

# Towards finite-state analysis of lexical cohesion

**Branimir K. Boguraev**

*IBM T.J. Watson Research Centre, P.O. Box 704, Yorktown Heights, NY 10598*

bkb@watson.ibm.com

**Abstract** In recent work, we have outlined a programme for enhancing a range of document content analysis tasks by means of exploiting lexical cohesion factors in the text. This paper describes a shallow parsing component, fully realised as an INTEX cascade of finite-state transducers, which jointly act as a syntactic analysis system for the identification of phrasal, configurational, and grammatical information in free text documents. The cascade functions as a subsystem, in a large scale environment for document content analysis and characterisation, providing the base level data for lexical cohesion analysis.

In principle, such analysis requires detailed morpho-syntactic information, fine-grained identification of phrasal constituents, configurational information (typically established with reference to a full syntactic parse), and at least some minimal determination of grammatical function. Instead of deploying a full syntactic parser, with its concomitant limitations and overheads, this work uses finite-state techniques for enabling fine-grained analysis of lexical cohesion. Ultimately, in contexts where real-time applications require topical highlights of documents, multi-threaded summaries for personal information delivery, or navigating through multiple-document spaces, INTEX offers highly scalable functionalities enabling deeper semantic analysis.