Korpus-Abfrage: Werkzeuge und Sprachen

Gastreferat zur Vorlesung
“Korpuslinguistik mit und für Computerlinguistik”

Charlotte Merz
3. Dezember 2002

Motivation

Lizentiatsarbeit:

A Corpus Query Tool for Automatically Annotated Corpora

- Corpus Query Tool
- Theoretical Part about Corpus Query

Overview

- Corpus Query Tools:
  - SARA
  - TIGERSearch
- Theoretical Considerations:
  - Parameters of Corpus Query
  - Corpus Query Languages
- My Own Corpus Query Tool

“Languages” of Corpus Query

- Scripting languages (perl, tgrep, etc.):
  - Not very intuitive or easy to use
- Corpus Query Languages
  - Formal construct designed to retrieve data from corpora
  - Emphasis on linguistic information (trade-off between linguistic correctness and performance)
- SQL
  - For database queries only

Corpus Query Tools: SARA

- SARA:
  “**S**GML-**A**ware **R**etrieval **A**pplication”
- Query Tool for British National Corpus (BNC: 100 Million words, PoS-tagged)
- Makes use of Corpus Query Language
- Graphical interface (“Query Builder”) as well as Corpus Query Language CQL

SARA Query Possibilities 1

- Word query
  - (e.g. ‘colour’ retrieves ‘colour’, ‘coloured’, ‘colouring’, etc.)
- Phrase query
  - ‘home _ centre’ retrieves ‘home loan centre’ or ‘home improvement center’
- Pattern query
  - ‘colo?r’ retrieves all instances of ‘color’ and ‘colour’

SARA Query Possibilities 2

- PoS-query
  - “colour”=NN1 retrieves all instances of ‘colour’ as a noun
  - “colour”=VVI retrieves all instances of ‘colour’ as infinitive
- SGML-query
  - ‘<body>’ retrieves all instances of the SGML-tag ‘<body>’

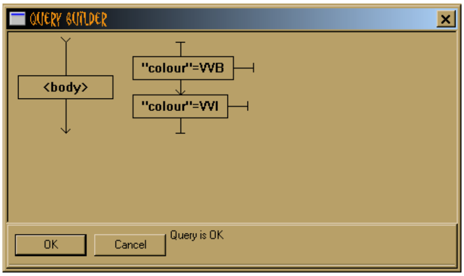
SARA Query Builder

Query Builder: visual interface to create complex queries

- Scope node (left)
  - e.g. search within the scope of a single SGML-element <body>
- Content node (right)
  - Find ‘colour’ in combination with PoS-tag ‘VVB’ or ‘VVI’

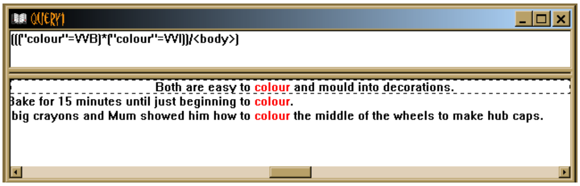
    (BNC Tagset: VVI is infinitive of lexical verb, VVB is base form of lexical verb, except infinitive)

SARA Query Builder



Query is OK

SARA Result Display



```
((("colour"=VVB)*("colour"=VVI))/<body>)
```

```
                   Both are easy to colour and mould into decorations.
Bake for 15 minutes until just beginning to colour.
big crayons and Mum showed him how to colour the middle of the wheels to make hub caps.
```

SARA CQL 1: Atomic Query

- Atomic query:
  - A word, punctuation mark, or delimited string (e.g. jam, ?, “Mrs.”)
  - A word-and-PoS pair (e.g. “CAN”=NN1)
  - A phrase (e.g. “not in your life”)
  - A pattern (e.g. colo?r)
  - An SGML query (e.g. <body>)
  - Wildcard character _ (e.g. home _ center)

SARA CQL 2: Unary Operators

- Unary operators:
  - Case: $ operator makes query case-sensitive
  - Header: @ operator makes query search within headers as well as bodies of texts
  - Not: ! Operator matches everything which is not a solution to the query
    (e.g. “!cat dog” finds occurrences of ‘cat’ not preceded by ‘cat’)

SARA CQL 3: Binary Operators

- Binary operators:
  - Sequence: blanks between two queries (e.g. cat dog)
  - Disjunction: operator | matches cases which satisfy either query (e.g. cat | dog)
  - Join: * (order matters) and # (order does not matter) operator match cases which satisfy both queries (e.g. cat * dog)

SARA Conclusion

- Disadvantages:
  - no syntactic mark-up in BNC
    -> retrieval options less comlex
  - no “delexicalized” search options for PoS
  - output functions restricted
- Advantages:
  - SGML search options
  - query builder
- **BNCWeb** refines BNC query

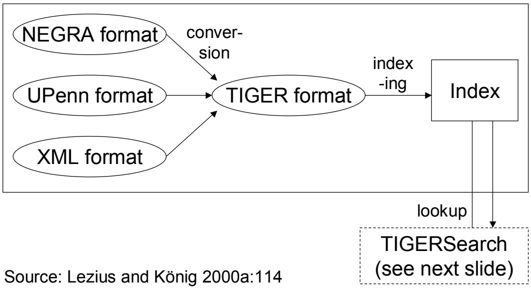
SARA

- Literature:
  - Burnard, Lou. 1996. “Introducing SARA: An SGML-Aware Retrieval Application for the British National Corpus” at http://www.hcu.ox.ac.uk/BNC/using/papers/burnard96a.htm
  - SARA handbook
- Internet Resources:
  - SARA trial version for 30 days at http://sara.natcorp.ox.ac.uk/
  - Simple Search online at http://sara.natcorp.ox.ac.uk/lookup.html

Corpus Query Tools: TIGERSearch

- Two-part system:
  TIGERRegistry and *TIGERSearch*
- TIGERRegistry:
  import and preprocessing of corpora
- TIGERSearch:
  querying, display and export of query results
- corpora:
  - NEGRA treebank (10’000 syntactically annotated sentences)
  - other corpora converted to TIGERXML-format

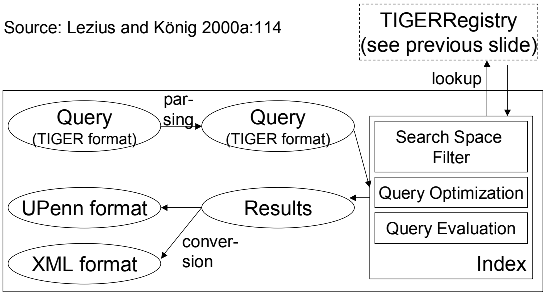
TIGERSearch Architecture TIGERRegistry



Source: Lezius and König 2000a:114

TIGERSearch Architecture TIGERSearch

Source: Lezius and König 2000a:114



TIGERSearch Description/Query Language 1

- TIGER Description Language serves two purposes:
  - to encode the syntactic annotation of the corpus
  - to define queries
- TIGER Description Language Levels:
  - node level
  - node relation level
  - graph description level

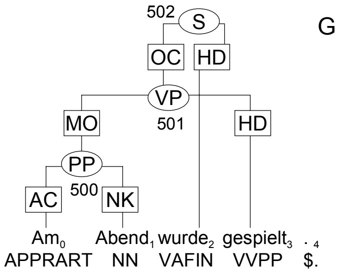
TIGERSearch Description/Query Language 2

- Node level:
  - nodes are feature-value pairs (e.g. word="Farbe", pos="NN")
  - combination of nodes with Boolean expressions (e.g. [word="Farbe" & pos="NN"])
- Node relation level:

  nodes are combined by the following two relations:
  - direct precedence (horizontal dimension)
  - direct dominance (vertical dimension, operator >)

    (e.g. [cat="PP"] > [pos="APPRART"])

TIGERSearch Description/Query Language 3



Graph description level:

(restricted) Boolean expressions combine node relations (e.g. [cat="VP"] > [pos="APPRART"] & [cat="VP"] > [pos="VVPP"])

TIGERSearch Conclusion

- Disadvantages:
  - TIGER Description Language has to be learned in order to carry out queries
  - only one output function (with syntactic annotation)
- Advantages:
  - conversion of different corpus formats to TIGERXML
  - graphical syntax output, highlighting of searched element

TIGERSearch

- Literature:
  - Lezius, Wolfgang and König, Esther. 2000. "Towards a Search Engine for Syntactically Annotated Corpora." KONVENS 2000.
  - Lezius, Wolfgang and König, Esther. 2000. "The TIGER Language."
  - Smith, George. 2002. "A Brief Introduction to the TIGER Sample Corpus"
- Internet Resources:
  - TIGER Project http://www.ims.uni-stuttgart.de/projekte/TIGER

Corpus Query Languages: Overview

- Formal construct designed to retrieve data from corpora
- Corpus query language depends on project; many different versions available
- Conflict between traditional linguistic description languages (i.e. grammar formalisms) and efficiency

Corpus Query Languages: Elements

Corpus query languages consist of the following elements:

- Symbols for constituents;
- Symbols to describe the order of these constituents (horizontally as well as vertically);
- Boolean operators to combine (sequences of) constituents;
- Further options such as case-sensitiveness, number, etc.

General Parameters of Corpus Query

- **Research question:**
  query for word, syntactic constituents, statistical information, etc.?
- **User:**
  beginner, intermittent user, experienced user?
- **Corpus annotation:**
  plain text, PoS-tagged, syntactically annotated, semantic tags?

Technical Considerations of Corpus Query

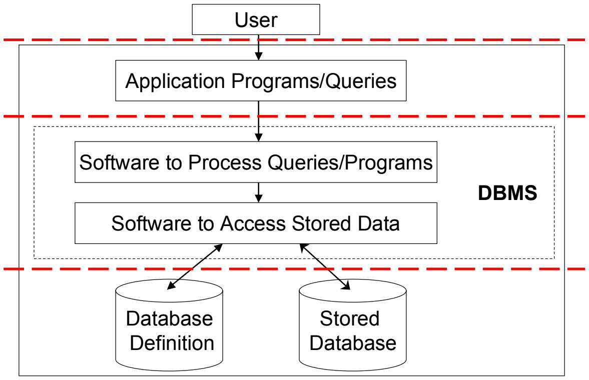
- **Data storage:**
  plain text, XML-encoded text,
  NEGRA Export Format, database, etc.
- **Architecture:**
  local program vs. client/server-architecture
- **Interface:**
  textual input vs. graphical interface
- **Output:**
  KWIC, PoS-tags, syntactic structures, graphical output, lemmas, etc.

My Own Corpus Query Tool

- User: beginner (can be extended to professional user)
- Architecture:
  - webbased query interface (PHP & HTML)
  - MySQL database on server at IFI
- Graphical query interface
- Corpus storage and retrieval from a database

Database Systems

- A database is a logically coherent collection of data with some inherent meaning
- A database is administered by a database management system (DBMS)
- Data in a database is modelled in a scheme which describes their meaning (meta-data)
- Relational Database Systems are based on “tables”

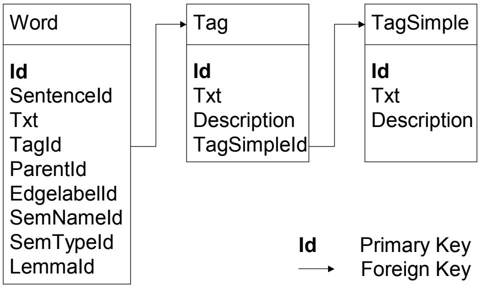

User
Application Programs/Queries
Software to Process Queries/Programs
Software to Access Stored Data
DBMS
Database Definition
Stored Database

Simplified Database Environment (Elmasri, Navathe 2000:6)

Advantages of Database Systems

- Centralized realization of all database functions (such as data definition, data organization, data integrity, access to specific data) allows consistent access to data
- Integration of all data avoids redundancy
- Data is independent of applications
- Database systems take measures to guarantee data integrity and control of multiple users
- “meta-data” informs about structure of data

Relational Database Schema (excerpt)



Relational Database: MySQL Tables (excerpt)

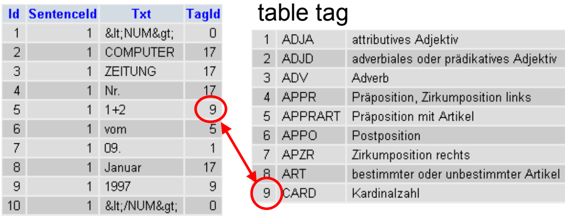


table word

Id	SentenceId	Txt	TagId
1	1	<NUM>	0
2	1	COMPUTER	17
3	1	ZEITUNG	17
4	1	Nr.	17
5	1	1+2	9
6	1	vom	5
7	1	09.	1
8	1	Januar	17
9	1	1997	9
10	1	</NUM>	0

table tag

1	ADJA	attributives Adjektiv
2	ADJD	adverbiales oder prädikatives Adjektiv
3	ADV	Adverb
4	APPR	Präposition, Zirkumposition links
5	APPRART	Präposition mit Artikel
6	APPO	Postposition
7	APZR	Zirkumposition rechts
8	ART	bestimmter oder unbestimmter Artikel
9	CARD	Kardinalzahl

SQL

- SQL (Structured Query Language) is a relational data definition and manipulation language
- SQL query structure:

  SELECT <attribute list>
  FROM <table list>
  WHERE <condition>

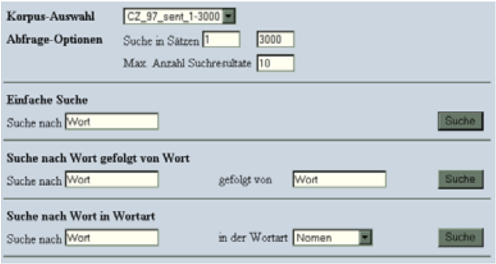
- example query for word “vom”

  SELECT Txt FROM word WHERE Txt=“vom”

Query Possibilities

- Query for words
  - single word
  - word followed by word in variable distance
- Query for PoS-tags
  - single PoS-tag
  - PoS-tag followeg by PoS-tag in variable distance
- Query for syntactical constituents
- Query for lemma
- Corpus-Browsing

Query Interface



Korpus-Auswahl CZ_97_sent_1-3000

Abfrage-Optionen Suche in Sätzen 1 3000

Max. Anzahl Suchresultate 10

Einfache Suche

Suche nach Wort

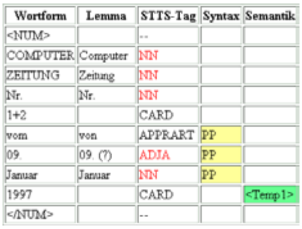
Suche nach Wort gefolgt von Wort

Suche nach Wort gefolgt von Wort

Suche nach Wort in Wortart

Suche nach Wort in der Wortart Nomen

Result Display



Wortform	Lemma	STTS-Tag	Syntax	Semantik
<NUM>		--		
COMPUTER	Computer	NN		
ZEITUNG	Zeitung	NN		
Nr.	Nr.	NN		
1+2		CARD		
vom	von	APPRART	PP	
09.	09. (?)	ADJA	PP	
Januar	Januar	NN	PP	
1997		CARD		<Temp1>
</NUM>		--		

- Simple Query:
  - KWIC
  - with PoS-tags
- Advanced/Lemma Query:
  - full annotation in verticalized sentence table (see left)

My Own Corpus Query Tool: Conclusion

- Disadvantages:
  - restricted versatility of query
- Advantages:
  - easy handling
  - different types of result display
- Performance with large corpora?

Literature

- Literature:
  - Plaehn, Oliver. 1998. “Datenbank-Dokumentation.”
  - Elmasri, Ramez and Navathe, Shamkant. 2000. *Fundamentals of Database Systems*.
- Internet Resources:
  - http://www.ifi.unizh.ch/chmerz/CorpusQuery/start.html