Dorota Smyk-Bhattacharjee
University of Zurich, Switzerland
dsmyk@es.unizh.ch

Bolek Umnicki
bolekum@gmail.com

INDIANA: A system for identifying lexical innovations

Neologisms have attracted increased linguistic attention recently (e.g. Plag 1999, Bauer and Renouf 2001). Within this context we want to observe lexical change in computer-mediated communication in English. For this purpose we have compiled a corpus of English language web-logs or blogs (i.e. personal diaries published on the Internet). The corpus currently amounts to one million words of running text, covering the period from 2000-2004. For such a small corpus the existing automated neologism selection methods are not applicable, and therefore we are proposing a new range of tools.

A software tool named INDIANA (INternet DIctionary ANAlyser) has been developed to extract neologisms from webfiles. It is a combination of a cumulative database and a series of online and offline filters. At any given moment the database consists of all the words that have already been processed. Each new input file of text is converted into a sorted list of words. INDIANA first checks every word from the list against the existing database. If the word is not already present in the database, it is checked against two external reference sources; the data of the British National Corpus (BNC) and the Webster online dictionary. If a match is found in one of these reference sources this information is added to the database, otherwise the word is marked as a potential neologism.

INDIANA also includes a variety of filters which enable not only a quick extraction of potentially new words but also offer information on type/token frequency, distribution across the input files, and easy view of the word in context. Additionally distribution information of every text file, grouped as information about the author, text, linguistic features etc. is encoded. This enables detailed cross-analysis using any combination of filters.

After a brief description of the corpus of web-logs, this presentation will introduce and describe the tools and methods used in the analysis.

References

Bauer, Laurie and Antoinette Renouf. 2001. A corpus-based study of compounding in English. *Journal of English Linguistics* 29: 101-123.

Plag, Ingo. 1999. *Morphological Productivity. Structural Constraints in English Derivation.* Berlin/New York: Mouton de Gruyter.