wortschatz lexikon

UNIVERSITÄT LEIPZIG

http://wortschatz.uni-leipzig.de

Linguistic Corpus Search

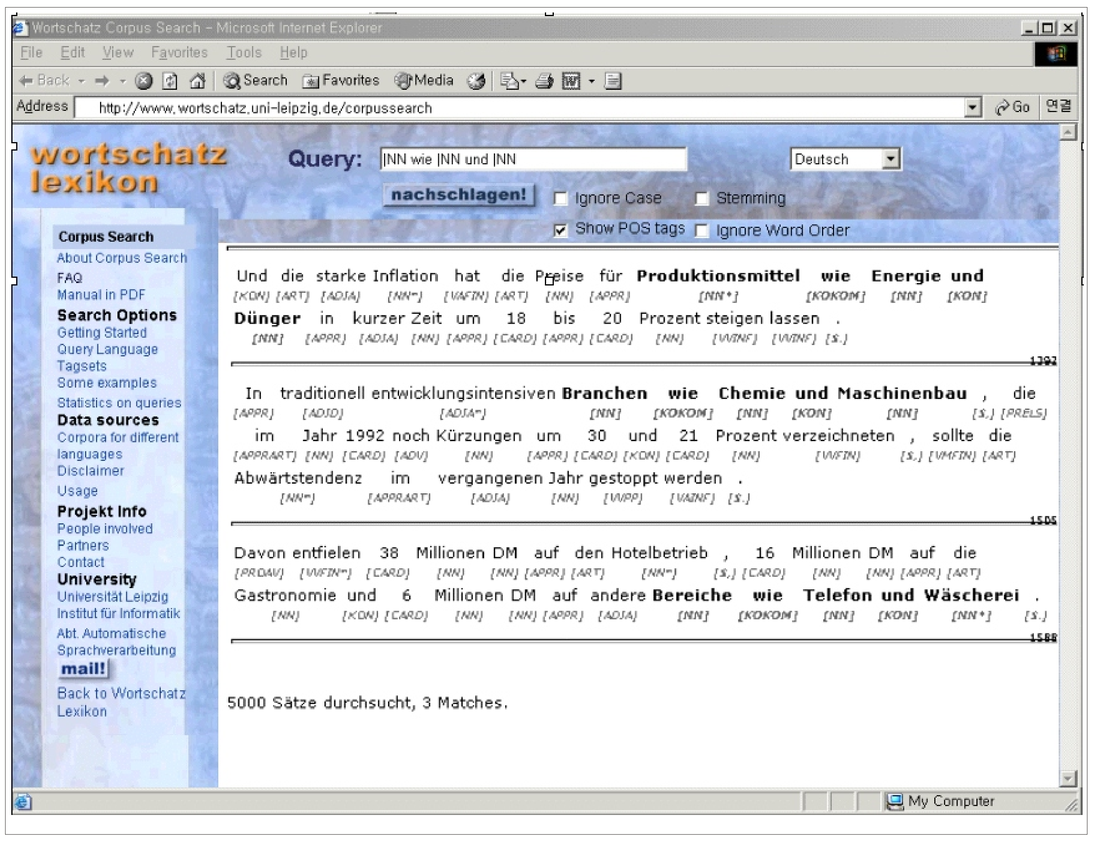
Chris Biemann, Uwe Quasthoff, Christian Wolff

Corpus Search
Simple but powerful

Using Wildcards: For *"ein* * vom Zaun brechen"*, we will find phrases like *"einen Streit vom Zaun brechen"*, *"einen Krieg vom Zaun brechen"* and so on.

Using POS-Tags: *"house*|NN*"* will match both *"house|NNL1"* and *"houses|NNL2"*, but not *"houses|VVZ the|AT camp|NNL1"*

Abbreviations: *"of|*"* can be abbreviated to *"of"*, *"*|VVAD"* to *"VVAD"* and *"*|*"* to *"*"*.



Linguistic Features

POS-Tagging: Done by TNT for English and German. More languages to come.

Base Form Reduction: Allows search for inflected forms: *_house* finds *house* and *houses*, *_Haus* finds *Haus, Hauses, Häuser* and *Häusern*

Free Word Order: *Jung und Alt* also finds *Alt und Jung*, *boys and girls* also finds *girls and boys*

Index Structure and Ranking

4+ -gram-Index: We index all n-grams for n 4 at the end of the word. For *Portugal,* we get *ortugal, rtugal, tugal* and *ugal*. This allows effective search for words with patterns like **rtug**

Ranking: Preferred are
- sentences containing additional collocations as typical objects,
- sentences containing the search patterns in the given order (in the case of variable word order), and
- sentences not containing subordinate clause separators such as ",", ";" and "-".

Resources and Corpora

Language	Newspaper	Web Text
English	10 M sentences	100 M sentences
German	30 M sentences	100 M sentences
Italian	3 M sentences	
Spanish	1 M sentences	
French	3 M sentences	
Swedish		10 M sentences
Norwegian		3 M sentences
Finnish		3 M sentences
Danish		3 M sentences
Korean	3 M sentences	

Tagger available for English and German
More languages in preperation

Applications

Phrases with variable Parts: Pos-Tags for variable Parts: *"to be in a tight |NN"* gives *"to be in a tight corner"*, *"to be in a tight squeeze"* and *"to be in a tight spot"*.

Terminology: Search for NPs described by words and POS-Tags. Example: *calculation of |ADJ eigenvalues* gives *calculation of simple eigenvalues*

Knowledge Extraction: Patterns of POS-Tags for hyponymy and co-hyponymy:
"|ADJ |N like |N and |N" for *"environmental problems like noise and exhaust"*

Institute of Computer Science, NLP Dept., Leipzig University, 04109 Leipzig, Germany
{biem, quasthoff, wolff}@informatik.uni-leipzig.de, WWW: http://wortschatz.uni-leipzig.de/