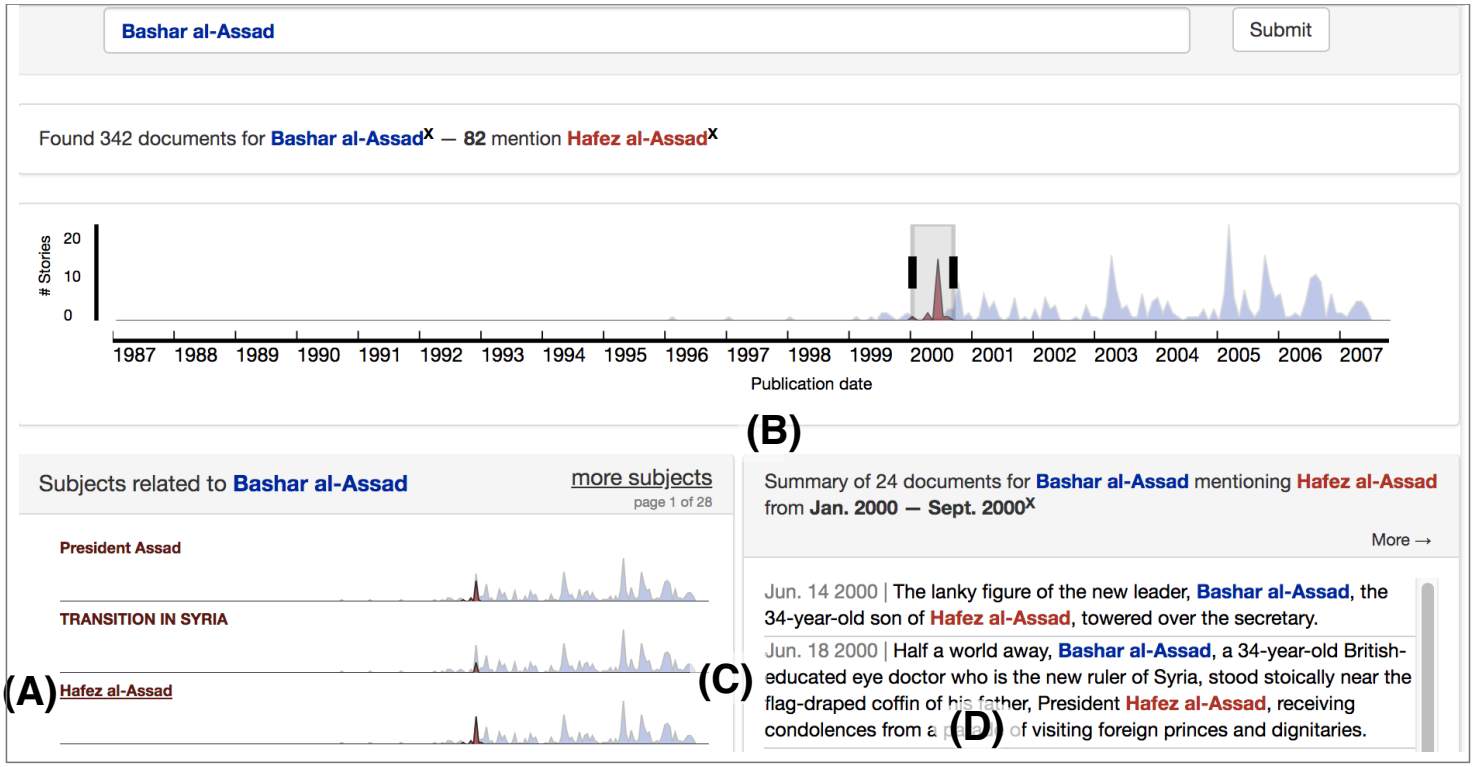

Figure 1: An analyst investigates the query *Bashar al-Assad* in a news corpus using our ROOKIE interface for lexical corpus analysis

Many text analysis problems begin with a large and unfamiliar corpus: a journalist is leaked a trove of documents, a historian tries to build a narrative with the congressional record, a marketer examines feedback forms after a sudden drop in sales. In such settings, practitioners cannot read all available evidence. They must investigate by searching, browsing and reading selections from a body of text.

To assist in this process, we propose lexical corpus analysis: in which analysts formulate, refine and answer qualitative research questions by investigating entities and concepts from a corpus. We present a collection of natural language processing methods for our proposed analytic technique.

For instance, Figure 1 presents ROOKIE, a particular tool for lexical corpus analysis. In the figure, an analyst investigates the roots of the Syrian civil war by querying for *Bashar al-Assad* in a collection of articles from the *New York Times*. Our ROOKIE system then **(A)** identifies an association between *Bashar al-Assad* and the related term *Hafez al-Assad* and **(B)** selects and displays sentences to summarize the relationship between this pair.

Unlike well-studied corpus analysis tools such as document search engines or topic models, our proposed approach relies on answers to fundamental language technology questions, which are poorly understood. What lexical items should be shown to an analyst? Could a computer summarize relationships between lexical items, to help make sense of a corpus? How might an automated system simplify sentences shown to an analyst in order to support rapid browsing?

To help answer, this work presents the following:

- We describe the ROOKIE interface [DS + J @KDD (2017)], and demonstrate that the system helps real users answer research questions from a historical archive.
- We define and begin work on the task of relationship summarization [NAACL (2018), NEWSUM @EMNLP (2019)], in which the goal is to summarize the relationship between a pair of terms in a lexicon, e.g. what is the relationship between *Bashar al-Assad* and *Hafez al-Assad* **(B)**?
- We describe a technique for shortening sentences which contain query terms [EMNLP (2019)], such as the lengthy sentence describing Bashar al-Assad and Hafez al-Assad **(C)**. We also offer a companion dataset which examines human perceptions of shortened sentences, informed by psycholinguistic studies of well-formedness.
- We present NP-FST [NLP + CSS @EMNLP (2017)], which efficiently extracts important multiword terms for a corpus lexicon, such as *Hafez al-Assad* **(A)**. We also present companion software `phrasemachine`, which enjoys wide use. (`https://github.com/slanglab/phrasemachine`)
- We propose new approaches to identifying different lexical items that refer to the same underlying entity [WNUT Abstract @EMNLP (2019)]. For instance, in **(D)**, the spans "Hafez al-Assad" and "his father" refer to the same person. Identifying such coreferent spans provides a more complete picture of the role of an underlying entity within a corpus.

Together, these methods support our proposed goal of lexical corpus analysis: a reliable, interpretable and trustworthy research method, that aims to support real users investigating diverse questions across different domains.