

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Übung 1: UNIX und CQP

PD Dr. Sabine Schulte im Walde

16. Januar 2010

1 Aufgabe

- Sie nutzen zwei einfache aber mächtige Werkzeuge für erste empirische linguistische Analysen: Im Betriebssystem UNIX erstellen Sie Häufigkeitsverteilungen, mit dem Corpus Query Processor erzeugen Sie Konkordanzen und ermitteln Indikatoren für Wortbedeutungen.
- Abgabe: Lösungen bis zum 14. Februar 2010 per Email an `schulte@ims.uni-stuttgart.de` schicken.

2 Daten

Sie arbeiten mit einem Ausschnitt des deutschen Web-Korpus *deWaC*. Der Ausschnitt enthält ca. 500,000 Wortformen.

3 Werkzeuge

1. UNIX

- Siehe Foliensatz.
- Nützliche Referenzen: Vieler (1992); Church (1994); Brew and Moens (2002); Baroni (2009)

2. CQP

- Siehe Foliensatz.
- Nützliche Referenzen: Evert (2009), STTS-Tagset

4 Übungen

4.1 UNIX

1. Vorbereitung:

Machen Sie sich mit den Daten vertraut: `less dewac.text`

2. (2 Punkte) Tokenisierung:

Tokenisieren Sie den Text: `cat dewac.text | tr ' ' '\n' > dewac.token`

Erklären Sie kurz, was der Tokenisierer macht.

3. (2 Punkte) Bestimmen Sie die Anzahl der Zeilen, Wörter und Zeichen in beiden Dateien, berichten Sie diese Zahlen und erklären Sie identische und unterschiedliche Ergebnisse.

```
wc dewac_text
wc dewac_token
```

4. (1 Punkt) Sortieren Sie die neu erstellte Datei `dewac_token` alphabetisch:

```
cat dewac_token | sort > dewac_token_sorted
```

Machen Sie dasselbe in umgekehrter Reihenfolge:

```
cat dewac_token | sort -r > dewac_token_sorted_rev
```

Welche Abkürzung ist die erste in dieser letzten Liste?

5. (1 Punkt) Erstellen Sie eine Liste von Typen:

```
cat dewac_token_sorted | uniq > dewac_types
```

Wieviele Typen gibt es?

6. (4 Punkte) Erstellen Sie eine Frequenz-Verteilung:

```
cat dewac_token_sorted | uniq -c | sort -nr > dewac_freq
```

- Erklären Sie das Vorgehen.
- Nennen Sie jeweils zwei Nomen, Verben und Adjektive mit einer Frequenz zwischen 10 und 50 (zusammen mit der jeweiligen Frequenz) sowie zwei Nomen, Verben und Adjektive mit einer Frequenz von 1.

4.2 CQP

1. Vorbereitung:

Starten Sie CQP und laden Sie das Korpus *DEWAC*.

2. (2 Punkte) Ermitteln Sie drei Beispiel-Nomen, die mindestens dreimal im Korpus innerhalb einer Nominalphrase von dem Adjektiv *rot* modifiziert werden. Dokumentieren Sie Ihre Anfragen.

3. (2 Punkte) Ermitteln Sie drei Verben, die mindestens einmal in einem Verb-Letzt-Satz vorkommen. Dokumentieren Sie Ihre Anfrage und je einen Verb-Letzt-Satz pro Verb.

4. (6 Punkte) Wählen Sie zwei ambige Worte (verschiedener Wortart) und ermitteln Sie durch Konkordanz-Analysen Ihrer Wahl Kontext-Indikatoren der Wortbedeutungen. (Diese Indikatoren können Wörter sein, Strukturen etc.) Dokumentieren Sie mindestens drei Indikatoren pro Wort.

References

Marco Baroni. Distributions in Text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, volume 2 of *Handbooks of Linguistics and Communication Science*. Mouton de Gruyter, Berlin, 2009.

Chris Brew and Marc Moens. Data-intensive linguistics. Manuscript, 2002.

Kenneth W. Church. Unix for Poets. Manuscript, AT&T Research, 1994.

Stefan Evert. *The CQP Query Language Tutorial*. Universität Osnabrück, 2009.

Jens Vieler. Einführung in UNIX. Technical Report A/037/9202, Universitätsrechenzentrum, FernUniversität Gesamthochschule in Hagen, 1992.