# Language-independent Methods for Compiling Monolingual Lexical Data

Christian Biemann,[1] Stefan Bordag,[1] Gerhard Heyer,[1]
Uwe Quasthoff,[1] Christian Wolff [2]

[1]Leipzig University
Computer Science Institute, NLP Dept.
Augustusplatz 10/11
04109 Leipzig, Germany

{biem, sbordag, heyer, quasthoff}@informatik.uni-leipzig.de
[2]University of Regensburg
PT 3.3.48
93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

**Abstract:** In this paper we describe a flexible, portable and language-independent infrastructure for setting up large monolingual language corpora. The approach is based on collecting a large amount of monolingual text from various sources. The input data is processed on the basis of a sentence-based text segmentation algorithm. We describe the entry structure of the corpus database as well as various query types and tools for information extraction. Among them, the extraction and usage of sentence-based word collocations is discussed in detail. Finally we give an overview of different applications for this language resource. A WWW interface allows for public access to most of the data and information extraction tools (*http://wortschatz.uni-leipzig.de*).

## 1 Introduction

We describe an infrastructure for managing large monolingual language resources. Several language independent methods are used to detect semantic relations between the words of a language. These methods differ in productivity and precision for different languages, but there are highly productive and accurate methods for all languages tested. The process starts with the collection of monolingual text corpora from the Web. Next, we identify collocations, i.e. words that occur significantly often together. These collocations form a network that is analyzed further to identify semantic relations. Because semantic features are often reflected in morphosyntactic structures, we apply classifiers that use the sequence of its characters for classification. Moreover, we use context information and POS-information, if available.

While it is clear that the above mentioned methods can be used to find semantic relations or, can be used to verify the corresponding hypotheses, we want to present abstract methods, specific application and results for different languages.