LSA 125 – Psycholinguistics and Syntactic Corpora

Today: Other ways to do corpus-based psycholinguistics

LSA Summer Institute 2009, UC Berkeley

Florian Jaeger

TAs: Judith Degen, Alex Fine, and Peter Graff

Today

- **Corpora vs. Experiments**
- **Other useful tools** and thoughts on …
  - how to use *non*-syntactic corpora for research on sentence production.
  - how to use corpora for work on *comprehension*.
  - Some really cool resources

Common claims against corpus-based research to language production

- … by linguists:
  - People make mistakes
  - Not clear that all speakers are native speakers

  - Rareness of relevant events → cannot argue from lack of occurrence
  - Grammaticality != judgments != occurrence in corpus
  - Corpora cannot be used to test usage-based gradience in data since they average over speakers

Common claims against corpus-based research to language production

- … by psychologists:
  - Corpora provide only heterogeneous data which make reliable hypothesis testing impossible
  - Corpora cannot be used to learn about the underlying mechanisms of language production
  - Corpora only allow post-hoc testing of hypotheses

Trade-offs

- Ecological validity
- Heterogeneity (genre, styles, subject populations)
- Distributions

What happens when we present weird stimuli to our subjects?

- Recently on the streets:

  “The plumber that the clerk saw in a pub bought a house at the river.”

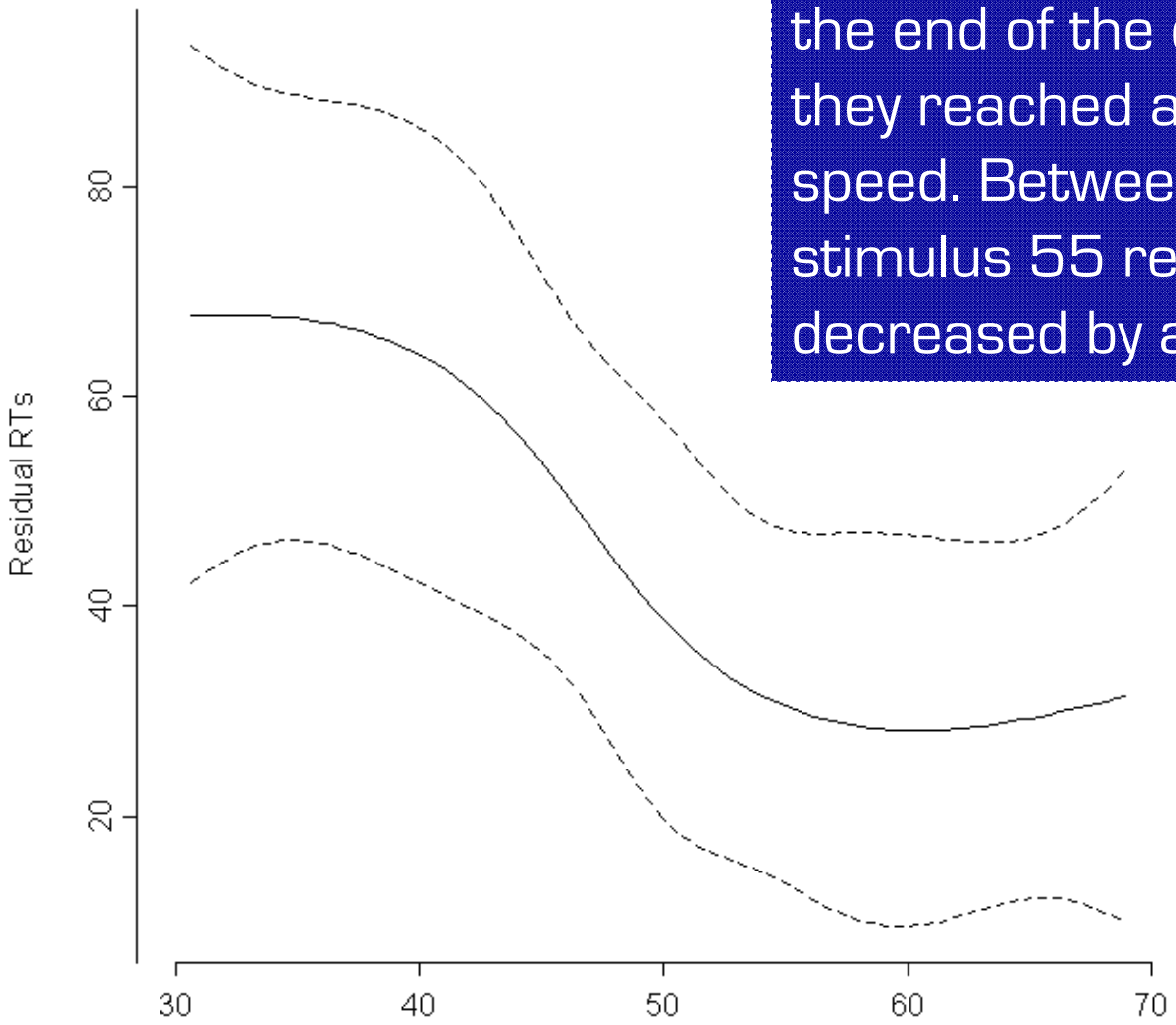
  “The boy that the girl kicked is tall.”

  “Who do you think believes that people get used to unnatural stimuli?”

Processing Effects in Superiority-Violating *wh*-questions

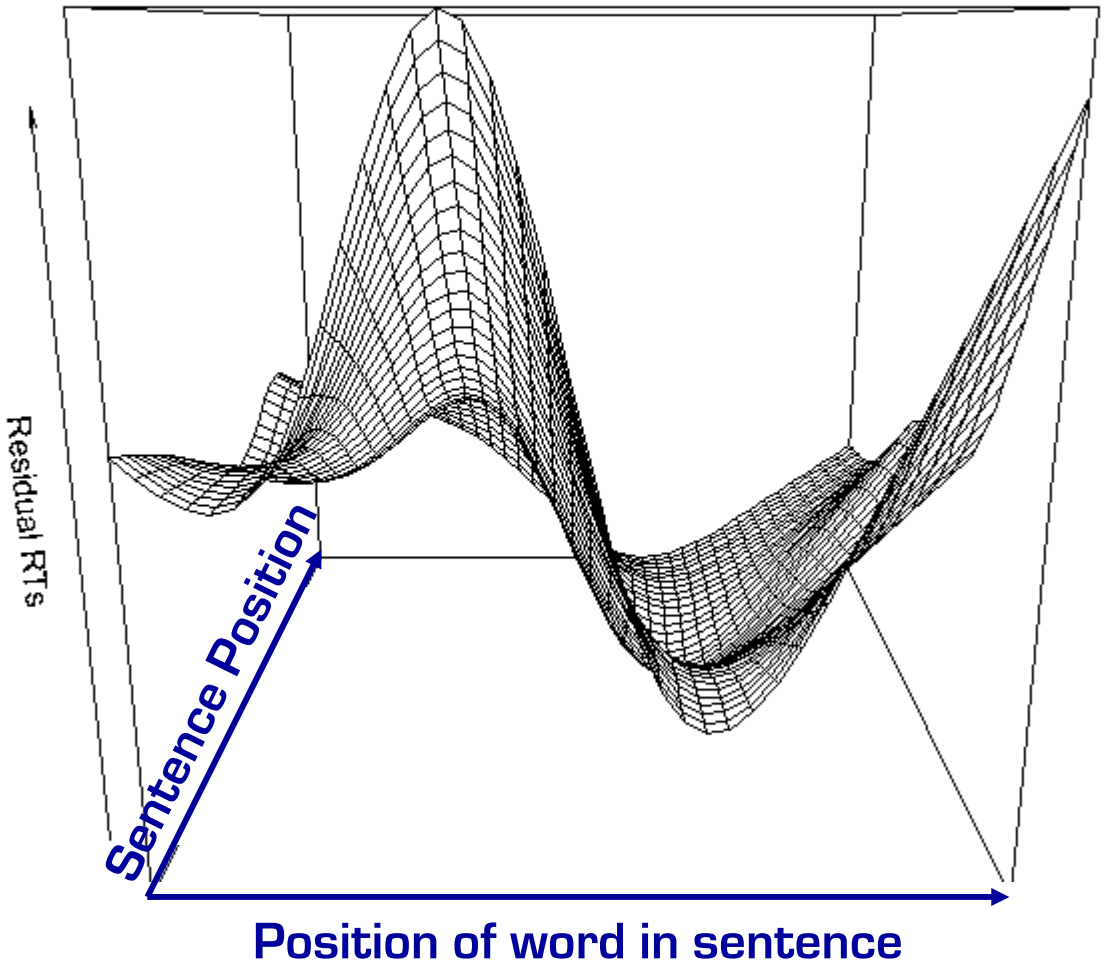
- Self-paced reading study investigating the effect of *wh*-phrase type on the processing of binary superiority-violating *wh*-questions (in the target region)
- 2 (1st *wh*-phrase) x 2 (2nd *wh*-phrase) design:
  - BARE_BARE
    Mary wondered what who read but later the teacher told her.
  - BARE_WHICH
    Mary wondered what which student read but …
  - WHICH_BARE
    Mary wondered which book who read but …
  - WHICH_WHICH
    Mary wondered which book which student read but …

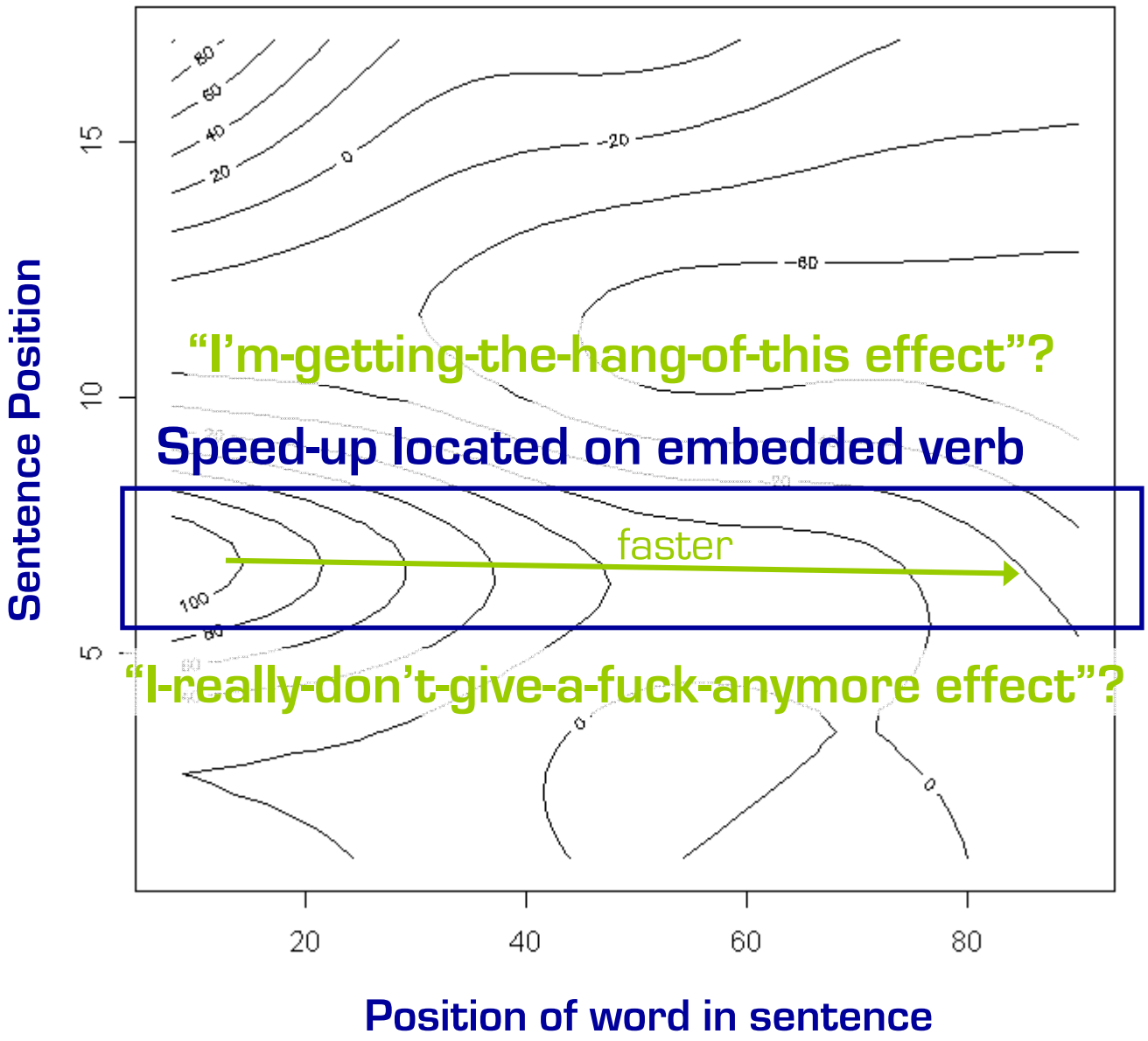
Participants read faster towards the end of the experiment until they reached a maximum reading speed. Between stimulus 40 and stimulus 55 residual reading times decreased by almost 50%.



Position of sentence in list that the subject saw

But there's more to it than meets the eye. The position of a sentence in the list interact with position of the word in the sentence.




Sentence Position
Position of word in sentence
15
10
5
20
40
60
80
-80
-60
-40
-20
0
-20
-60
100
80
0
0
"I'm-getting-the-hang-of-this effect"?
Speed-up located on embedded verb
faster
"I-really-don't-give-a-fuck-anymore effect"?

So what can be do?

- Why not sample from corpora when creating experiments, so that the stimuli are representative of natural language?

- The problem of unbalanced data can be addressed with modern statistical methods

UID's prediction for *that*-omission

- Speakers' decisions to pronounce or omit optional words should depend on processing complexity:
  - Optional words should be inserted, whenever processing would otherwise be difficult
  - Optional words should be omitted, whenever processing would be easy anyway

$$\log \frac{1}{p(RC \mid ctxt)}$$

Phrase onset

$$\log \frac{1}{p(w_i \mid RC, ctxt)}$$

Lexical content

How big is the family { *that* / **you** } *cook for?*

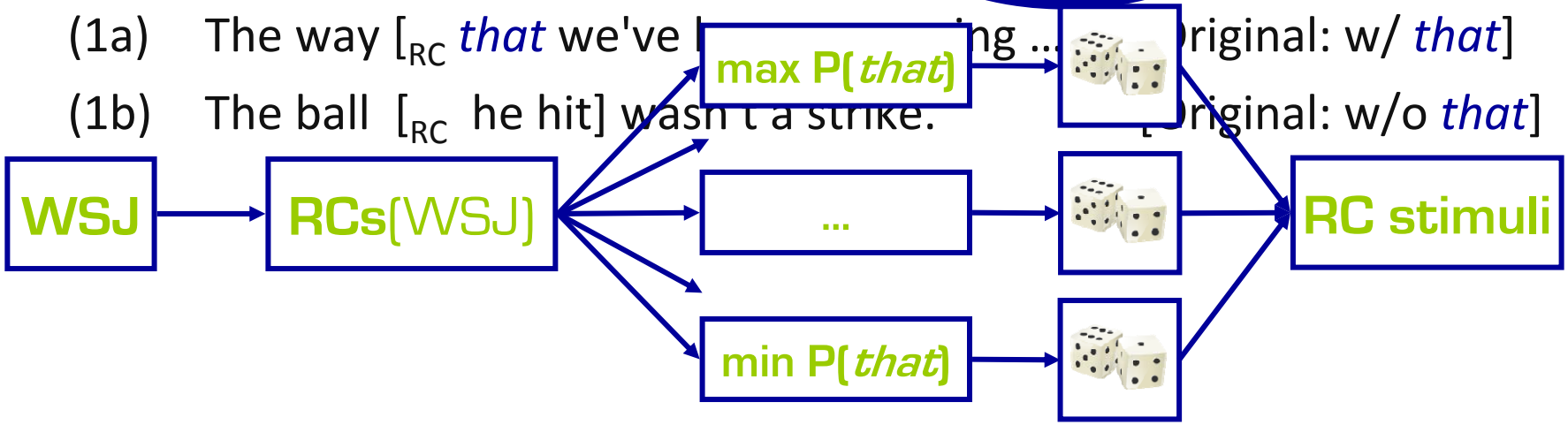
Self-paced reading experiment :: Materials

- 24 RCs sampled from the parse Street Journal (Penn Treebank III release: M tive of the corpus in t
  - 12 of with R
  - 12 without R

Used written corpus because this is a reading time study and because same patterns of *that*-omission have been observed in speech and writing (JaegerWasow05)

(1a) The way [$_{RC}$ *that* we've ng … riginal: w/ *that*]

(1b) The ball [$_{RC}$ he hit] wasn't a strike. riginal: w/o *that*]



Self-paced reading experiment :: Materials

(1a) The way [$_{RC}$ *that* we've been managing … [Original: w/ *that*]

(1b) The ball [$_{RC}$ he hit] wasn't a strike. [Original: w/o *that*]

- For each stimulus, a matched stimulus was created that differed only in that-presence, (2a) for (1a) and (2b) for (1d):

(2a) The way [$_{RC}$ we've been managing … [Original: w/ *that*]

(2b) The ball [$_{RC}$ *that* he hit] wasn't a strike. [Original: w/o *that*]

Self-paced reading experiment :: Materials & Subjects

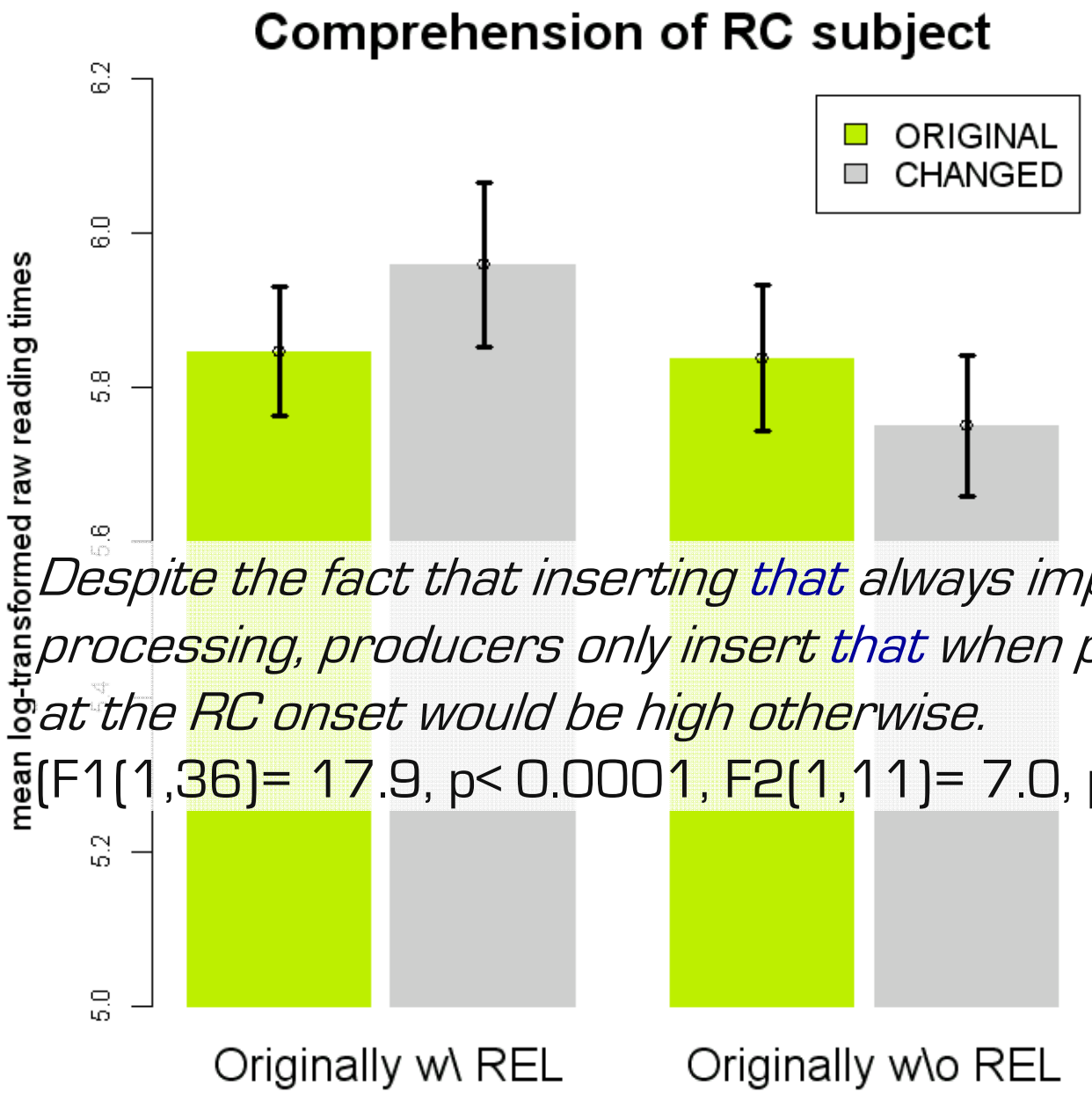
- 48 fillers (24 *wh*-question; 24 other types of stimuli resembling the experimental stimuli in complexity)

- 37 participants

Self-paced reading experiment :: Procedure

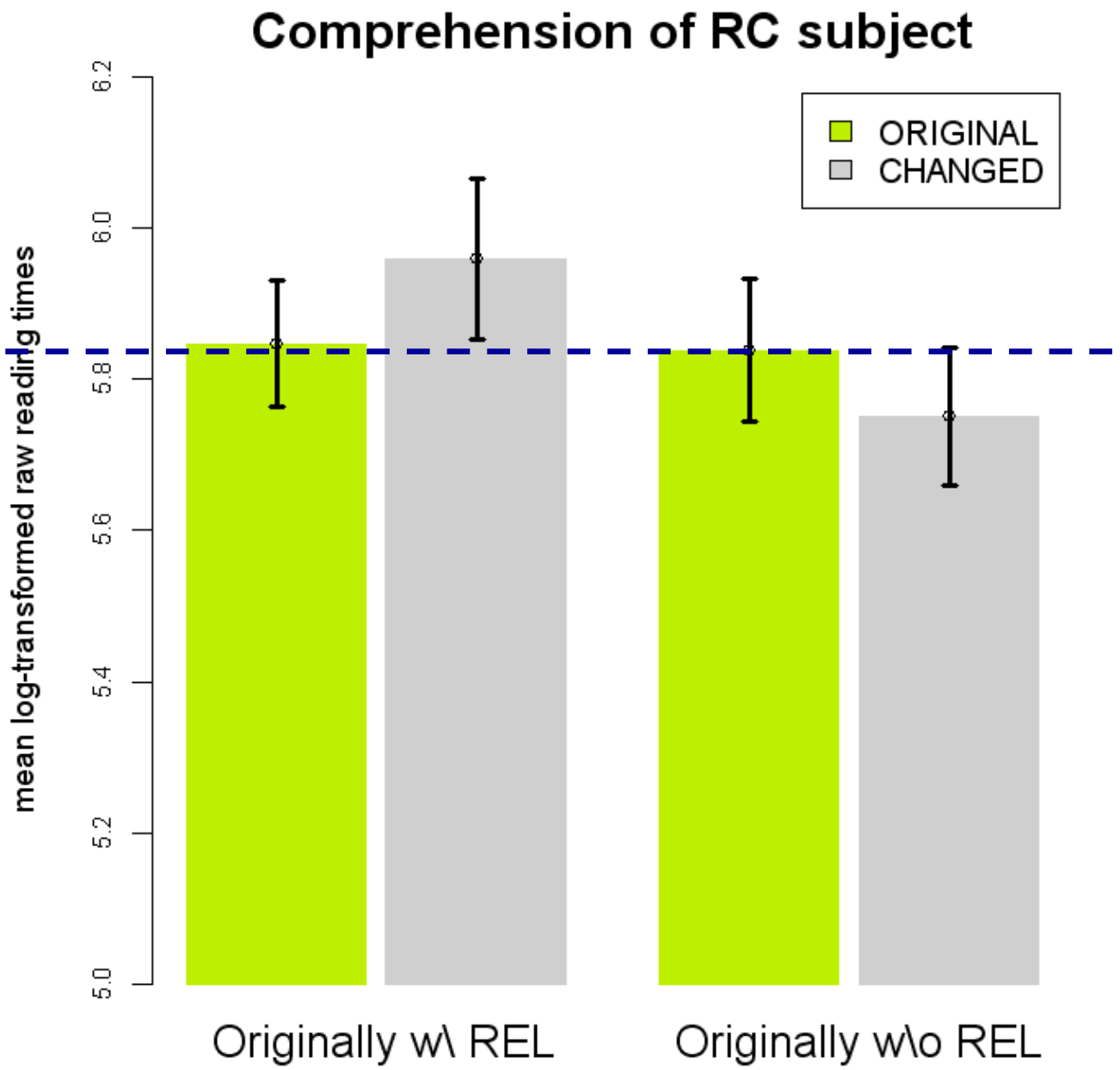
- Stimuli presented word-by-word moving window

```
The way we've been managing
```



Despite the fact that inserting that always improves processing, producers only insert that when processing at the RC onset would be high otherwise.

(F1(1,36)= 17.9, p< 0.0001, F2(1,11)= 7.0, p< 0.023)


Comprehension of RC subject
mean log-transformed raw reading times
6.2
6.0
5.8
5.6
5.4
5.2
5.0
ORIGINAL
CHANGED
Originally w\ REL
Originally w\o REL

What else is out there?

- Treebanks – online and offline
- Automatically parsed corpora
- Non-syntactic online corpora with interesting search interfaces

- Other syntactic search tools: TigerSearch

Available Treebanks

- Offline
  http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml

  or search LDC catalogue: http://www.ldc.upenn.edu/

  - American English (spoken & written)
  - British English (spoken & written)
  - German, Nordic Treebanks
  - Mandarin Chinese, Korean, Japanese
  - Czech, Bulgarian (planned), Russian
  - Standard Arabic

Relative clauses in Mandarin Chinese :: Corpus

- Chinese Treebank (version 5.1), distributed by LDC
  http://www.cis.upenn.edu/~chinese/
  - 18,525 graphs (≈ sentences)
  - 522,683 terminals (≈ words)

- Segmented
- Annotated for:
  - Part-of-speech
  - Syntactic bracketing (traces, fillers, PRO)
  - Functional structure (subject, topic, etc.)
  - In progress: *Semantic role labeling, discourse structure*

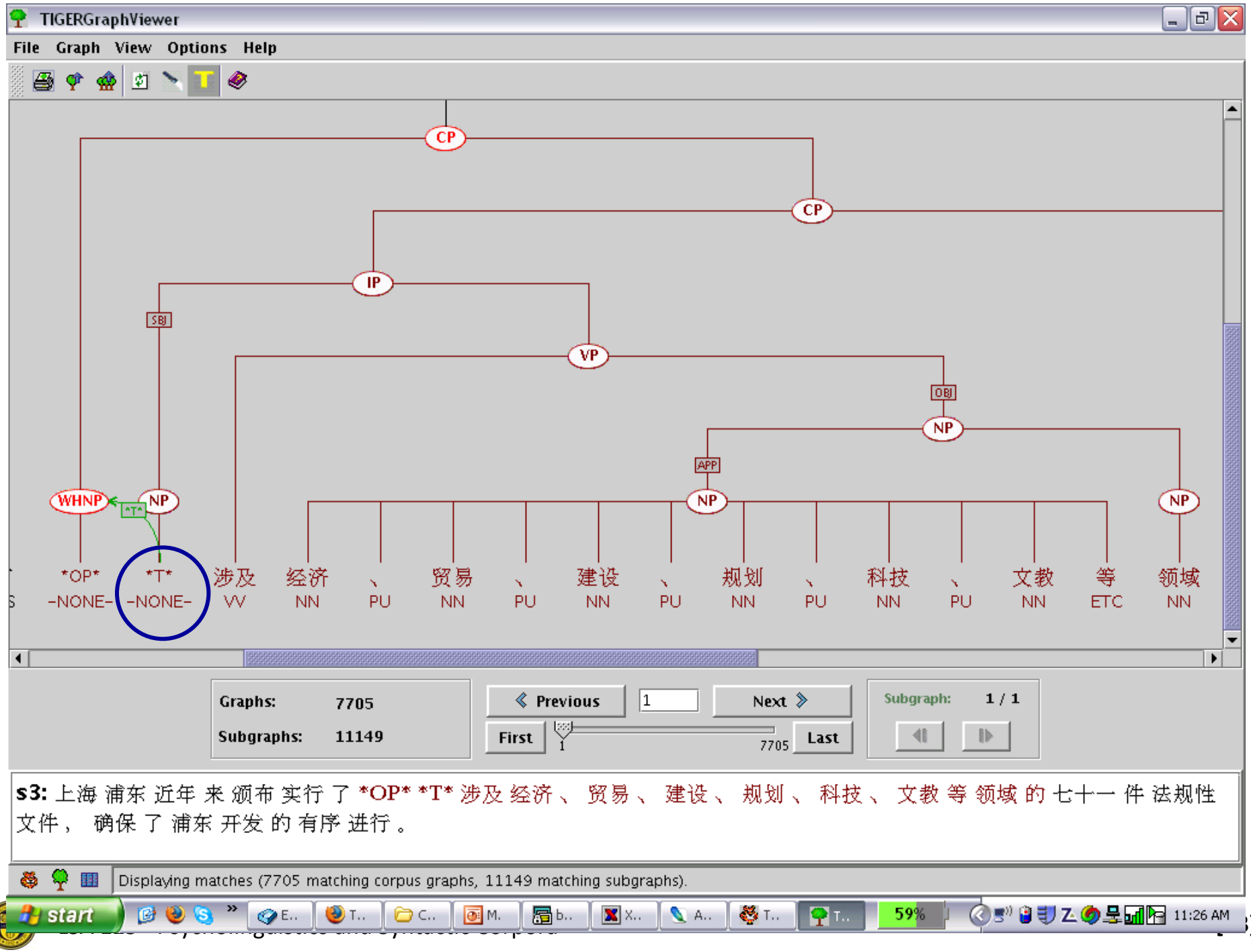
Mandarin *de*-omission

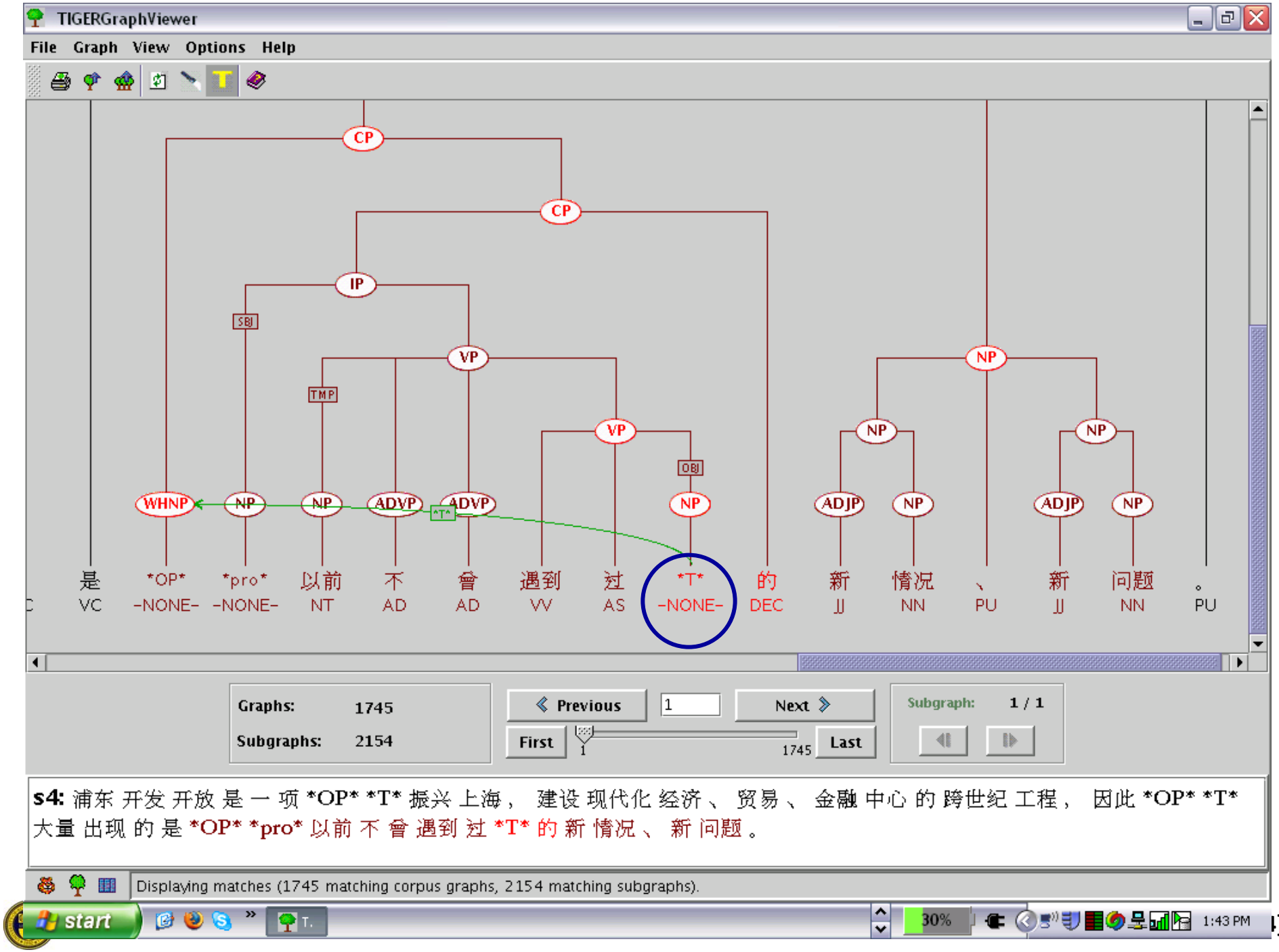
- Mandarin Chinese relative clauses are not always marked with a complementizer (*DE*).
  - $[_{NP}\ [_{RC}$ … *de*] … N]
- Used TigerSearch to extract all 10,546 headed RCs (http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/)

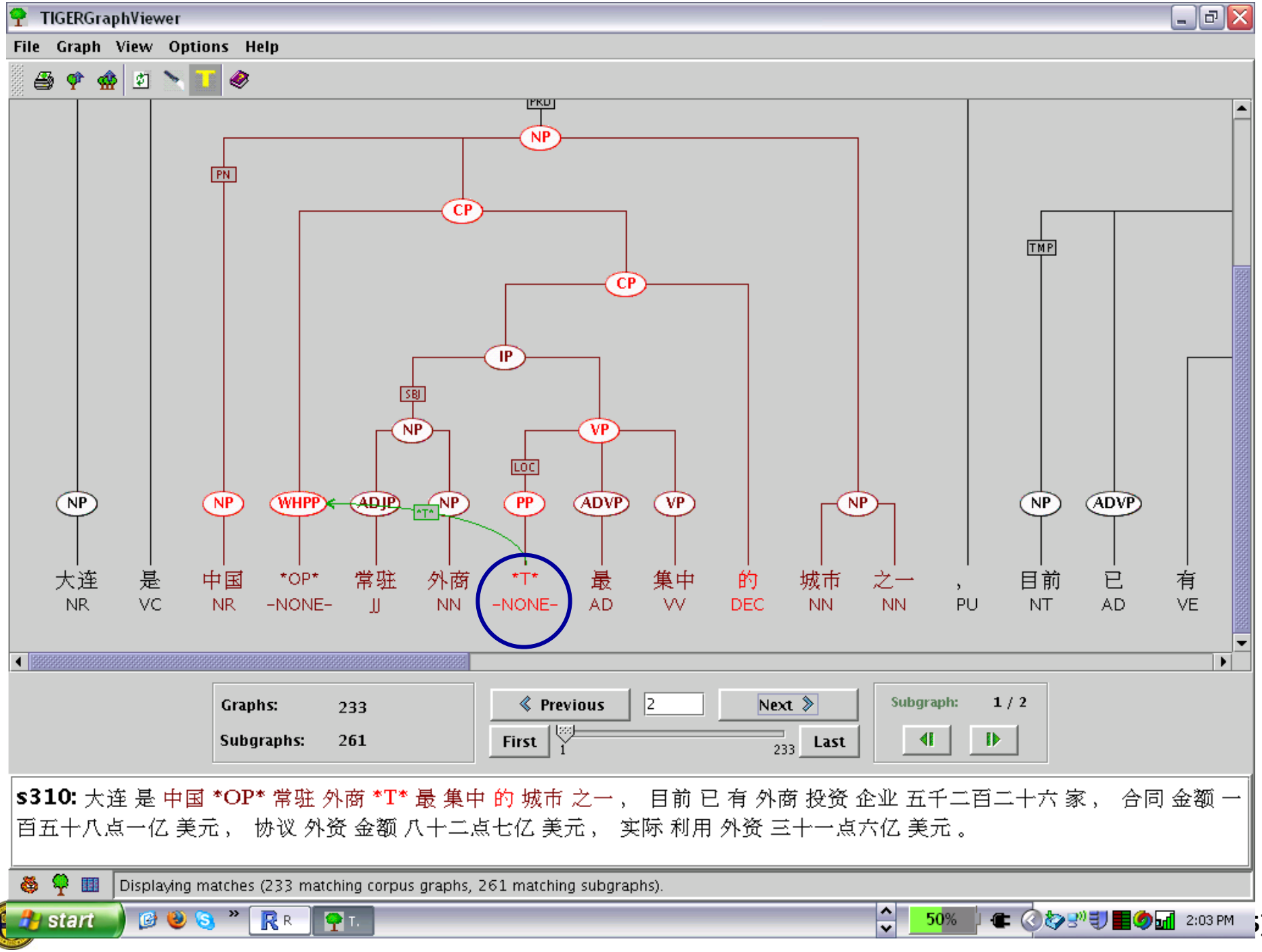
  10,401 have *de*

  145 don't have *de*

  → *omission is rare* (1.4%; cf. 60-80% in Standard American English written corpora)


TIGERGraphViewer
File Graph View Options Help
CP
CP
IP
SBJ
VP
OBJ
NP
APP
NP
WHNP
NP
T
NP
OP -NONE-
T -NONE-
涉及 VV
经济 NN
、 PU
贸易 NN
、 PU
建设 NN
、 PU
规划 NN
、 PU
科技 NN
、 PU
文教 NN
等 ETC
领域 NN
Graphs: 7705
Subgraphs: 11149
Previous
1
Next
First
1
7705
Last
Subgraph: 1 / 1
s3: 上海 浦东 近年 来 颁布 实行 了 *OP* *T* 涉及 经济 、 贸易 、 建设 、 规划 、 科技 、 文教 等 领域 的 七十一 件 法规性 文件 ， 确保 了 浦东 开发 的 有序 进行 。
Displaying matches (7705 matching corpus graphs, 11149 matching subgraphs).
start
59%
11:26 AM


TIGERGraphViewer
File Graph View Options Help
CP
CP
IP
SBJ
VP
TMP
VP
OBJ
WHNP NP NP ADVP ADVP NP
T
NP
NP NP
ADJP NP ADJP NP
是 VC
OP -NONE-
pro -NONE-
以前 NT
不 AD
曾 AD
遇到 VV
过 AS
T -NONE-
的 DEC
新 JJ
情况 NN
、 PU
新 JJ
问题 NN
。 PU
Graphs: 1745
Subgraphs: 2154
Previous
1
Next
First
1
1745
Last
Subgraph: 1 / 1
s4: 浦东 开发 开放 是 一 项 *OP* *T* 振兴 上海， 建设 现代化 经济 、 贸易 、 金融 中心 的 跨世纪 工程， 因此 *OP* *T* 大量 出现 的 是 *OP* *pro* 以前 不 曾 遇到 过 *T* 的 新 情况 、 新 问题 。
Displaying matches (1745 matching corpus graphs, 2154 matching subgraphs).
start
30%
1:43 PM


TIGERGraphViewer
File Graph View Options Help
PRD
NP
PN
CP
CP
TMP
IP
SBJ
NP
VP
LOC
NP
WHPP
ADJP
NP
PP
ADVP
VP
NP
NP
ADVP
大连 是 中国 *OP* 常驻 外商 *T* 最 集中 的 城市 之一 , 目前 已 有
NR VC NR -NONE- JJ NN -NONE- AD VV DEC NN NN PU NT AD VE
Graphs: 233
Subgraphs: 261
Previous
2
Next
First
1
233
Last
Subgraph: 1 / 2
s310: 大连 是 中国 *OP* 常驻 外商 *T* 最 集中 的 城市 之一， 目前 已 有 外商 投资 企业 五千二百二十六 家， 合同 金额 一百五十八点一亿 美元， 协议 外资 金额 八十二点七亿 美元， 实际 利用 外资 三十一点六亿 美元 。
Displaying matches (233 matching corpus graphs, 261 matching subgraphs).
start
50%
2:03 PM

Results :: *DE*-rate by type of extracted element

Extraction type	Total N	Percent *DE*
Topic NP	747	99.2%
Subject NP	8,002	98.9%
Object NP	2,002	97.8%
PP	263	93.2%

× 1.3

× 2.3

× 3.1

Available Treebanks (2)

- Smaller online corpora (e.g. http://corp.hum.sdu.dk/) also available for e.g.
  - Danish
  - Estonian
  - Portuguese
  - French
  - Esperanto

Automatically Parsed Corpora

- Beware of statistical biases, e.g. 100 million words of British English
  - Over exclusion
  - Over inclusion
  - NB: I would like to send you guys a paper some of the coming weeks that talks about this issue.

- Additional annotation may be needed

POS-annotated corpora

- CQP

Online Flat Text Corpora

- Web as corpus: http://www.webcorp.org.uk/
  - Limited regular expression syntax
- BYU collection of online searchable corpora (American and British English, Spanish, Portuguese) -- http://view.byu.edu/
  - Limited regular expression syntax
  - Part of speech
  - Lemmatizer
  - Synonym search
- Automated Google searches: http://www.linguistics.ucla.edu/people/hayes/QueryGoogle/qgapplet.html

More links?

- David Lee's awesome list of links to corpora and tools: http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm
- LDC pages: http://www.ldc.upenn.edu/
- International corpus linguistics email list: http://gandalf.aksis.uib.no/corpora/