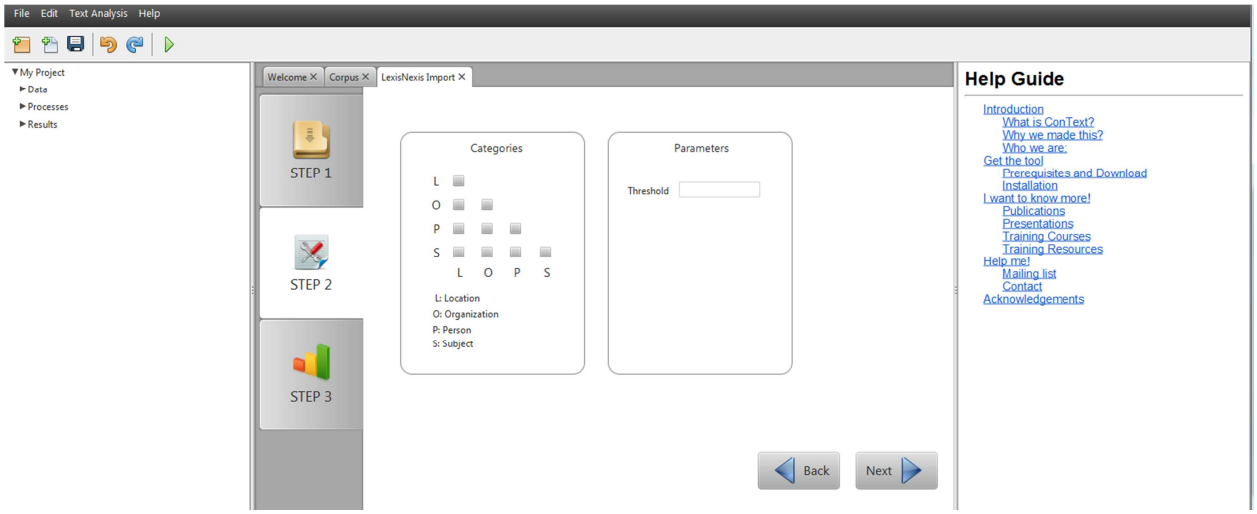


Figure 3 gives an example for one such task: After previously downloaded data from LexisNexis[1] are automatically organized into a) a curated dataset of articles and b) a database of meta-data (step 1 in the GUI below), the user can construct various types of one-more or multi-mode networks from the meta-data (step 2). The various text analysis techniques supported by ConText (listed below) can then be used to analyze the substance of the underlying bodies of the articles, and the meta-data networks can be enhanced or fused with the outcome of these processes (step 3).

Figure 3: Construction of Meta-Data Networks



In the following, we give a brief overview on the main routines for data import, curation, pre-processing and analysis provided in ConText.

1. Data import:
- Publicly available data from social media, currently Facebook and Twitter. In the near future, we will also support the import of data from YouTube.
- Client-side text data sets (Figure 2).

2. Data curation and pre-processing:
- Structured text data and meta-data:
    - Conversion of downloaded batches of data from LexisNexis (http://www.lexisnexis.com/) into text corpora readily available for further analysis by (Figure 3):
        - Splitting up and disambiguating data into individual text data files with and without (up to user) pertinent meta-data.
        - Organizing pertinent meta-data in database.
- Unstructured natural language text data:
    - Creation and application of stop word lists

[1] ConText does not support data collection from LexisNexis, only the curation and analysis of data already acquired from LexisNexis.

    - Stemming
    - Parts of speech tagging

3. Analysis of unstructured, natural language text data:

- Summarization techniques:
    - Corpus statistics, including absolute and weighted term frequencies
    - Topic modeling
    - Sentiment Analysis
    - Integrated visualization of topic modeling and sentiment analysis
- Entity and Relation extraction techniques:
    - Entity Detection
    - Construction and application of codebooks, thesauri and dictionaries
    - Relation extraction based on semantics, syntax, co-occurence and meta-data
    - Construction of one-mode networks (association networks, semantic networks) and multi-mode networks

Besides providing ConText for free, we are also using this tool for conducting data collection and analysis tasks in our lab, currently mainly for studying the impact of social justice documentaries and media on society (J. Diesner, Aleyasen, Kim, Mishra, & Soltani, 2013; J Diesner, Pak, Kim, Soltani, & Aleyasen, 2014).

Acknowledgement: This work is supported by the FORD Foundation, JustFilms division, grant No. 0125-6162. The author is grateful to the graduate student team from UIUC who works on this tool: Amirhossein Aleyasen, Jinseok Kim, Shubhanshu Mishra, Kiumars Soltani and Sean Wilner. Further gratitude goes to Dr. Susie Pak, St John's University, Orlando Bagwell, former director of JustFilms, and Joaquin Alvarado, Chief Strategy Officer at the Center for Investigative Reporting.

References:

Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM, 55*(7), 60-70.

Chang, J., Boyd-Graber, J., & Blei, D. (2009). *Connections between the lines: augmenting social networks with text.* Paper presented at the 15th ACM SIGKDD International Conference, Paris, France.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Danowski, J. A. (1993). Network Analysis of Message Content. *Progress in Communication Sciences, 12*, 198-221.

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts.* Carnegie Mellon University. (CMU-ISR-12-101, PhD Thesis)

Diesner, J. (2013). From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. *Künstliche Intelligenz/ Artificial Intelligence, 27*(1), 75-78. doi: 10.1007/s13218-012-0225-0

Diesner, J., Aleyasen, A., Kim, J., Mishra, S., & Soltani, S. (2013). *Using Socio-Semantic Network Analysis for Assessing the Impact of Documentaries.* Paper presented at the WIN (Workshop on Information in Networks), New York, NY.

Diesner, J., & Carley, K. M. (2011a). Semantic Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 766-769): Sage.

Diesner, J., & Carley, K. M. (2011b). Words and Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 958-961): Sage.

Diesner, J., Pak, S., Kim, J., Soltani, K., & Aleyasen, A. (2014). *Computational Assessment of the Impact of Social Justice Documentaries* Paper presented at the iConference, Berlin, Gemany.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211-244.

Gruzd, A. (2009). *Automated discovery of social networks in text-based online communities.* Paper presented at the ACM International Conference on Supporting Group Work.

Kirchner, C., & Mohr, J. W. (2010). Meanings and relations: An introduction to the study of language, discourse and networks. *Poetics, 38*(6), 555-566.

McCallum, A., Wang, X., & Mohanty, N. (2007). Joint Group and Topic Discovery from Relations and Text *Statistical Network Analysis: Models, Issues, and New Directions. Lecture Notes in Computer Science 4503* (pp. 28-44).

Mihalcea, R. F., & Radev, D. R. (2011). *Graph-based Natural Language Processing and Information Retrieval*: Cambridge University Press.

Roth, C., & Cointet, J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16-29.

Sowa, J. (1992). Semantic Networks. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2nd ed., pp. 1493 - 1511). New York, NY, USA: Wiley and Sons.