Today: Extracting and importing data from syntactic corpora into a database

Florian Jaeger & Judith Degen
Today

• Get your feet wet:
  – What is a **syntactically-annotated corpus**?
  
  – **TGrep2** :: a tool to search syntactically-annotated corpora

• Next week:
  – **TDT lite** :: a set of scripts we wrote to combine TGrep2 output into a database that can be handed to Excel or a stats program of your choice (e.g. R).
Timeline for Corpus-based Project

- What is the structure of interest?
- What are the mark-up conventions of the corpus?
- Define & refine patterns (TGrep2; TigerSearch; Tregexp):
  - avoid over-inclusive (easy, except for large databases)
  - avoid over-exclusive (hard)
  - cost-accuracy-tradeoff (less clean-up → noisier data)
- Extraction of variables of interest:
  - May need annotation (Edinburgh Nite Toolboxes)
  - May need scripting (TGrep2 Database Tools)
  - cost-accuracy-tradeoff (cheap estimates → noisier estimates)
- Additional processing (smoothing; LSA)
- Statistical analysis (R software package; R-lang email list)
  - Clusters require mixed models, bootstrap, … (lmer(), bootcov())
**that-omission**

- Non-subject-extracted relative clauses in English allow optional *that*-omission:

  \[
  \text{How big is the family } \{ \begin{align*}
  \text{for?} \\
  \text{you cook for?}
  \end{align*} \}
  \]

  \[
  \text{that you cook}
  \]
Timeline for Corpus-based Project

- What is the structure of interest?

- What are the mark-up conventions of the corpus?

- Define & refine patterns (TGrep2; TigerSearch; Tregexp):
  - avoid over-inclusive (easy, except for large databases)
  - avoid over-exclusive (hard)
  - cost-accuracy-tradeoff (less clean-up → noisier data)

- Extraction of variables of interest:
  - May need annotation (Edinburgh Nite Toolboxes)
  - May need scripting (TGrep2 Database Tools)
  - cost-accuracy-tradeoff (cheap estimates → noisier estimates)

- Additional processing (smoothing; LSA)

- Statistical analysis (R software package; R-lang email list)
  - Clusters require mixed models, bootstrap, … (lmer(), bootcov())
TGrep2

• Search tools for syntactic corpora developed by Doug Rohde (2005)
  – Downloadable for free:
    http://tedlab.mit.edu/~dr/Tgrep2/
  – Online tutorial:
    http://www.bcs.rochester.edu/people/fjaeger/teaching/tutorials/TGrep2/LabSyntax-Tutorial.html

• Parsed Switchboard in Penn Treebank format
  – 800,000 word syntactically annotated telephone conversation corpus (Switchboard, Treebank III)
A common syntactic annotation standard

- Syntactic structure annotation
  - Hierarchical dependencies
  - Linear order
  - Traces
  - Syntactic categories

- Predicate argument structure annotation
  - Grammatical functions (e.g. SUBJ, TOP, ADV, ...)
  - Modification types (e.g. NP-TEMP, ADV-LOC, ...)
  - Case marking preposition (e.g. PP-DTV)

- Part-of-speech (POS) annotation

- In Switchboard: disfluency (reparandum, repair)

- Genre, speaker, etc. information
WSJ

(TOP (S (NP-SBJ (NP (NNP Pierre)
       (NNP Vinken)))
       (, ,)
       (ADJP (NP (CD 61)
              (NNS years))
              (JJ old))
       (, ,))
(VP (MD will)
   (VP (VB join)
      (NP (DT the)
      (NN board))
   (PP-CLR (IN as)
      (NP (DT a)
      (JJ nonexecutive)
      (NN director)))
   (NP-TMP (NNP Nov.)
    (CD 29))))
   (. .)))
(TOP (S (NP-SBJ (NNP Mr.)
       (NNP Vinken)))
(VP (VBZ is)
   (NP-PRD (NP (NN chairman))
   (PP (IN of)
      (NP (NP (NNP Elsevier)
      (NNP N.V.)) ...)
(TOP (CODE (SYM SpeakerA1) 
  (. .)))

(TOP (INTJ (UH Okay) 
  (. .) 
  (-DFL- E_S)))

(TOP (S (INTJ (UH Uh)) 
  , , 
  (ADVP-TMP (RB first)) 
  , , 
  (INTJ (UH um)) 
  , , 
  (NP-SBJ-1 (PRP I)) 
  (VP (VBP need) 
    (S (NP-SBJ (-NONE- *-1)) 
      (VP (TO to) 
        (VP (VB know) 
          , , 
          (INTJ (UH uh)) 
          , , 
          (SBARQ (WHADVP-2 (WRB how)) 
            (SQ (VBP do) 
              (NP-SBJ (PRP you)) 
              (VP (VB feel) 
                (ADVP (-NONE- *T*-2)) 
                (EDITED (RM (-DFL- [])) 
                  (PP-UNF (IN about))))))

SWBD
Annotations in SWBD: NITE XML

- Combination of annotations from different projects in one big data structure
- Nodes can
  - have children (hierarchical relationship)
  - point at other nodes (arbitrary relationship)
- Some nodes have timing information from original sound files
(SBAR (WHADVP (N 400B34) (WDT that))
  (S (NP-SBJ_MARKABLE_human (N 400B21) (PRP we))
    (VP (VBD had)
      (S (NP-SBJ_MARKABLE (-NONE- (N 400B21)))
        (VP (TO to)
          (VP (VB do)
            (NP_MARKABLE_nonconc (PRP it))
            (ADVP-TMP (-NONE- (N 400B34))))))))
(S (NP-MARKABLE (-NONE- (N 40121A))))
(S (WHNP_MARKABLE (N 401508) -NONE-)
  (S (NP-SBJ_MARKABLE_human (N 401623) (PRP we))
    (VP (MD could)
      (VP (VB have)
        (VP (VBN done)
          (NP_MARKABLE (-NONE- (N 401608)))))))
(SBAR (WHADVP (N 405458) (WDT that))
  (S (NP-SBJ_MARKABLE_human (PRP they))
    (VP (VBD were)
      (ADJP-PRD (JJ concerned))
      (ADV (NONE- (N 405458)))))))
(SBAR (ADVP (RB especially))
  (WHADVP (N 40721B) (WRB where)))
(S (NP-SBJ_MARKABLE_human (PRP they))
  (VP (VBP 've)
    (VP (VBN had)
      (INTJ (UH uh))
      (NP_MARKABLE_human (JJ extended)))))

Tgrep2 search pattern for RC*s
/^SBAR/  >  /^NP/
< (^/WH/ != /PP/)
< (^/S/ < (^/-SBJ/ !< ````-NONE-''))
!< IN\WDT\DT
!< ````-NONE-''
Data

• Over 3,700 RC*s (RCs with obligatory *that* were excluded) from approximately 350 different speakers
Timeline for Corpus-based Project

• What is the structure of interest?
• What at the mark-up conventions of the corpus?
• Define & refine patterns (TGrep2; TigerSearch; Tregexp):
  – avoid over-inclusive (easy, except for large databases)
  – avoid over-exclusive (hard)
  – cost-accuracy-tradeoff (less clean-up → noisier data)

• Extraction of variables of interest:
  – May need annotation (Edinburgh Nite Toolboxes)
  – May need scripting (TGrep2 Database Tools)
  – cost-accuracy-tradeoff (cheap estimates → noisier estimates)

• Additional processing (smoothing; LSA)
• Statistical analysis (R software package; R-lang email list)
  – Clusters require mixed models, bootstrap, … (lmer(), bootcov())
• Extracting all RC*s with a pronoun subject:

tgrep2 -af -m "\%xm\n" "/^SBAR/ > /^NP/ < (/^WH/ != /PP/) < (^S/ < (-SBJ/ < /^PRP/)) !< IN|WDT|DT !< `\-NONE-`"

outputs:

5:73
21:68
31:28
41:25
236:62
331:168
589:30
651:9
...

Variables in the model

- Use a set of scripts *(TGrep2 Database Tools)* to combine the output of many TGrep2 searches into a database of cases.

- Probabilities:
  - RC Predictability; Predictability of RC onset
  - Frequency of words immediately preceding and following RC onset
Variables in the model

- Continuous syntactic variables, e.g.
  - Lengths of each of 3 regions (pre-NP, between head noun and RC, & RC)
- Categorical structural variables, e.g.
  - Embedding within the RC
  - Properties of RC subject (NP type, animacy)
  - Properties of matrix clause (negation, verb)
- Structural priming, e.g.
  - Within speakers
  - Across speakers
  - Distance-based; Lemma-based; etc …
Variables in the RC* model

• Phonological variables, e.g.
  • segmental properties of preceding segment
  • stress structure of preceding segment
• Speech variables, e.g.
  • Speech rate, Pauses
  • Rate of disfluency in different regions
  • (Prosodic phrases & accents)
• Social variables, e.g.
  • Age
  • Speaker gender
  • Education
Timeline for Corpus-based Project

- What is the structure of interest?
- What are the mark-up conventions of the corpus?
- Define & refine patterns (TGrep2; TigerSearch; Tregexp):
  - avoid over-inclusive (easy, except for large databases)
  - avoid over-exclusive (hard)
  - cost-accuracy-tradeoff (less clean-up → noisier data)
- Extraction of variables of interest:
  - May need annotation (Edinburgh Nite Toolboxes)
  - May need scripting (TGrep2 Database Tools)
  - cost-accuracy-tradeoff (cheap estimates → noisier estimates)
- Additional processing (smoothing; LSA)
- Statistical analysis (R software package; R-lang email list)
  - Clusters require mixed models, bootstrap, … (lmer(), bootcov())
Results of model

• Predictability one of the most influential factors
  • Both RC* predictability and the predictability of the RC* onset affect *that*-rates even when many other factors are considered

• As predicted by Uniform Information Density
Contemporary American English with Penn Treebank III annotation – Text

- Parts of ATIS-3
- Parsed Brown corpus, release 3
  - approx. 24,000 sentences & 396,000 words
  - 15 different written text categories of (good standard reference; like BNC).
- Parts of Wall Street Journal corpus (WSJ), release 3
  - approx. 24k sentences & 505,000 words [1 million out of 30 million]
  - Newspaper articles
  - Also available:
    - RST discourse annotation (for parts)
    - Propositional/event structure annotation (113,000 verb tokens; 3,200 verb types)
    - Automatically annotated extension to 30 million words
Contemporary American English with Penn Treebank III annotation – Speech

- **International Corpus of English** (ICE-GB)
  - approx. 84,000 sentences & 1 million words
  - Speech and written language
  - Not quite Treebank III annotation style

- Parts of **Switchboard corpus** (Swbd), release 3
  - approx. 100k sentences & 800,000 words [1 million out of 2 million]
  - Spontaneous speech
  - Also available:
    - Disfluency annotation (all)
    - Sound files (all)
    - Phonetic & phonological annotation (~38,000 words)
    - Animacy annotation (~140,000 NPs)
    - Information Structure annotation (~60,000 NPs)
Diachronic American English with Penn Treebank III annotation

- The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)
  - approx. 110,000 sentences & 1.5 million words
  - Also available:
    - Text source, genre, dialect, and publication date information

- Helsinki Parsed Corpus of Middle English, second edition (PPCME2)
  - Over 100,000 sentences & 1.3 million words
  - Prose text samples of Middle English
  - Also available:
    - Text source, genre, dialect, and publication date information
POS & Syntactically annotated corpora of other languages - (1)

- Parsed **NEGRA corpus**, version 2
  - German
  - approx. 200,000 sentences
  - Newspaper articles (Frankfurter Rundschau)
  - Also available:
    - Morphological analysis (first 60,000 words)

- Parsed **TIGER corpus**
  - German
  - approx. 40,000 sentences & 700,000 words
  - same source as NEGRA

- **Prague Dependency Treebank**, version 1.0
  - Czech
  - approx. 1.8 million words
POS & Syntactically annotated corpora of other languages - (2)

- **Penn Chinese Treebank**, version 6
  - approx. 600,000 words
  - Newswire text

- **Penn Arabic Treebank**, Part 3, version 1.0
  - approx. 340,000 words
  - Newswire text
  - Also available:
    - Vocalization and Lemmatization information
    - Aligned translations into English (for parts)

- **Penn Korean Treebank**, 
  - approx. 5,000 sentence & 55,000 words
  - 33 constructed texts in Korean (translated into English) for purposes of language training in a military setting.
Let’s do some practice

• Login to/login into/log into the corpus server
  ssh <username>@slate.hlp.rochester.edu
Sanity check

• Type `env` (and press enter):
  TGREP2_CORPUS=/p/hlp/corpora/TGrep2able/swbd.t2c.gz
  TGREP2ABLE=/p/hlp/corpora/TGrep2able/
  TDTlite=/p/hlp/tools/TDTlite/
  TDT_DATABASES=/p/hlp/tools/TDT/databases/
  PATH=...:/p/hlp/tools/TDTlite
TGrep2

- Type `tgrep2`

- `tgrep2 -c <corpus> -af <output-options|output-formatting> <macro-file> <pattern|pattern-file>`

  - `-c <corpus>` defaults to TGREP2_CORPUS
  - `-af` gives all matches exactly once
  - `-i` makes TGrep2 case-insensitive (default is case-sensitive)

  `<output-options>` and `<macro-file>` are optional
TGrep2

• ... a very simple call: let’s find sentences in the default corpus (Switchboard)

\texttt{tgrep2 \textquote{TOP} \textpipe more}

[\textit{more} gives output page-by-page – press ENTER or SPACE]
Tgrep2

• let’s find NPs
  tgrep2 “NP” | more

• Now let’s count:
  tgrep2 “NP” | wc -l

[wc -l counts lines of the output; Tgrep2 defaults to one match per line]
TGrep2 – Different outputs

• We can format the output:
  
  \texttt{tgrep2 -l "NP" | more}
  
  \texttt{tgrep2 -t "NP" | more}
  
  \texttt{tgrep2 -u "NP" | more}
  
  \texttt{[be cautious with the tgrep2 \texttt{-l | wc \texttt{-l}]}

• There are more options for later …
TGrep2 – Regular Expressions

• Let’s count all instances of any type of NP in the corpus:
  
  `tgrep2 -af “NP” | wc -l`
  
  `tgrep2 -af “/^NP/” | wc -l`

• Investigate why there is a difference:
  
  `tgrep2 -af “/^NP/” | more`
Across Corpora

• Count all instances of any type of NP in the Wall Street Journal, Brown, and Switchboard corpus

```
l$TGREP2ABLE
brown.t2c.gz
wsj_mrg.t2c.gz
swbd.t2c.gz
tgrep2 -c $TGREP2ABLE/<corpus-file> -af "/^NP/" | wc -l
```

• What’s the ration of NPs (/^NP/) to VPs (/^VP/) in the three corpora?
How many of these NPs have lexical content (as opposed to traces)?

```
tgrep2 -af "/^NP/ <= (/^\{0,1\}[a-zA-Z].*/ @< *)" | wc -l
```

• NB:
  – Left-headedness
Time to get real: PP-ordering in English
(Hawkins, 1999; taken from Hawkins, 2007:97)

(19)  a. The man vp[wanted pp1[for his son] pp2[in the cold but not unpleasant wind]]
1 2 3 4 5
----------------------------------

b. The man vp[wanted pp2[in the cold but not unpleasant wind] pp1[for his son]]
1 2 3 4 5 6 7 8 9
----------------------------------

Structures like (19) were selected from a corpus on the basis of a permutation test (Hawkins, 2000, 2001): the two PPs had to be permutable with truth-conditional equivalence (i.e. the speaker had a choice). Only 15% (58/394) of these English sequences had long before short. Among those with at least a one-word weight difference (excluding 71 with equal weight), 82% had short before long, and there was a gradual reduction in the long before short orders, the bigger the weight difference (PPS = shorter PP, PPL = longer PP):

(22) PPL > PPS by 1 word by 2 4 by 5 6 by 7 +
[V PPS PPL] 60% (58) 86% (108) 94% (31) 99% (68)
[V PPL PPS] 40% (38) 14% (17) 6% (2) 1% (1)
Time to get real ...

• What should be the cases we extract to get **all and only** the relevant cases? (avoid inclusion and exclusion errors)

• VPs
• VPs with PPs
• VPs with PPs that are sisters to each other
• VPs with adjacent PPs that are sisters to each other
• VPs with exactly two adjacent PPs that are sisters to each other
Cheat sheet

- TGrep2 is left-headed!
- Syntactic relations: < > << >> $ ~ =
- Linear relations: , .
- Labeling of nodes: =xx
- Disjunction | []
- Negation: !
/\^VP/=VP1 < (\^PP/=PP1
 $.. (\^PP/=PP2 !<$ (\^PP/ != =PP1)
 !,, (* !< * ,, =PP1
 !>> (EDITED|UH|PRN)/-
 >> =VP1))))

UNF/
Macros

- Macros keep those precious fingers soft and smooth by avoiding too much typing

@ NP          /^NP/;
@ VP          /^VP/;
@ PP          /^PP/;
@ AP          /^\(ADJ|ADV\)P/;
@ WH          /^WH/;
@ SBJ_ZERO    (@SBJ) < (@ZERO);
@ SBJ_NERO    (@SBJ) !< (@ZERO);
@ SSBJ_ZERO   S < (@SBJ_ZERO);
@ SSBJ_NERO   S < (@SBJ_NERO);