Intensivwochenende Computerlinguistik

Universität Zürich, Oktober 2015
Martin Volk, Simon Clematide

Korpusabfragen mit TIGERSearch - Musterlösung

Bitte entwickle geeignete Abfragen zur Beantwortung der folgenden Fragen. Durchsuche das englische Korpus zu ***Sophies World*** (528 Bäume, 7829 Tokens).

1. **Wie oft kommt das Wort „*garden*“ im Korpus vor? Kommt das Wort auch im Plural vor?**
   - `[word="garden"]` ➔ 14 Treffer
   - `[word="gardens"]` ➔ 1 Treffer
   - `[word=/[Gg]ardens/]` ➔ (RegEx Suche) 16 Treffer

2. **Wie oft kommt das Wort „*have*“ als Verb in der Grundform (VB) und wie oft als Verb im Präsens (VBP) vor? Kommt es noch mit einem anderen Wortarten (Part-of-Speech) vor?**
   - `#w:[word="have"]` ➔ 34 Treffer in 33 Bäumen (= graphs)
   - Das Statistikmodul sagt uns: "have" kommt mit folgenden Wortarten vor:
     - 18 * VB (= Verb, base form)
     - 15 * VBP (= Verb, non-3rd person singular present)
     - 1 * VBZ (= Verb, 3rd person singular present) → falsch annotiert!

3. **Wie viele Token enthalten eine Ziffer? Nutze einen regulären Ausdruck in Deiner Suche. (Dabei steht \d für eine beliebige Ziffer.)**
   - `[word=/.*\d.*/]` ➔ (RegEx Suche) 3 Treffer
   - Diese Suche heisst: Suche ein Wort, das eine Ziffer (=digit) enthält und optional einige Symbole davor oder dahinter.

4. **Ermittle die Rangfolge der häufigsten Substantive. Ermittle danach für 3 der 10 häufigsten Substantive jeweils das am häufigsten zusammen auftretende Adjektiv.**
   - `#p:[pos=/NNS?/]` ➔ 1107 Treffer in 439 Bäumen (= graphs)
   - Das heisst: Suche alle Wörter mit Wortart = NN (Nomen im Singular) oder Wortart=NNS (Nomen im Plural).
   - Das Statistikmodul zeigt uns die häufigsten Substantive:
     - 39 * world

- 21 * something
- 18 * people
- 18* time
- 16 * rabbit

- Beachte: In der Häufigkeitsliste werden unterschiedliche Formen des selben Nomens separat gezählt. Beispiel: *16 * rabbit* und *1 * rabbits*
- Die Adjektive finden wir mit speziellen Suchen für ausgewählte Nomen:
- `#a:[pos=/JJ[RS]?/] .[word="world"]` ➔ 1 Treffer mit dem Adjektiv *extraordinary*
- `#a:[pos=/JJ[RS]?/] .[word="rabbit"]` ➔ 9 Treffer, davon 5 * *white,* und je 1 * *cozy, live, same, large*

5. **Wie viele Wörter enthalten einen Bindestrich? Stelle sicher, dass Du Wörter mit Bindestrich am Anfang oder Ende (z.B. deutsch *Sicherheits*-) nicht mitzählst.**
   - `[word=/.+\-.+/]` ➔ (RegEx Suche) 12 Treffer
   - Diese Suche heisst: Suche ein Wort, das einen Bindestrich enthält und mindestens ein Symbol davor oder dahinter. Diese Suche findet jedoch auch Wörter, die am Anfang einen Bindestrich haben, wenn sie im Innern noch einen Bindstrich haben. Deshalb ist die folgende Suche genauer:
   - `[word=/[^-]+.*\-.*[^-]+/]` ➔ (RegEx Suche) 11 Treffer
   - Diese Suche garantiert, dass am Anfang und Ende des gefundenen Wortes keine Bindestriche stehen und dass mindestens ein Bindestrich im Wortinnern vorkommt.

6. **Welche Substantive nehmen am häufigsten ein Satzkomplement (SBAR)? In anderen Worten: Welche Substantive werden von einem SBAR-Nebensatz näher bestimmt? Gebe die 3 häufigsten mit je einem Beispielsatz an.**
   - `#np:[cat="NP"] > [cat="SBAR"] &`
   - `#np > #n:[pos=/NNS?/]` ➔ 3 Treffer
   - Diese Suche heisst: Suche eine NP, die ein "SBAR" (= modifizierender Nebensatz) direkt dominiert und die ausserdem ein Substantiv direkt dominiert. Das Resultat sind nur die 3 Substantive *something, thought, thing*. Andere Substantive, die ebenfalls ein "SBAR" nehmen, werden in dieser Baumbank nicht direkt sondern indirekt von der NP dominiert. Deshalb können wir die Suche allgemeiner formulieren als:
   - `#np:[cat="NP"] > [cat="SBAR"] &`
   - `#np >* #n:[pos=/NNS?/]` ➔ 130 Treffer in 56 Bäumen
   - Beispiel: s44 *"... the fact, that her mother never got home ...*". Dieser Satz liefert aber noch 5 weitere Treffer, da auch alle Substantive im SBAR-Nebensatz als mögliche Substantive gezählt werden. Deshalb präzisieren wir die Suche:

- `#np:[cat="NP"] > #sbar:[cat="SBAR"] &`
- `#np >* #n:[pos=/NNS?/] &`
- `#n . #sbar` ➔ 48 Treffer in 50 Bäumen
- Die häufigsten Substantive, die von einem SBAR-Nebensatz näher bestimmt werden, sind
    - 6 * something
    - 4 * questions + 1 * question
    - 3 * thing
    - 3 * fact

7. **Welche Konstituenten haben nur eine Tochterkonstituente (= unäre Knoten)? Welches sind die 5 häufigsten Konstituenten?**
    - `#c:[cat=/.*/] &`
    - `tokenarity(#c, 1)` ➔ 2044 Treffer in 508 Bäumen
    - Die obige Suche findet Konstituenten, die genau ein Token umfassen.
    - `#c:[cat=/.*/] &`
    - `arity(#c, 1)` ➔ 2044 Treffer in 508 Bäumen
    - Die obige Suche mit **arity** statt **tokenarity** findet Konstituenten, die genau einen Tochterknoten haben. Aufgrund der gleichen Treffermenge sehen wir, dass diese Tochterknoten immer Token (~ Wörter) sind. Wir können diese Suchanfrage noch vereinfachen mit dem Sondersymbol **NT** für Nicht-Terminal (= Konstituente im Baum).
    - `#c:[NT] &`
    - `arity(#c, 1)` ➔ 2044 Treffer in 508 Bäumen
    - Das Statistikmodul sagt uns, dass die folgenden Konstituenten am häufigsten nur 1 Tochterkonstituente haben:
        - 1295 * NP
        - 350 * ADVP
        - 92 * WHNP
        - 67 * ADJP
        - 64 * PRT

8. **Wie viele Sekundärkanten enthält das Korpus? Welches sind die drei häufigsten Zielknoten einer Sekundärkante?**
    - `#source >~ #target` ➔ 366 Treffer in 250 Bäumen
    - Das Statistikmodul sagt uns, dass die folgenden Konstituenten die häufigsten Zielknoten sind:
        - 170 * NP
        - 88 * WHNP
        - 59 * WHADVP

- 25 * S
- 11 * SBAR

9. Welche Konstituenten (cat) sind Bestimmungen des Ortes (Funktion LOC)? Bitte mit Häufigkeit angeben.

- `#cat1 >LOC #cat2` ➔ 141 Treffer in 114 Bäumen
- Das Statistikmodul sagt uns, dass die folgenden Konstituenten in unserer Baumbank Ortsbestimmungen sind:
  - 109 * PP
  - 29 * ADVP
  - 2 * SBAR
  - 1 * UCP

10. Wieviele Subjekte bestehen nur aus einem Personalpronomen?

- `#cat >SBJ [pos="PRP"]` ➔ 1 Treffer
- Die obige Suche resultiert nur in einem Treffer, da Pronomen normalerweise eine NP bilden und diese NP in Subjektfunktion steht (und nicht das Pronomen selbst). Dieser eine Treffer ist also ein Annotierungsfehler. Die korrekte Suche muss also die Zwischenkategorie mit berücksichtigen.
- `#cat1 >SBJ #cat2 &`
- `#cat2 > [pos="PRP"]` ➔ 412 Treffer in 291 Bäumen

11. Wieviele Verb-Partikel Paare (*look up, look out, figure out*) enthält diese Baumbank? Wie oft stehen Verb und Partikel nicht unmittelbar nebeneinander?

- `#cat > [cat="PRT"] &`
- `#cat > [pos=/VB.?/]` ➔ 64 Treffer in 60 Bäumen
- Die obige Suche heisst: Suche eine (beliebige) Konstituente #cat, die eine Konstituente mit Namen "PRT" dominiert. Die selbe Konstituente #cat soll auch ein Verb dominieren. Diese Suche findet alle Verb-Partikel Paare. Wenn man das Verb mit einem Namen (= Variablen-Bindung) versieht und vom PRT-Knoten auf das Partikel weitergeht, kann man mit Hilfe des Statistikmoduls zählen, welche Paare wie häufig vorkommen. Also:
- `#cat > #p:[cat="PRT"] &`
- `#p > #part:[pos="RP"] &`
- `#cat > #v:[pos=/VB.?/]`
- Die häufigsten sind:
  - 3 * pulled out
  - 3 * grow up
  - 2 * put on

    - 2 * figure out
- Nun wollen wir die Suche einschränken auf alle Verb-Partikel Paare, wo zwischen dem Verb und dem Partikel mindestens ein Wort steht. Eine mögliche Formulierung der Suche ist:
- `#cat > #p:[cat="PRT"] &`
- `#cat > #v:[pos=/VB.?/] &`
- `#cat > #middle &`
- `#v . #middle &`
- `#middle .* #p` ➔ 5 Treffer in 5 Bäumen
- Das heisst: Suche eine (beliebige) Konstituente #cat, die eine Konstituente mit Namen "PRT" dominiert (die wir hier #p nennen). Die selbe Konstituente #cat soll auch ein Verb dominieren (das wir hier #v nennen) sowie eine weitere (beliebige) Konstituente #middle. Das Verb #v soll unmittelbar vor #middle stehen, und #middle soll (unmittelbar oder mittelbar) vor dem Partikel #p stehen.

12. Eine Folge von Adjektiv – Nomen – Konjunktion – Nomen (z.B. *old men and women*) ist im Englischen potentiell mehrdeutig. Eine solche Wortfolge kommt einmal in diesem Korpus vor. Wie ist die Mehrdeutigkeit dort aufgelöst?

- `[pos=/JJ[RS]?/] . #n1:[pos=/NNS?/] &`
- `#n1 . #conj:[pos="CC"] &`
- `#conj . [pos=/NNS?/]`
- Gefunden wird die Sequenz "*delicious food and drink*". In dieser Sequenz bezieht sich "*delicious*" nur auf "*food*" und nicht auf "*drink*".

Ziel:

- Erlernen einer Korpusabfragesprache.
- Erfahrungen beim Durchsuchen einer Baumbank.