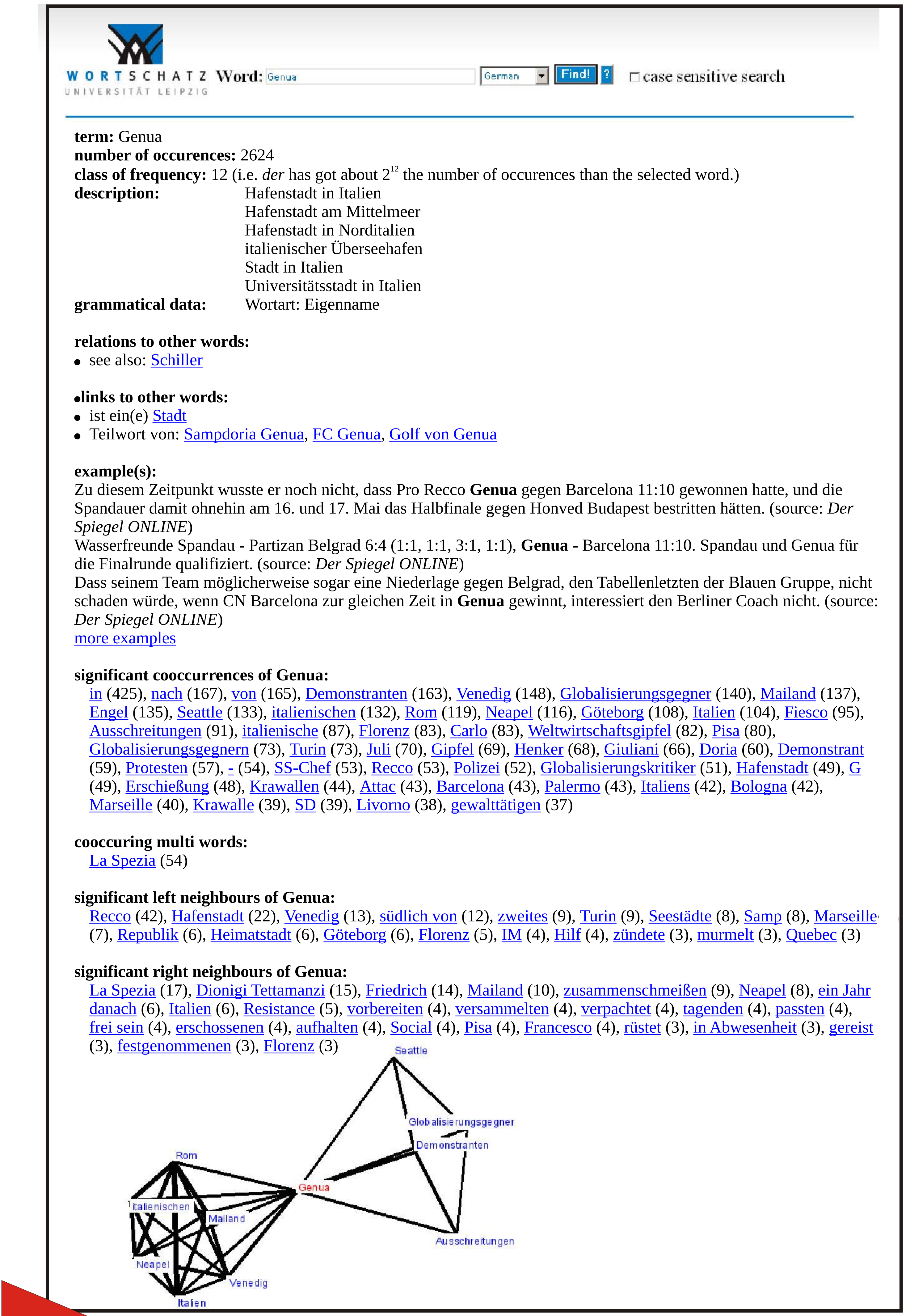
Corpus Portal for Search in Monolingual Corpora

Uwe Quasthoff, Matthias Richter, Chris Biemann

NLP Group, University of Leipzig - http://corpora.uni-leipzig.de



WORTSCHATZ UNIVERSITÄT LEIPZIG **Word:** Genua German Find! ? □ case sensitive search

term: Genua
number of occurences: 2624
class of frequency: 12 (i.e. *der* has got about 2^{12} the number of occurences than the selected word.)
description: Hafenstadt in Italien
Hafenstadt am Mittelmeer
Hafenstadt in Norditalien
italienischer Überseehafen
Stadt in Italien
Universitätsstadt in Italien
grammatical data: Wortart: Eigenname

relations to other words:
- see also: Schiller

•links to other words:
- ist ein(e) Stadt
- Teilwort von: Sampdoria Genua, FC Genua, Golf von Genua

example(s):
Zu diesem Zeitpunkt wusste er noch nicht, dass Pro Recco **Genua** gegen Barcelona 11:10 gewonnen hatte, und die Spandauer damit ohnehin am 16. und 17. Mai das Halbfinale gegen Honved Budapest bestritten hätten. (source: *Der Spiegel ONLINE*)
Wasserfreunde Spandau **-** Partizan Belgrad 6:4 (1:1, 1:1, 3:1, 1:1), **Genua -** Barcelona 11:10. Spandau und Genua für die Finalrunde qualifiziert. (source: *Der Spiegel ONLINE*)
Dass seinem Team möglicherweise sogar eine Niederlage gegen Belgrad, den Tabellenletzten der Blauen Gruppe, nicht schaden würde, wenn CN Barcelona zur gleichen Zeit in **Genua** gewinnt, interessiert den Berliner Coach nicht. (source: *Der Spiegel ONLINE*)
more examples

significant cooccurrences of Genua:
in (425), nach (167), von (165), Demonstranten (163), Venedig (148), Globalisierungsgegner (140), Mailand (137), Engel (135), Seattle (133), italienischen (132), Rom (119), Neapel (116), Göteborg (108), Italien (104), Fiesco (95), Ausschreitungen (91), italienische (87), Florenz (83), Carlo (83), Weltwirtschaftsgipfel (82), Pisa (80), Globalisierungsgegnern (73), Turin (73), Juli (70), Gipfel (69), Henker (68), Giuliani (66), Doria (60), Demonstrant (59), Protesten (57), - (54), SS-Chef (53), Recco (53), Polizei (52), Globalisierungskritiker (51), Hafenstadt (49), G (49), Erschießung (48), Krawallen (44), Attac (43), Barcelona (43), Palermo (43), Italiens (42), Bologna (42), Marseille (40), Krawalle (39), SD (39), Livorno (38), gewalttätigen (37)

cooccuring multi words:
La Spezia (54)

significant left neighbours of Genua:
Recco (42), Hafenstadt (22), Venedig (13), südlich von (12), zweites (9), Turin (9), Seestädte (8), Samp (8), Marseille (7), Republik (6), Heimatstadt (6), Göteborg (6), Florenz (5), IM (4), Hilf (4), zündete (3), murmelt (3), Quebec (3)

significant right neighbours of Genua:
La Spezia (17), Dionigi Tettamanzi (15), Friedrich (14), Mailand (10), zusammenschmeißen (9), Neapel (8), ein Jahr danach (6), Italien (6), Resistance (5), vorbereiten (4), versammelten (4), verpachtet (4), tagenden (4), passten (4), frei sein (4), erschossenen (4), aufhalten (4), Social (4), Pisa (4), Francesco (4), rüstet (3), in Abwesenheit (3), gereist (3), festgenommenen (3), Florenz (3)

Corpora and Sizes

Language	size(sentences)	origin
Catalan	10 M	web
Danish	3 M	web
Dutch	1 M	newspapers
English	10 M	newspapers
Estonian	1 M	newspapers
Finnish	3 M	web
French	3 M	newspapers
German	30 M	newspapers
Icelandic	1 M	newspapers
Italian	3 M	newspapers
Japanese	300 k	web
Korean	1 M	newspapers
Norwegian	3 M	web
Sorbian	300 k	newspapers
Spanish	1 M	web
Swedish	3 M	web
Turkish	1M	web

Features

- uniform format and web interface for all corpora
- comparable data sets for different languages
- corpora in several pre-defined sizes
- statistical information including co-occurrence data
- standardized visualization of co-occurrence data
- access to a collection of linguistic resources free of charge
- seamless integration into applications via SOAP-based Web Services

Application Areas

- preparing monolingual dictionaries
- researching linguistic questions
- comparing different languages on a statistical basis
- parameterizing language models e.g. for speech recognition
- expanding queries with statistically similar words
- extracting significant terms from documents by comparison against a reference corpus
- selecting balanced word sets for experiments, e.g. in psycholinguistics

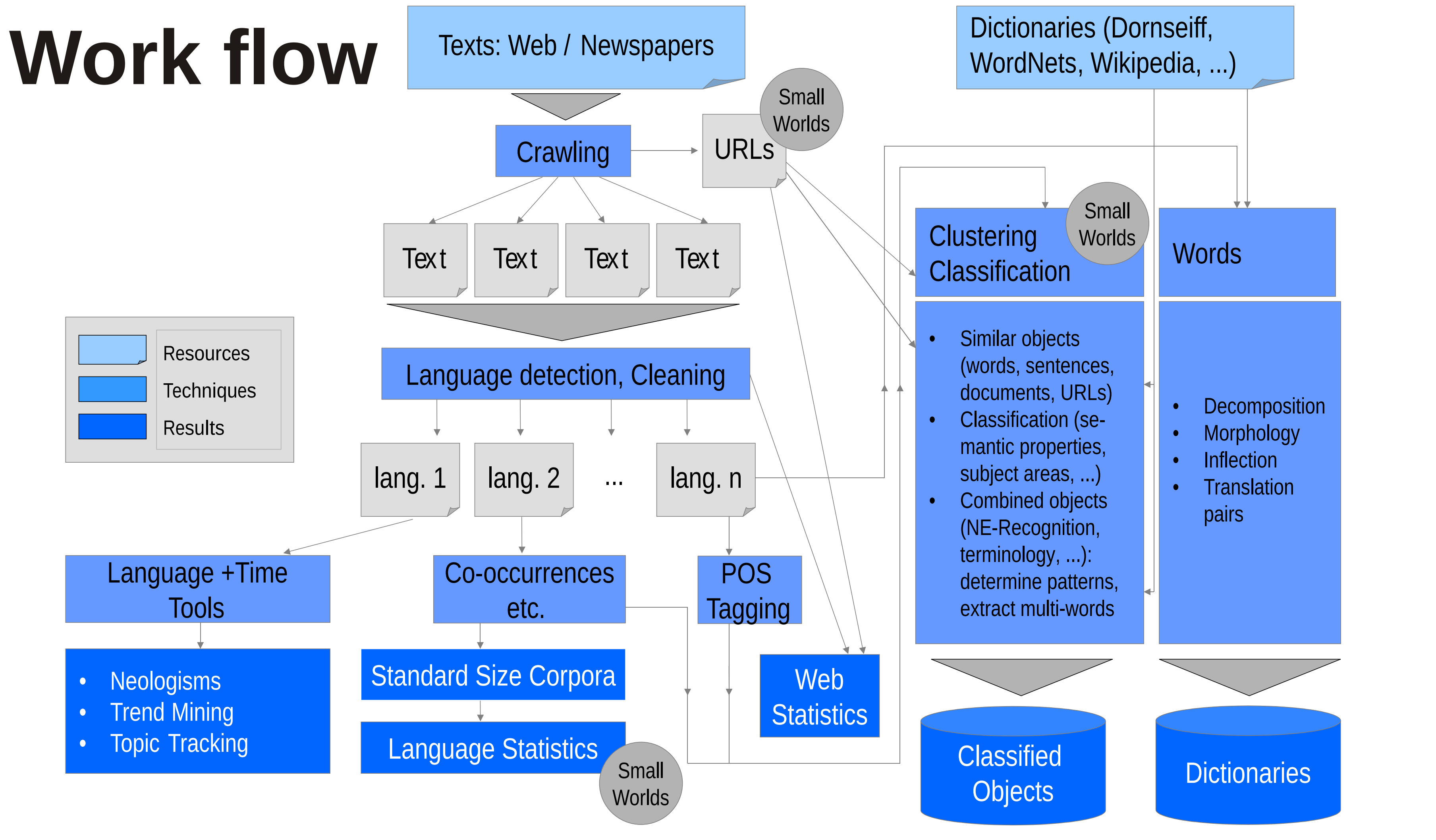
Standard size corpora
in 15 languages:
Catalan, Danish, Dutch,
English, Estonian, Finnish,
French, German, Italian,
Japanese, Korean, Norwegian,
Sorbian, Swedish, Turkish

For Linux®, MacOS X®,
and MS Windows®

LEIPZIG CORPORA COLLECTION
http://corpora.uni-leipzig.de
Version 1.0

UNIVERSITÄT LEIPZIG

Work flow



Texts: Web / Newspapers

Crawling

URLs

Small Worlds

Text
Text
Text
Text

Language detection, Cleaning

lang. 1
lang. 2
...
lang. n

Language +Time Tools

- Neologisms
- Trend Mining
- Topic Tracking

Co-occurrences etc.

Standard Size Corpora

Language Statistics

Small Worlds

POS Tagging

Web Statistics

Dictionaries (Dornseiff, WordNets, Wikipedia, ...)

Clustering Classification

Small Worlds

- Similar objects (words, sentences, documents, URLs)
- Classification (semantic properties, subject areas, ...)
- Combined objects (NE-Recognition, terminology, ...): determine patterns, extract multi-words

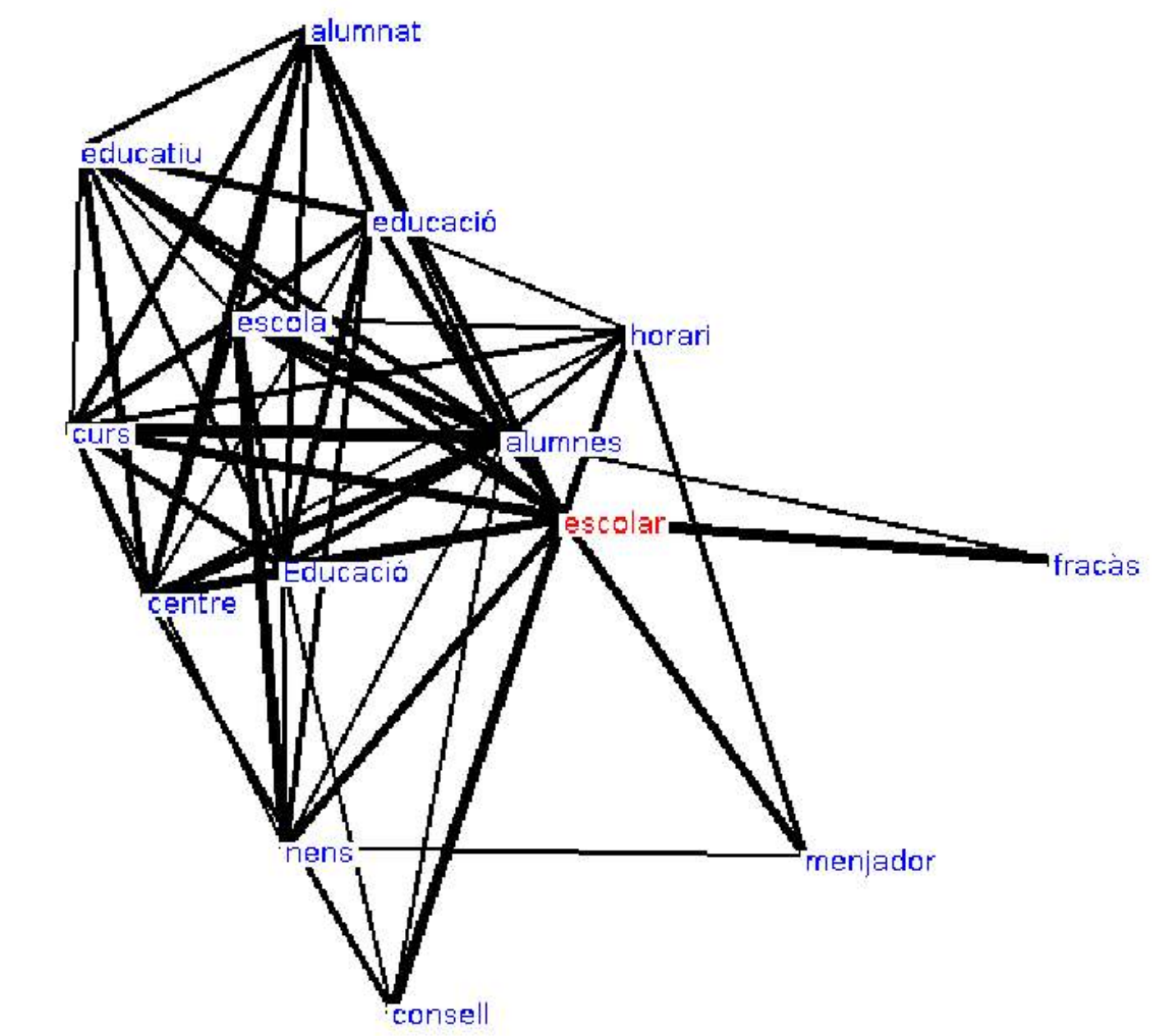
Classified Objects

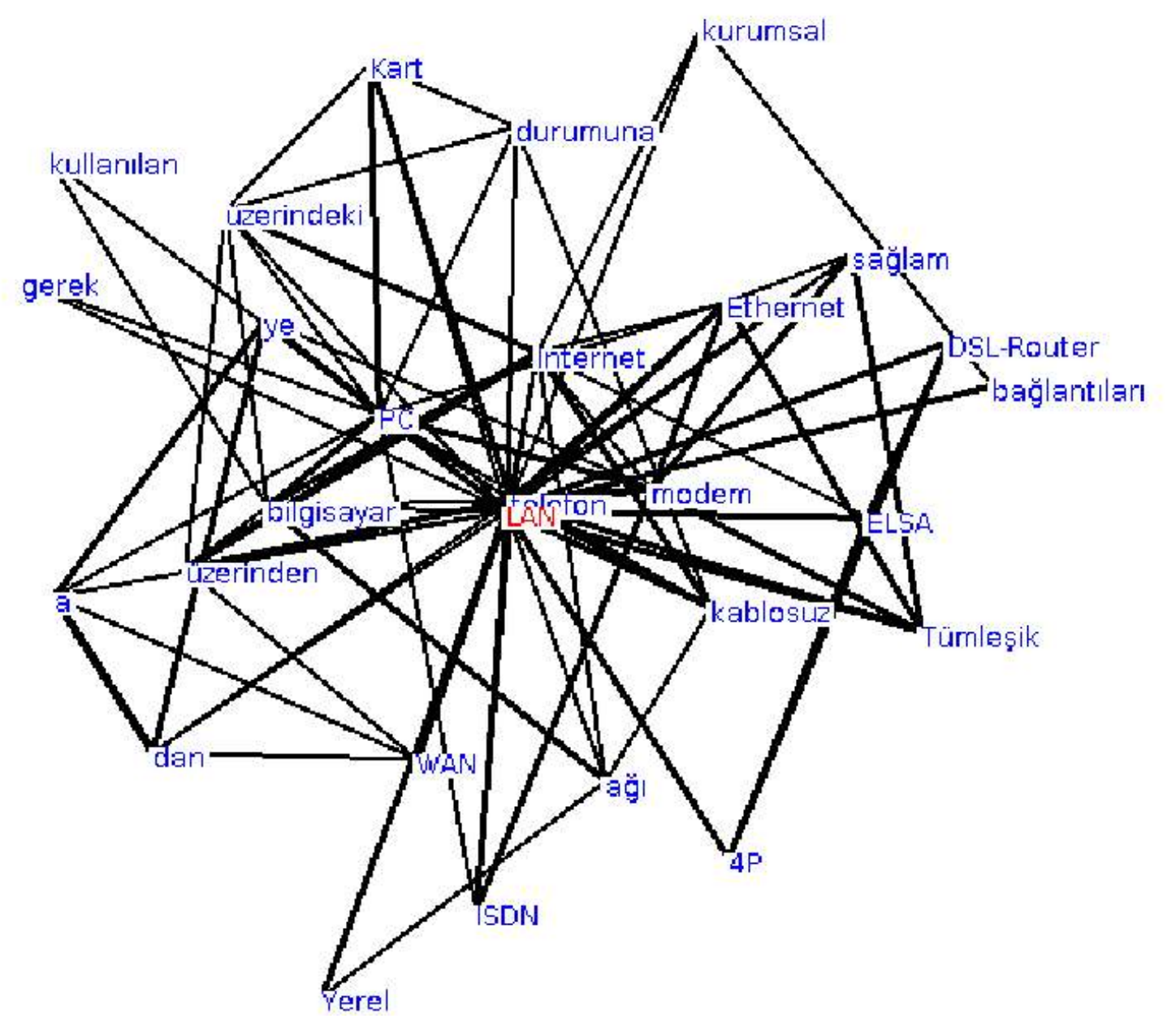
Words

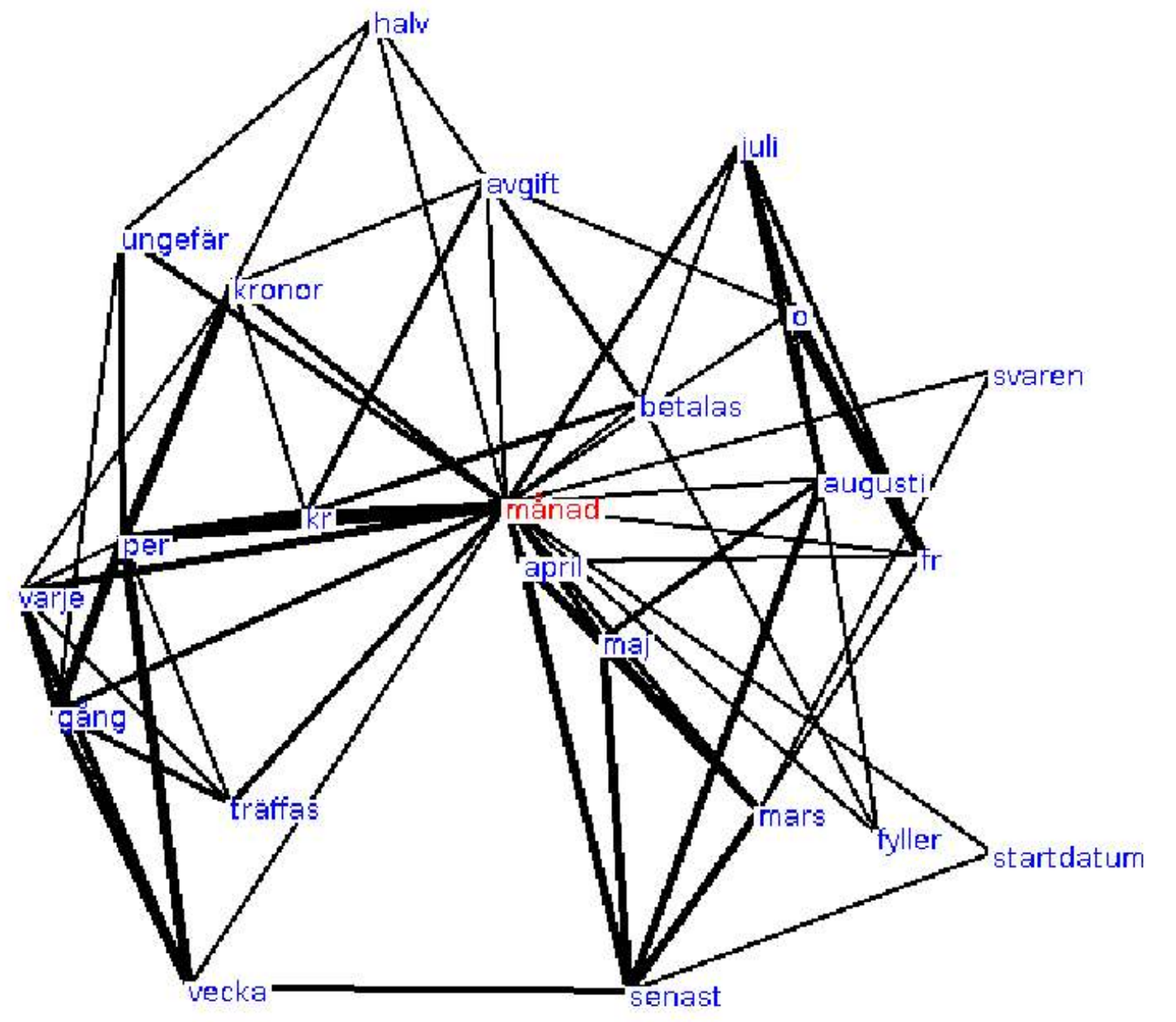
- Decomposition
- Morphology
- Inflection
- Translation pairs

Dictionaries

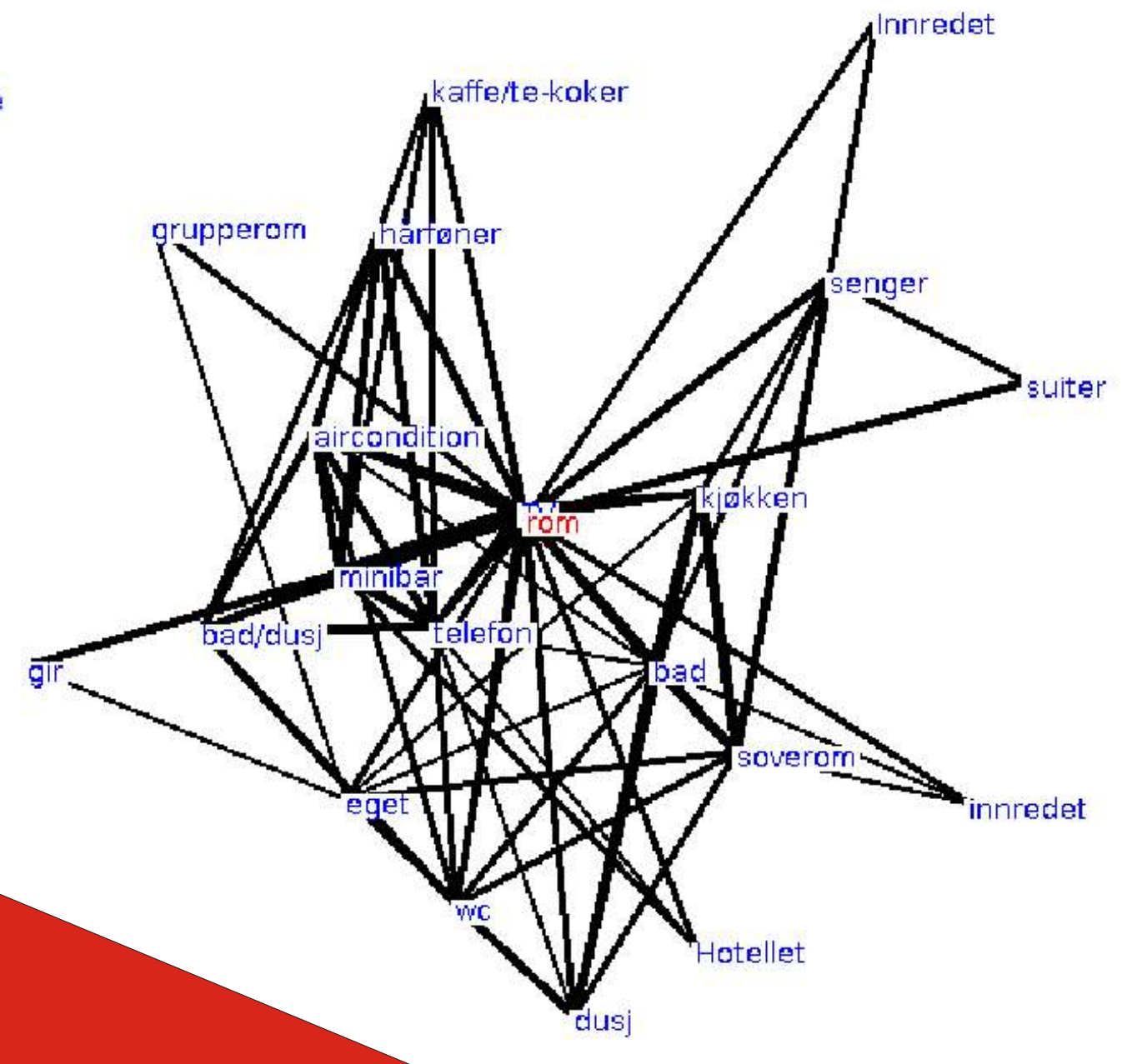
Resources
Techniques
Results

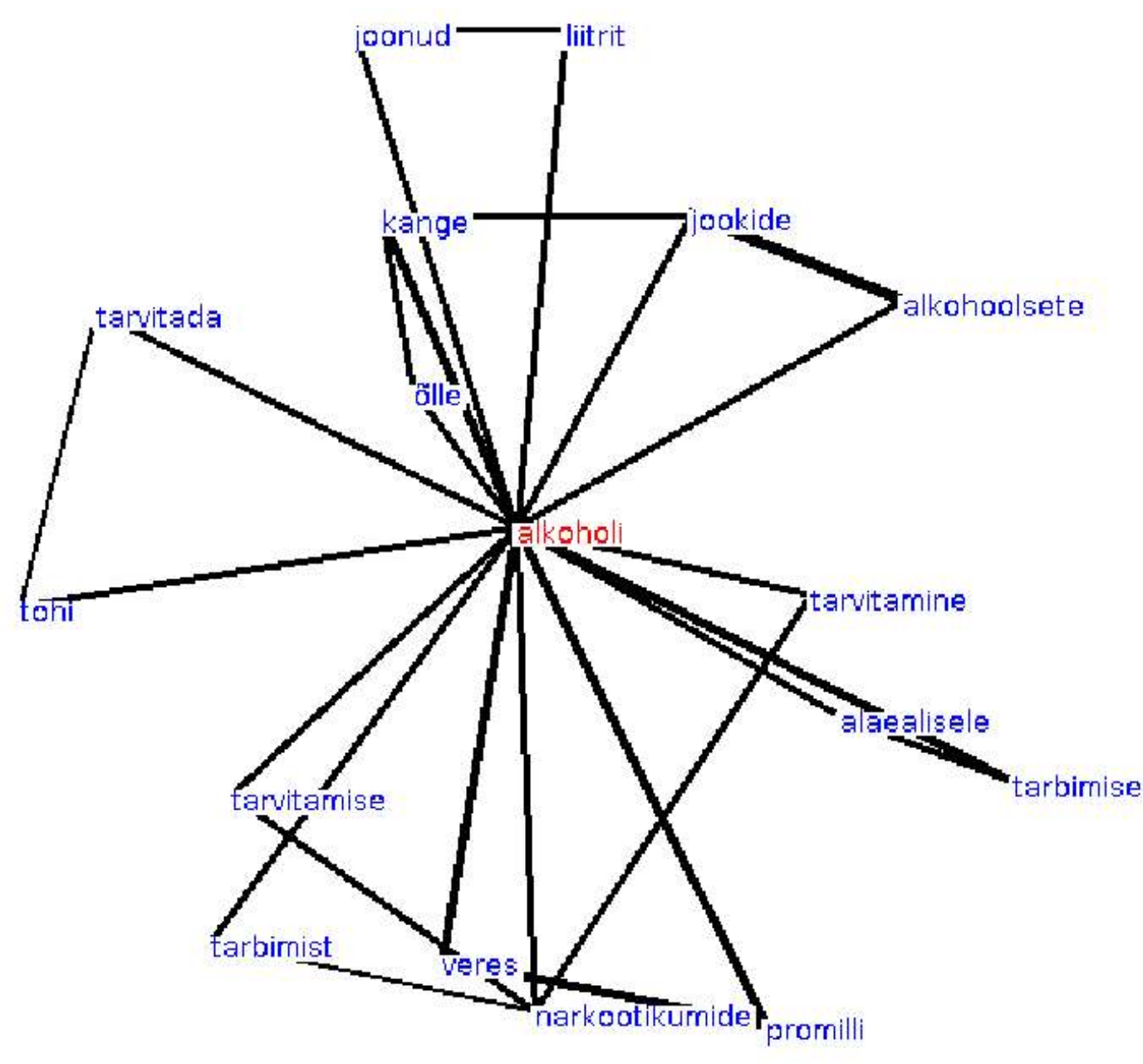


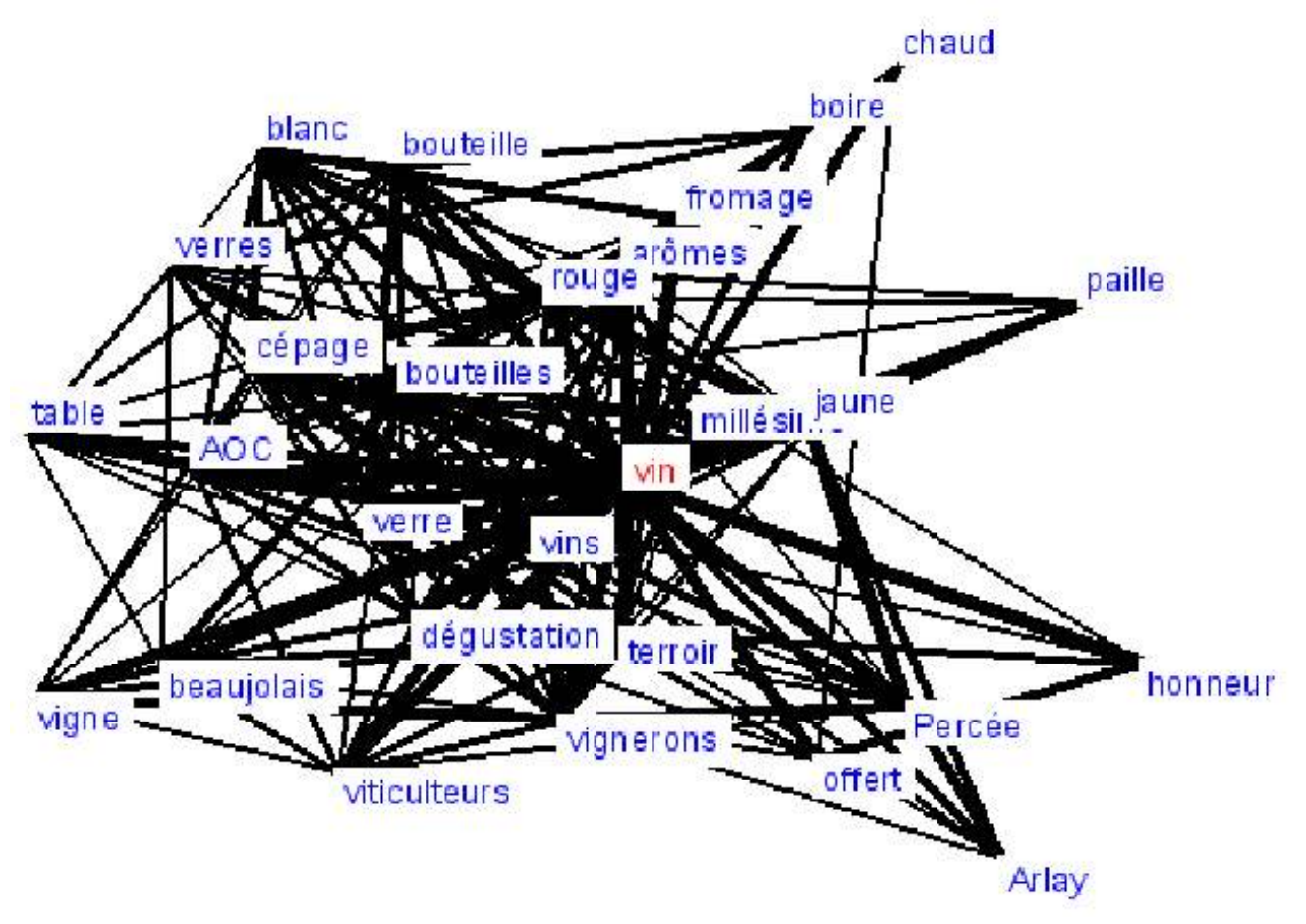


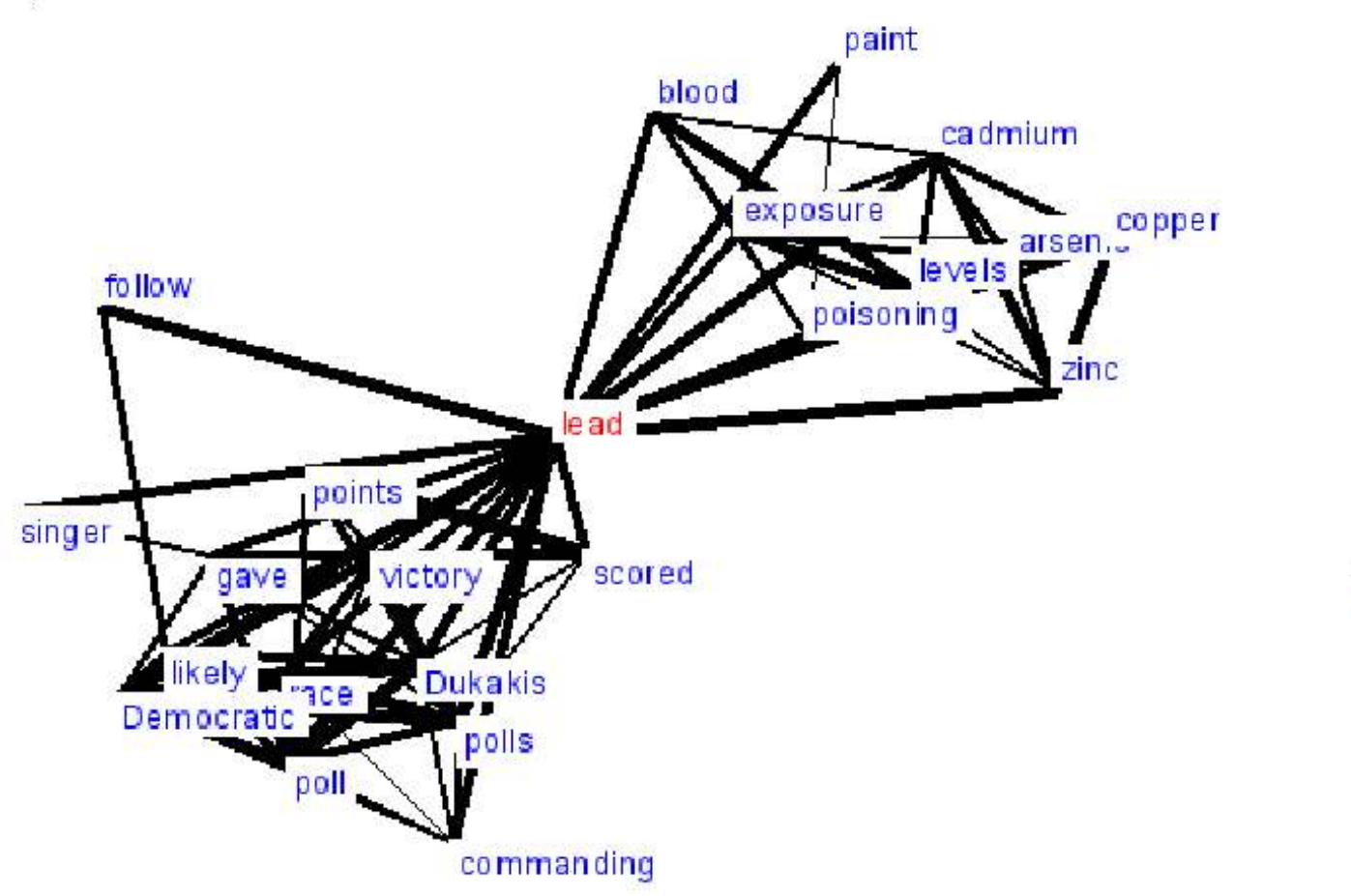


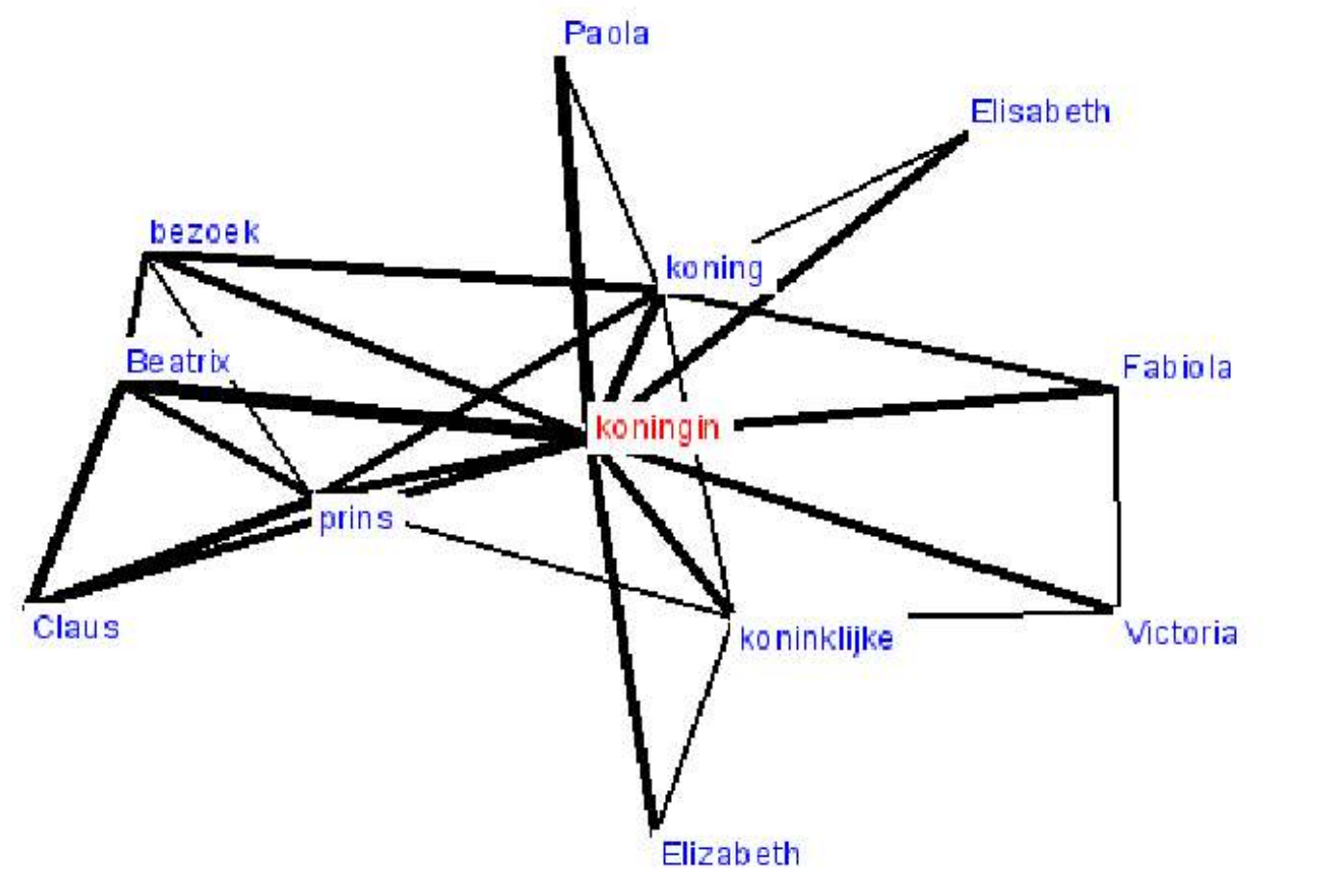
House-Senate
committee
sponsored
peace
meeting
conference
international
attend
samfund
demokratisk
politiske
demokratist
diktatur
menneskerettigheder
magt











Free DVD !

Printed by Universitätsrechenzentrum Leipzig