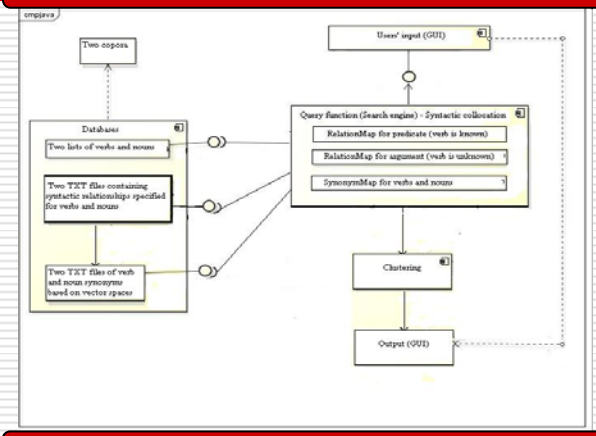
# WORD ASSISTANT

**Zheng Ye, Philipp Busch, Pei Lu**
**Department of Computational Linguistics, University of Heidelberg, Germany**

## Introduction and Project Architecture

**Problem**: Due to limited vocabulary, non-native writers are often unsure about specific expressions.
**Goal**: The purpose of this software project is build a writing support system as one of possible solutions to the problem.
**Architecture**: The system consists of three modules, three databases, the query function and the GUI interface.

## Process

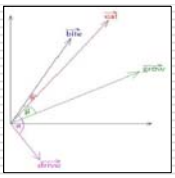



## Approaches

To achieve our goal, three special approaches are employed.

1. Three databases are extracted from the UK Web Archiving Consortium (UKWAC) and the British National Corpus (BNC). The UKWaC is a wide-ranged corpus covering diverse domains in English, while the BNC is manually annotated which avoids finding the wrong words.

2. The third database of synonyms is built with the help of two dependency vectors which are generated by DEPENDENCYVECTORS 2.5 (Pado 2010).



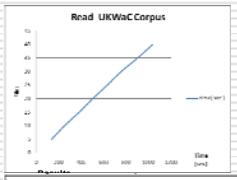
$$sim_{cos}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

3. A search engine is set up on the basis of two hash maps, i.e. RelationMap and SynonymMap, and uses frequency of occurrence of word combination to filter words.

4. The GUI is built as an interface to provide the user a direct access to Word Assistant with a dialog-box.

## Evaluation

As it takes much time to read all files from the UKWaC and BNC, the software was tested with several mini examples. Herein is one of those examples.




**Read resource**
relationfre_noun.txt
relationfre_verb.txt
verbsynomy.txt
nounsynomy.txt
verb_bnc.txt (same to the file in the reaource)
noun_bnc.txt (same to the file in the reaource)

**relationfre_noun.txt**
dog love obj 7
cat love obj 9
dog love subj 5
i love subj 10
you love obj 12
i want subj 9
book want obj 14
fun want obj 12

| | Example | result |
|---|---|---|
| The user should type a "#" in the position of the missing word. | i # you<br>Input: verb | love<br>want |
| | i love #<br>Input: noun | cat<br>you<br>dog |
| | # love dog<br>Input: noun | I |
| The user should type a "?" before the word for which the program should find the synonyms. | love ?you<br>Input: noun | him |
| | i ?love you<br>Input : verb | miss<br>think<br>want |
| | i love ?dog<br>Input: noun | cat |

**relationfre_verb.txt**
love you obj 11
love i subj 9
love dog subj 10
love cat obj 16
want fun obj 12
want book obj 13
want i subj 10

**verbsynomy.txt**
love miss 0.632455532033675 9
love think 0.738394847748393
love want 0.839473739273273

**nounsynomy.txt**
dog cat 0.997031700969424
you him 0.620712720134246

## Conclusion

This software has the potential to be expanded despite its limitation on verbs and nouns. The experiment with verbs and nouns serves as a basis for future research. The obvious disadvantage is that it takes much time to read the files from the corpora due to their vast size.

## References

A Ferraresi, E Zanchetta,M Baroni, S Bernardini (2008). *Introducing and evaluating ukWaC, a very large web-derived corpus of Englis*h. http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukwac.pdf (Consulted on 15 January 2011).

S Pado, U Pado and K Erk (2010). A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics* 36: 4.

S Pado (2010). *Manual for DependencyVectors 2.5*. http://www.nlpado.de/~sebastian/dv/dv_manual_2.5.pdf (Consulted on 20 December 2010).

philippbusch@gmx.net; pamela_yz@yahoo.de; lupei.hd@gmail.com