Corpus linguistics, assignment 2

Using TIGERSearch to analyze a syntactically annotated corpus

To solve the assignment, you should select at least one of the following treebanks: Talbanken (Swedish), TIGER (German), The Brown part of Penn Treebank (English), or Turin University Treebank (Italian). We have installed these treebanks for use with TIGERSearch in the lab room. Please contact Richard if you want to use some other treebank.

Solve the exercises below and write a short answer for each exercise. You should solve the tasks marked **ALL** or with the language of your treebank. Send your solutions to me (`richard.johansson@gu.se`) before December 13. I will also be happy to answer your questions about the assignment.

Some hints:

- There may be many possible ways these tasks can be solved.
- If you are unsure of how some linguistic phenomenon is annotated in your treebank, try to search for examples.
- In some cases, the solutions cannot be found using a single query: then you will have to make several queries and compile the result manually.
- You may also at times need to inspect your results manually.
- There may be some cases where you will have to go for an approximate solution. If you think your solution is approximate, please explain why.

Here are some references you may find useful:

- The TIGERSearch manual:

  `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/manual_html.html`
- TIGERSearch quick reference:

  `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/QueryLanguage_QuickReference.html`
- Swedish:
  - Word classes: `http://stp.lingfil.uu.se/~nivre/swedish_treebank/pos.html`
  - Phrase labels: `http://stp.lingfil.uu.se/~nivre/swedish_treebank/PS.html`
  - Function labels: `http://stp.lingfil.uu.se/~nivre/swedish_treebank/GF.html`
- English: short summary `http://bulba.sdsu.edu/jeanette/thesis/PennTags.html` or the full manual `https://svn.spraakdata.gu.se/repos/richard/pub/kl2013_web/penn-manual.pdf`
- German: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_introduction.pdf`

Your tasks:

1. **ALL** Consider some word with word class ambiguity, for instance *man* in Swedish, *that* in English, *das* in German, *che* in Italian. How often do the different word classes occur?

2. **ALL** Determine the 3 most frequent nouns, and in each case determine the most frequently occurring adjective together with this noun.

3. **ALL** Which are the three nouns most frequently occurring with an infinitive clause? Here are examples of such nouns.

   (English) *her desire to be in show business*
   (Swedish) *rätten att göra avdrag*
   (German) *die Verdächtigung, ein Ladendieb zu sein*

(Italian) *l'obbligo di pagare*, *un edificio da costruire*

4. **ALL** Out of all the grammatical subjects in your treebank, how many are pronouns? What is the proportion of objects that are pronouns?

5. **ALL** A clause (S) occurring inside another clause is said to be *nested*. How deep is the most deeply nested clause in your treebank?

6. **ALL** A *cleft sentence* is a sentence of the type:

   (English) *It was from me that she heard the news.*

   (Swedish) *Det var jag som tog den!*

   (German) *Es war Klaus, der atemlos über den Bahndamm stolperte.*

   (Italian) *Sei tu che fai così!*

   Can you find any such sentences in your treebank?

7. **Swedish, English** Which phrase types are most frequently used as location adverbials? Which verbs most frequently take location adverbials?

8. **Swedish, German** Which are the most common phrase types used in the fundament position of a clause (that is, the position just before the finite verb)? How often is the phrase in the fundament a subject?

9. **Swedish, English, German** How often does a subject occur after the finite verb of the clause? How often does the object occur before the verb?

10. **Swedish, English, German** How many verb–particle pairs does your treebank contain? Which phrase type occurs most frequently between a verb and a particle?

11. **Swedish, English** When a preposition appears alone, without an adjacent complement, we say that we have *preposition stranding*. Which are the most frequent prepositions in your treebank for which this occurs?

    (English) *You have nothing to worry about.*

    (Swedish) *Tomten tror alla barn på.*

12. **German** In German, a verb and its particle can occur quite far from each other. For instance, in this sentence there are three phrases between the verb and the particle.

    Wie stellen [Sie] [sich] [das] vor?

    What is the highest number of phrases you can find between a verb and its particle?

13. **Italian** Which are the most common reflexive verbs? Hint: note that the Turin treebank has a "funny" way of annotating contracted words. For instance, a verb with a clitic pronoun such as *farsi* will appear twice, once tagged as a verb and once as a pronoun.

    Example: *La stessa disposizione si applica se il deperimento ...*

    Example: *Egli può sempre opporsi opporsi a chi non è munito ...*

14. **Italian** Which are the words that most frequently appear just before a clause in the conjunctive mood?

    Example: *Non risulta che quest'articolo sia stato mai applicato.*