Complexity effects in the English comparative alternation: corpus-based evidence

Gero Kunter
Heinrich-Heine-Universität Düsseldorf
gero.kunter@uni-duesseldorf.de

Research supported by
*Deutsche Forschungsgemeinschaft (*grant KU 2896/1-1)

ICAME 36, Trier
May 31, 2015

English adjective paradigms

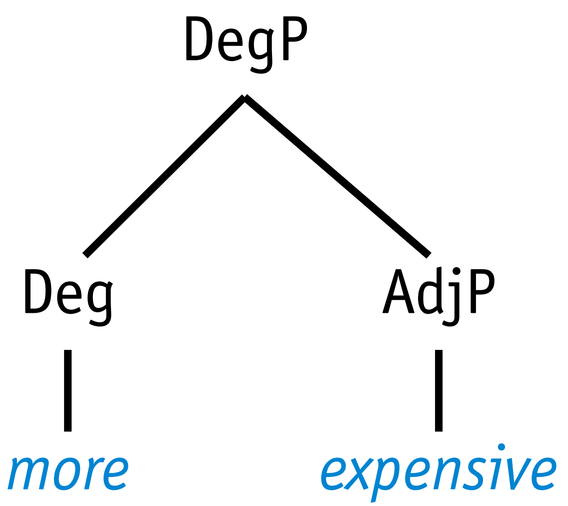
	HIGH	EXPENSIVE
Positive	*high*	*expensive*
Comparative	*high-er*	*more expensive*
Superlative	*high-est*	*most expensive*

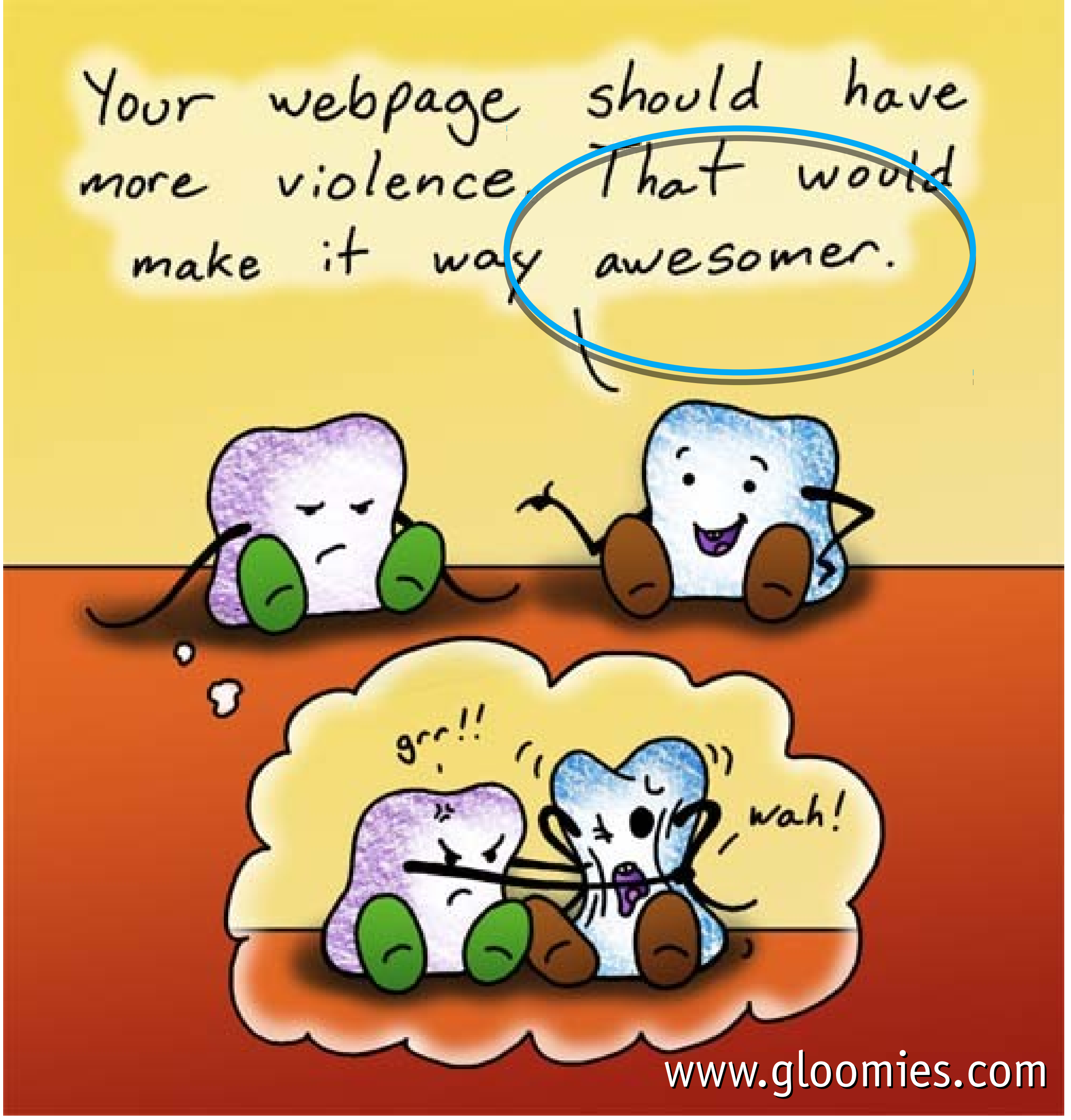
English adjective paradigms

Synthetic
(word)

{high} +	{er} →	*higher*
[Adj]	[Deg]	[Adj, Deg]

Analytic
(phrase)




Your webpage should have more violence. That would make it way awesomer.
grr!!
wah!
www.gloomies.com

www.americantaitai.com

They say happier times are even more happy because you've had some hard times.

(COCA SPOK_2010: NBC_Dateline)

So Stephon Marbury though is back with the Knicks, $180,000 more poor - poorer - more poor. I don't know.

(COCA SPOK_2007: NPR_Park)

Some factors determining comparative alternation

(e.g. Szmrecsanyi 2005, Hilpert 2008, Mondorf 2003, 2009)

Phonological	Number of syllables, final elements of base, stress pattern
Morphological	Number of morphemes, compound adjectives
Lexical	Frequency of adjective, comparative/positive ratio
Syntactic	*to*-infinitive complementation, premodification, predicative vs. attributive position
Semantic	Abstract vs. concrete meaning
Pragmatics	End-weight

More-support

in cognitively more demanding environments which require an increased processing load, language users [...] *tend to compensate for the additional effort by resorting to the analytic form*

(Mondorf 2009: 6)

Hypothesis

More-support acts as a compensatory strategy for increased processing complexity.

Therefore, adjectives that are difficult to process should favour analytic comparatives.

Independent support for *more*-support?

- Probability of analytic comparatives increases if syntactic or semantic complexity is increased (Boyd 2007)
- Speakers prefer analytic comparatives with cognitively complex adjectives in sentence completion task (Kunter 2015)

This paper

Do cognitively complex adjectives occur more frequently with analytic comparatives also in corpus data?

Data

Cognitive complexity data

Assumption

Words that are cognitively complex have longer processing times

English Lexicon Project (Balota et al. 2007)

- ~40.000 English words
- mean reaction times from lexical decision tasks
- 800 participants, 34 observations per item

Corpus frequencies

Contemporary Corpus of American English (COCA, Davies 2008-)

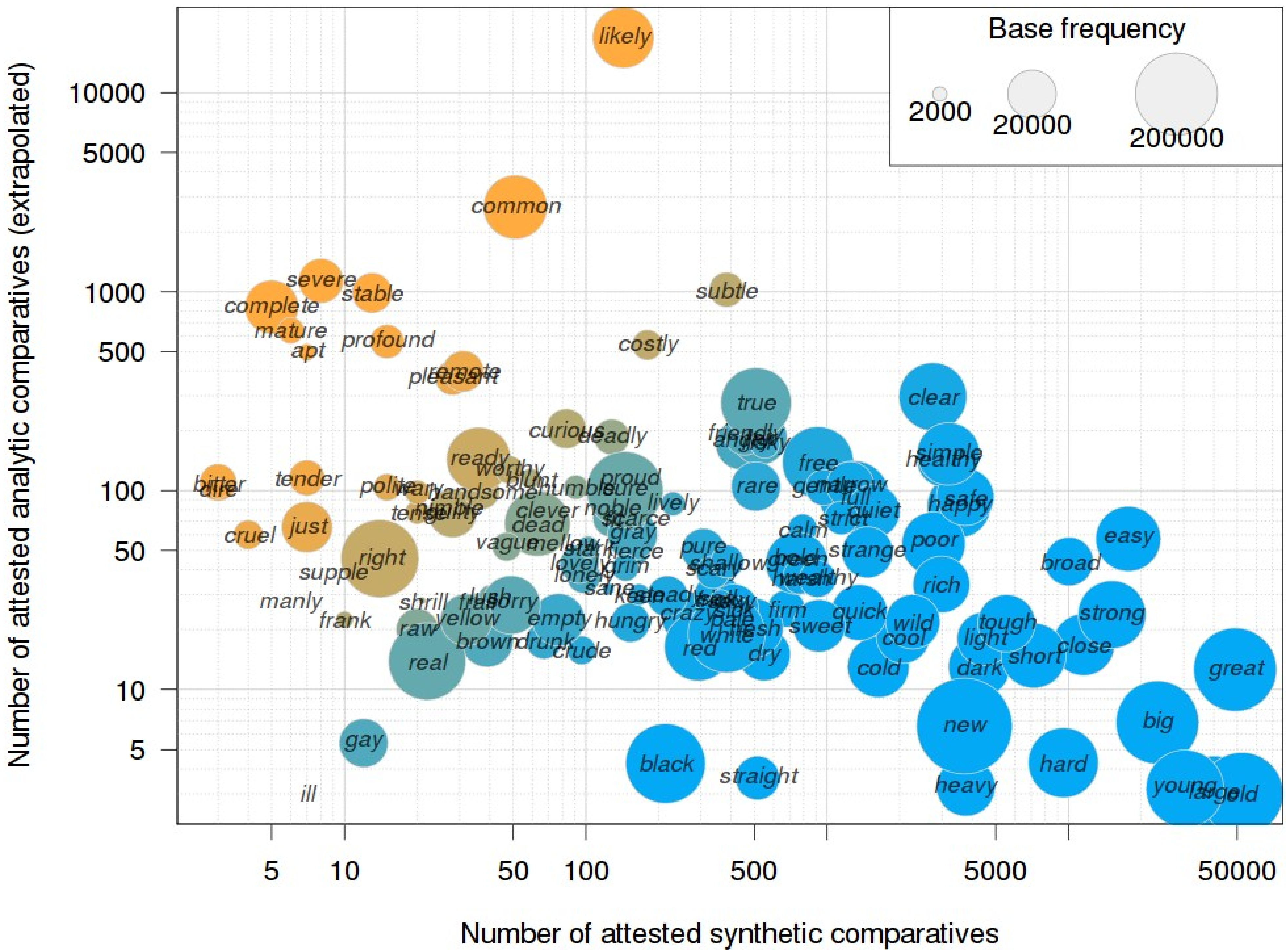
- 450 million words
- written and spoken texts
- time range: 1990-2012
- part-of-speech tags (uncorrected)
- queries: [*deadly*].[J*], *more deadly*.[J*]

Adjective selection

All adjectives listed in the CELEX lexical database (Baayen et al. 1995) that were…

- attested at least 3 times as synthetic comparatives in COCA
- attested at least 3 times as analytic comparatives in COCA
- also listed in the English Lexicon Project

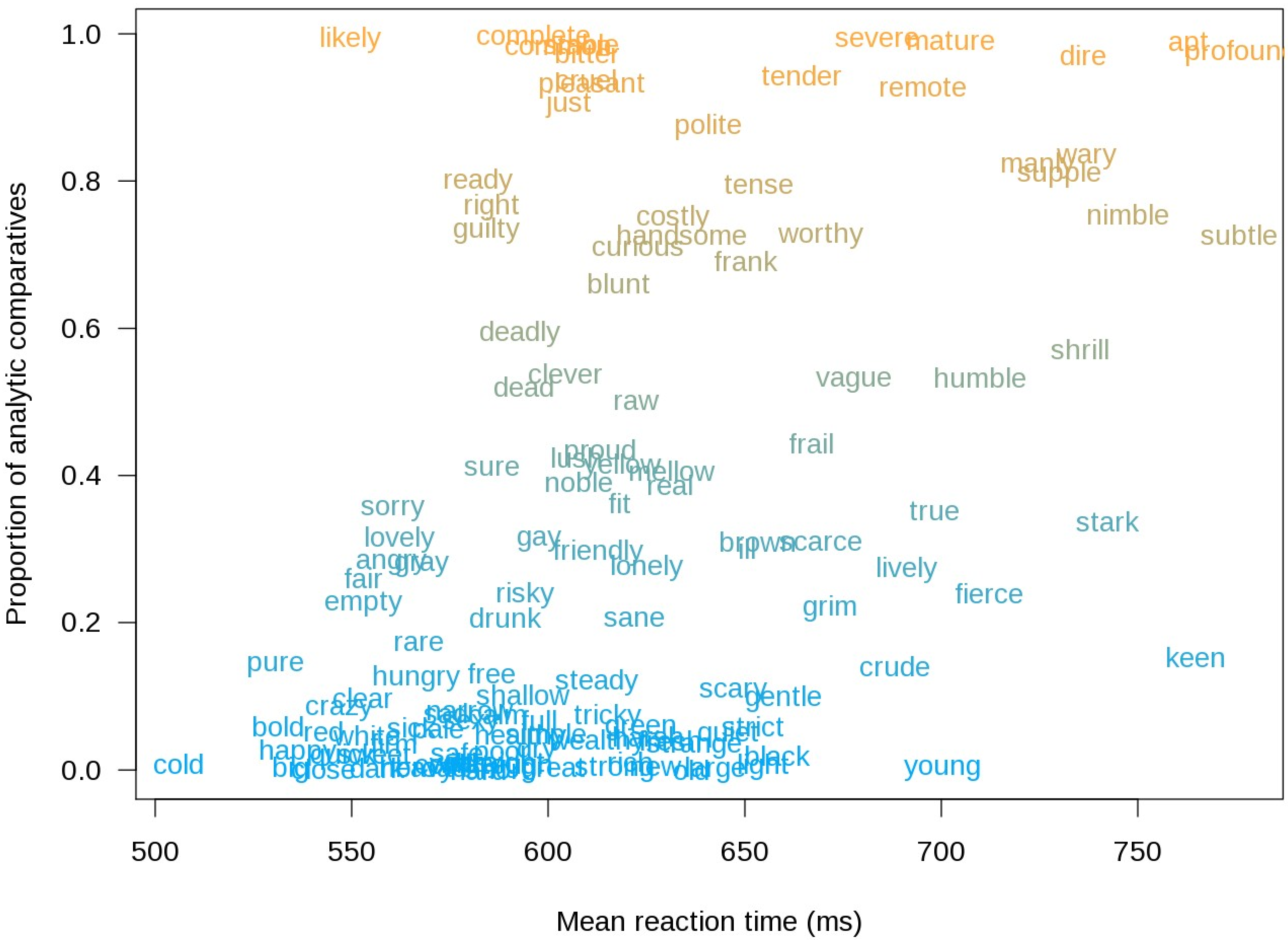
128 adjective types



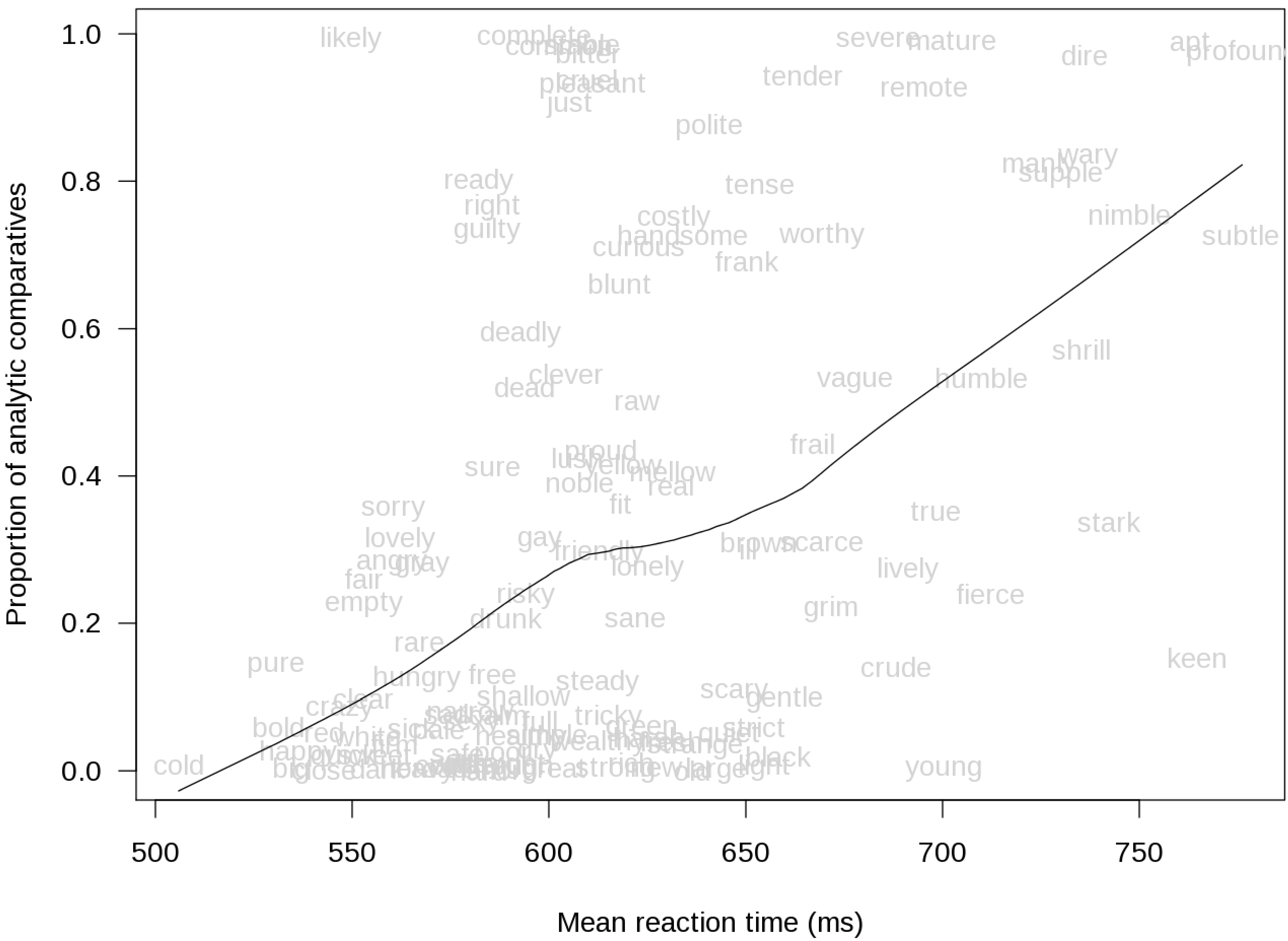
Base frequency
2000
20000
200000
Number of attested analytic comparatives (extrapolated)
Number of attested synthetic comparatives
10000
5000
1000
500
100
50
10
5
5
10
50
100
500
5000
50000
likely
common
severe
stable
complete
mature
apt
profound
subtle
costly
pleasant
remote
true
clear
curious
deadly
friendly
simple
healthy
ready
worthy
blunt
proud
free
bitter
dire
tender
polite
handsome
humble
rare
narrow
full
safe
happy
noble
lively
quiet
strict
clever
scarce
dead
calm
poor
easy
cruel
just
vague
mellow
gray
pure
strange
broad
right
supple
fierce
grim
lonely
lovely
scary
wealthy
rich
sane
manly
shrill
yellow
empty
hungry
firm
quick
strong
frank
raw
brown
drunk
crude
crazy
white
fresh
sweet
wild
tough
light
cool
close
red
dry
real
cold
dark
short
great
gay
new
big
black
straight
hard
heavy
young
ill

Examples

Adjective	Reaction time (ms)	Base frequency	Synthetic frequency	Analytic frequency	Proportion of analytic comparatives
new	627	491932	3709	7	0.00
risky	594	5558	552	172	0.24
deadly	592	8411	128	186	0.59
nimble	747	767	27	83	0.75
remote	695	12076	31	398	0.93



Proportion of analytic comparatives
Mean reaction time (ms)
1.0
0.8
0.6
0.4
0.2
0.0
500
550
600
650
700
750
likely
severe
mature
dire
apt
tender
remote
just
polite
wary
supple
ready
right
guilty
tense
costly
handsome
curious
worthy
frank
nimble
subtle
blunt
deadly
shrill
clever
dead
raw
vague
humble
frail
sure
proud
mellow
noble
real
fit
sorry
lovely
true
stark
gay
friendly
lonely
scarce
ill
lively
fair
empty
risky
drunk
sane
grim
fierce
rare
pure
hungry
free
steady
crude
keen
shallow
scary
gentle
crazy
bold
tricky
strict
cold
young



Proportion of analytic comparatives
Mean reaction time (ms)
0.0
0.2
0.4
0.6
0.8
1.0
500
550
600
650
700
750
likely
complete
bitter
severe
mature
dire
apt
tender
remote
pleasant
just
polite
wary
supple
ready
tense
right
costly
nimble
guilty
handsome
worthy
subtle
curious
frank
blunt
deadly
shrill
clever
vague
humble
dead
raw
frail
proud
sure
mellow
noble
real
fit
sorry
true
stark
lovely
gay
friendly
scarce
lonely
lively
fair
risky
fierce
empty
sane
grim
drunk
rare
pure
keen
crude
hungry
free
steady
scary
gentle
shallow
clear
tricky
strict
bold
cold
young

Analysis

Statistical model

Analysis	Beta regression model (Grün et al. 2012)
Dependent variable	Proportion of analytic comparatives
Main predictor	Mean lexical decision times

Results

(mean component of beta regression)

	Est.	Std. Err.	*z*	P (>|*z*|)
(Intercept)	-75.288	15.921	-4.729	0.000
Final elements of base (reference level: /i/)				
/CC/	1.673	0.404	4.156	0.000
/l/	0.964	0.[illegible]	[illegible]	[illegible]
/li/	0.731	0.		
/r/	1.673	0.		
other	0.528	0.		
Metrical structure (reference level: S)				
Sw	1.244	0.		
wS	2.485	0.		
Log Base freqency	-0.413	0.088	-4.656	0.000
Logit Comparative-positive ratio	-0.621	0.071	-8.738	0.000
Log Reaction Time	5.612	1.169	4.802	0.000

Log-likelihood: 142.800 on 22 Df

Pseudo R^2: 0.709

In line with previous research (e.g. Hilpert 2008, Mondorf 2009)

Results

(mean component of beta regression)

	Est.	Std. Err.	z	P (>|z|)
(Intercept)	-75.288	15.921	-4.729	0.000
Final elements of base (reference level: /i/)				
/CC/	1.673	0.404	4.156	0.000
/l/	0.964	0.372	2.595	0.009
/li/	0.731	0.436	1.676	0.094
/r/	1.673	0.350	4.720	0.000
other	0.528	0.296	1.785	0.074
Metrical structure (reference level: S)				
Sw	1.244	0.293	4.989	0.000
wS	2.485	0.716	3.469	0.000
Log Base freqency	-0.413	0.088	-4.656	0.000
Logit Comparative-positive ratio	-0.621	0.071	-8.738	0.000
Log Reaction Time	5.612	1.169	4.802	0.000

Log-likelihood:	142.800	on 22 Df
Pseudo R^2:	0.709	

Correlation between Base Frequency and RT

- More frequent words are easier to process than less frequent words (e.g. Taft 1979)
- Notable correlation between Base Frequency and Reaction Time in the data (r_S = -0.5, $p < 0.001$)

Is the effect of Reaction Time just a statistical artifact?

Probably not.

Independent contribution of RT

Model without Reaction Time

Only small changes in the remaining coefficients
($r_S = 0.96$, $p < 0.001$)

Model with Reaction Time residualized against Base Frequency

No changes in the remaining coefficients
($r_S = 0.99$, $p < 0.001$)

Model with only Reaction Time as predictor

Coefficient for Reaction Time like in full model
(5.490 vs. 5.612, Pseudo $R^2 = 0.16$)

Discussion and conclusion

more-support

Hypothesis
More-support acts as a compensatory strategy for increased processing complexity.

Therefore, adjectives that are difficult to process should favour analytic comparatives.

Discussion

- Reaction time (as proxy of cognitive complexity) is a significant predictor of analyticity
- Effect of Reaction time is independent of other variables
- Other predictors behave as expected
- **Cognitively complex adjectives occur with a higher proportion of analytic comparatives in a corpus.**

more-support: overly reductive?

Mondorf (2003, 2009):
English comparative alternation is (largely) driven by complexity effects

Hilpert (2008):
more-support hypothesis may be overly reductive – not all effects can be attributed to complexity (e.g. phonologcial effect of final segment)

This paper
Processing complexity is an independent factor that affects comparative alternation

Conclusion

Empirical support for *more*-support

Speakers of English are more likely to use analytic comparatives with cognitively complex adjectives.

Corpora and psycholinguistic effects

Corpus data and lexical databases can be used successfully to investigate complexity effects in methodologically principled ways.

References

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. *The CELEX lexical database*. Philadelphia, PA: Linguistic Data Consortium.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson & Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods* 39(3). 445–459.

Boyd, Jeremy. 2007. *Comparatively speaking. A psycholinguistic study of optionality in grammar*. Ph.D. dissertation, University of California, San Diego.

Davies, Mark. 2008--. *The Corpus of Contemporary American English (COCA): 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/: .

Grün, Bettina, Ioannis Kosmidis & Achim Zeileis. 2012. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software* 48(11). 1–25.

Hilpert, Martin. 2008. The English comparative. Language structure and language use. *English Language and Linguistics* 12(3). 395–417.

Kunter, Gero. 2015. *Effects of processing complexity in perception and production. The case of English comparative alternation*. Pirrelli, Vito, Claudia Marzi and Marcello Ferro eds., *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon*, Pisa, Italy. 32–36.

Mondorf, Britta. 2003. *Support for more-support*. In Rohdenburg, Günter & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 251–304. Berlin: Mouton de Gruyter.

Mondorf, Britta. 2009. *More support for more-support*. Amsterdam: John Benjamins.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150.

Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory & Cognition* 7(4). 263–272.