

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

CQP

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

16. Januar 2010

- **Corpus Query Processor (CQP)**
- Werkzeug für Konkordanzen
- *Key Word in Context (KWIC)*
- Merkmale:
 - binäre Kodierung
 - Indexierung
 - Wortformen plus potentiell mehrfache Annotationsebenen

Reguläre Ausdrücke

- Ursprung: Automatentheorie und formale Sprachen
- Definition von Mustern: Zeichenkette, die Menge von Zeichenketten beschreibt, ohne alle Elemente der Liste zu benennen
- Anwendung: Suche und Manipulation von Text, basierend auf Mustern
- Beispiele von Programmiersprachen, die reguläre Ausdrücke inkorporieren: Perl, Python, Tcl
- Basiskonzepte (Beispiele):
 - Menge: [aeiou] [0-9] [A-Za-z]
 - Alternativen: grey|gray gr(e|a)y überschw(e|ä)nglich
 - Quantifizierung: colour?r [0-9]+[.] ([0-9])*[.][0-9]+

CQP: Beispiele

```
‘alt.*’;  
‘sag(e|st|t|en|te|ten)’;
```

```
‘altes’ ‘Haus’;  
‘altes’ []? ‘Haus’;  
‘altes’ []* ‘Haus’;  
‘altes’ []1,3 ‘Haus’;  
‘altes’ [word != ‘(.|,|!)’] ‘Haus’;
```

CQP: Beispiele

```
‘‘altes’’ ‘‘Haus’’;
```

```
[lemma=‘‘alt’’] [lemma=‘‘Haus’’];  
[pos=‘‘ADJA’’ & lemma=‘‘alt’’] [lemma=‘‘Haus’’];
```

```
[pos=‘‘ADJA’’] [lemma=‘‘Haus’’];  
[pos=‘‘ADJA’’] [pos=‘‘ADJA’’]* [lemma=‘‘Haus’’];  
[pos=‘‘ADJA’’] [pos=‘‘ADJA’’]+ [lemma=‘‘Haus’’];  
[pos=‘‘ADJA’’] [pos=‘‘ADJA’’]+ [pos=‘‘NN’’];
```

- Definieren Sie die Variable \$CORPUS_REGISTRY:
`export CORPUS_REGISTRY="/proj/courses/ambig-ws0910/CQP/registry"`
- Prüfen Sie nach, ob die Variable korrekt definiert ist:
`ls $CORPUS_REGISTRY`
Als Antwort müssten Sie 'dewac' bekommen.
- Falls es noch keinen Link in Ihrem Verzeichnis für 'cqp' gibt (bei `ls` erscheint u.a. 'cqp -> ...'), legen Sie ihn an:
`ln -s /proj/courses/ambig-ws0910/CQP/cqp`
Jetzt ist der Link angelegt.
- Starten Sie cqp:
`./cqp -eC`

CQP: Nützliches

- Jeder CQP-Befehl muss mit einem Semikolon abgeschlossen werden.
- Korpus-Namen immer in Großbuchstaben schreiben.
- Liste der vorhandenen Korpora: `show corpora;`
- Information zu einem bestimmten Korpus, z.B. 'dewac':
`info DEWAC;`
- Korpus aufrufen, z.B. DEWAC; Bei erfolgreichem Aufruf zeigt der Prompt nun den Korpus-Namen an.
- Ändern der Kontextgröße (standardmäßig 5 Wörter links und rechts):
`set context 50;` (50 Wörter links und rechts)
`set context s;` (Satz)
- Annotation zu-/abschalten:
`show +lemma;`
`show -lemma;`
`show +pos;`
`show -pos;`
- siehe auch Tutorium zur CQP-Anfragesprache von Stefan Evert



Stefan Evert.

The IMS Open Corpus Workbench (CWB) – Corpus Encoding Tutorial.

Universität Osnabrück, 2008.



Stefan Evert.

The CQP Query Language Tutorial.

Universität Osnabrück, 2009.