Using Semantic Vector Space Models to investigate lexical replacement – a corpus based study of ongoing changes in intensifier systems

Dr. Martin Schweinberger
www.martinschweinberger.de

Habilitation (in progress)
Acquisition, Variation, and Diachronic Development of Intensification in English

(1) yeah... just it would make it so awkward eh you know (ICE-NZ S1A-001:1$M)

(2) um... sara's got a really nice sleeveless green... you know coat jacket (ICE-NZ S1A-002:1$Q)

(3) she was a very nervous sort of a woman (ICE-NZ S1A-018:1$A)

Intensification

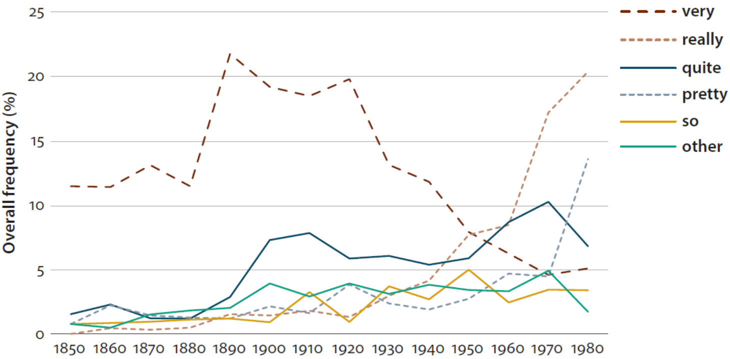
Intensification is related to the semantic category of *degree* (degree adverbs) and ranges between very low intensity (downtoning) and very high (amplifiers) (Quirk et al. 1985: 589–590).

- Amplifiers (Tagliamonte 2008)
  - Maximizers (e.g. *completely*)
  - Boosters (e.g. *very much*)
- Downtoners
  - Approximators (e.g. *almost*)
  - Compromisers (e.g. *more or less*)
  - Diminishers (e.g. *partly*)
  - Minimizers (e.g. *hardly*)

Previous Research

- Intensification...
  - major area of grammatical change (cf. Brinton and Arnovick 2006: 441)
  - crucial for the “social and emotional expression of speakers” (Ito and Tagliamonte 2003: 258)
  - teenage talk and young(ish) speakers (Bauer and Bauer 2002; Macaulay 2006)
  - female speakers (Tagliamonte 2006, 2008; D’Arcy 2015)
  - recently amplifier-adjective bigrams have come more into focus (e.g. Wagner 2017; Schweinberger 2017)
  - Intensifying *really* replaces *very* (lexical replacement) (cf. D’Arcy 2015; Ito and Tagliamonte 2003; Tagliamonte 2005, 2008)

Previous study of intensification in New Zealand English (D'Arcy 2015)



(D'Arcy 2015: 468)

Previous study of intensification in New Zealand English

- “Precisely this kind of bleaching has regularly been invoked as an explanation for the recycling that characterizes intensification [. . .]: *very* lost its pragmatic strength and *really* was recruited in its place. If this is correct, then the mechanism of change is arguably lexical replacement[. . .].”

Q_1

How does the lexical replacement
of *very* by *really* proceed?

→ Is it lexical replacement (semantic similarity)
or rather two processes
(decrease of *very*; increase of *really*)
that are linked but remain
separate developments (semantic dissimilarity)

Data Processing

- Split spoken ICE NZ data into utterances
- Removal of meta information
- Part–of-speech tagging
- Retrieving adjectives (PoS–tag JJ)
- Determining whether adjective is preceded by an intensifying adverb (PoS–tag RB)

Data Processing

- Removal of
    - negated adjectives
    - comparative and superlative forms
    - non–intensifiable forms
      (categorical, e.g. nationalities | locations: *asian*, *Asia*)
- Manual cross–evaluation of automated classification
- Adding speaker information (age, sex, etc.).

Data Summary: ICE-NZ data

Age	**Sex**	**Speakers** (N)	**Adj.** (N)	**Int.** (N)	**Int.** (%)
16-24	female	39	1102	140	12.7
16-24	male	29	811	81	10.0
25-39	female	23	629	65	10.3
25-39	male	16	481	35	7.3
40-49	female	16	509	60	11.8
40-49	male	9	172	7	4.1
50+	female	7	259	27	10.4
50+	male	6	236	25	10.6
Total		145	4199	440	10.5

Data Summary: Intensifiers ICE-NZ

Intensifier	**N**	**%**	**Int.** (%)
∅ Intensification	3759	89.52	
really	150	3.57	34.09
very	96	2.29	21.82
so	66	1.57	15.00
too	34	0.81	7.73
pretty	29	0.69	6.59
real	18	0.43	4.09
well	7	0.17	1.59
absolutely, right, totally	5	0.36	3.42
bloody	4	0.10	0.91
crazy, particularly	2	0.10	0.90
actually, badly, completely, definitely, dreadfully, enormously, entirely, excruciatingly, fucking, fully, horrendously, incredibly, obviously, purely, shocking, true, wicked	1	0.34	3.91
Total	4199	10.48	100

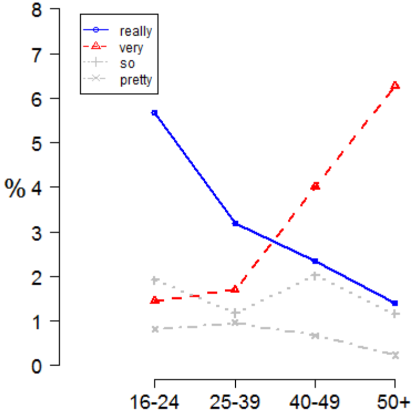


Figure: intensifiers : age (ICE NZ)

Comparison: Intensifiers ICE-Ireland

- To analyze the development of the intensifier system in NZE, I chose to compare to the intensifier system of IrE.

Q_2

Do we find the same trajectories in IrE that we observe in NZE??

Data Summary: ICE-Ireland data

Age	**Sex**	**Speakers** (N)	**Adj.** (N)	**Int.** (N)	**Int.** (%)
19-25	female	72	1072	96	8.22
19-25	male	8	182	8	4.21
26-33	female	51	790	89	10.13
26-33	male	4	48	5	9.43
34-49	female	8	145	28	16.18
34-49	male	6	187	18	8.78
50+	female	14	238	18	7.03
50+	male	9	133	7	5
Total		172	2795	269	9.6

Data Summary: Intensifiers ICE-Ireland

Intensifier	**N**	**%**	**Int.** (%)
∅ Intensification	2526	90.38	
very	78	2.79	29.00
really	58	2.08	21.56
so	41	1.47	15.24
too	28	1	10.41
quite	21	0.75	7.81
absolutely	8	0.29	2.97
real	7	0.25	2.60
fairly, pretty	4	0.28	1.49
awfully, bloody, exactly, pure, totally	2	0.35	0.74
completely, extra, extremely, fierce, mega, perfectly, proper, severely, terribly, truly	1	0.4	0.37
Total	2795	9.62	100

Intensifiers across Age Cohorts

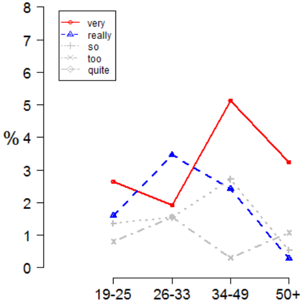


Figure: intensifiers : age (ICE IRE)

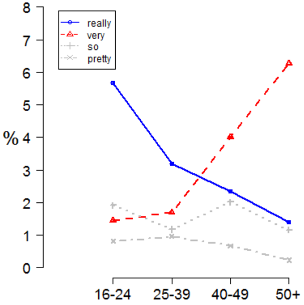


Figure: intensifiers : age (ICE NZ)

H_1

The regional differences in the developments of the intensifier systems reflect differences in collocation patterns (semantic dissimilarity).

→ “Lexical replacement” in NZE should be accompanied by a high degree of semantic similarity while there should less semantic similarity in IrE (no replacement).

Data processing for semantic analysis

Data editing

- Subset data: only speakers between 25/26 and 49 are considered (locus of change)
- Remove all biodata
- Tabulate intensifiers against adjectives (co-occurrence matrix)
- Convert counts into a binary variable (0=no co-occurrence; 1= co-occurrence) to counter frequency effects

Semantic Vector Space Models

- Distributional approach to semantics: “You shall know a word by the company it keeps”(Firth 1957).
- Words that share collocates are semantically similar (Stefanowitsch 2010: 368–370).
- Semantic similarity can thus be measured in collocate frequency.
- Collocate frequency is captured by cross-tabulating frequency counts.

	extremely	pretty	real	really	so	totally	very
basic	0	1	1	0	0	0	1
black	0	0	0	0	0	1	1
brave	0	0	0	0	0	0	1
busy	0	1	0	1	1	0	1
careful	0	0	0	1	0	0	1
clear	0	0	1	0	0	0	1

Semantic Vector Space Models

The analysis follows Levshina (2015)

- Each intensifier has an idiosyncratic vector of counts.
- The following steps are performed in the analysis
  1. Based on the vectors, we weight the data by calculating (a) expected co-occurrence scores and then (b) (positive) pointwise mutual information scores (PPMI)
  2. Based on the PPMI we calculate similarity scores between the resulting PPMI vectors with the help of the cosine measure.
  3. Semantic similarity of intensifiers can be visualized using dendrograms (clusters based on weighted frequencies) or networks (shared collocates based on non-weight type co-occurrence).

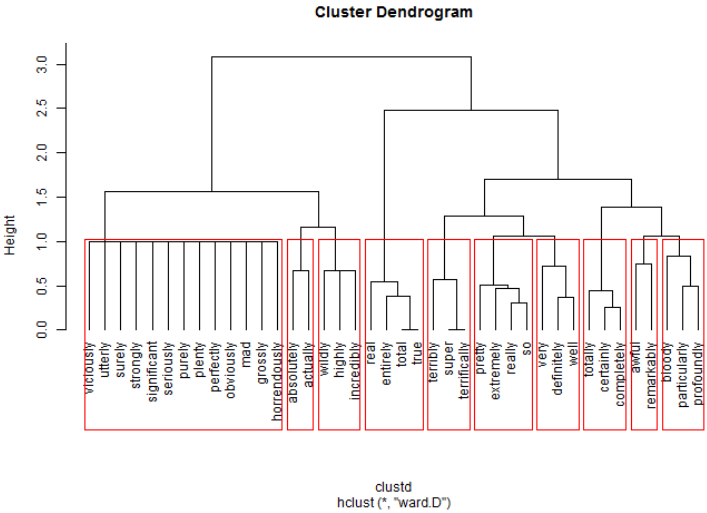
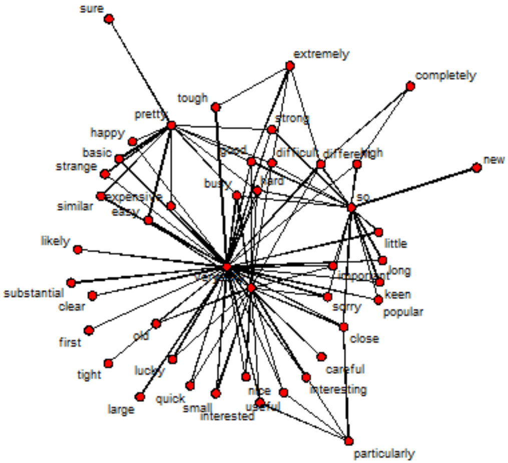


Figure: Clustering of intensifiers by adjective collocates (NZE)


sure
extremely
completely
tough
pretty
strong
happy
basic
good
difficult
different
high
new
strange
busy
hard
similar
expensive
so
easy
likely
little
long
very
important
substantial
keen
clear
sorry
popular
first
old
close
tight
lucky
careful
interesting
large
quick
nice
small
useful
interested
particularly

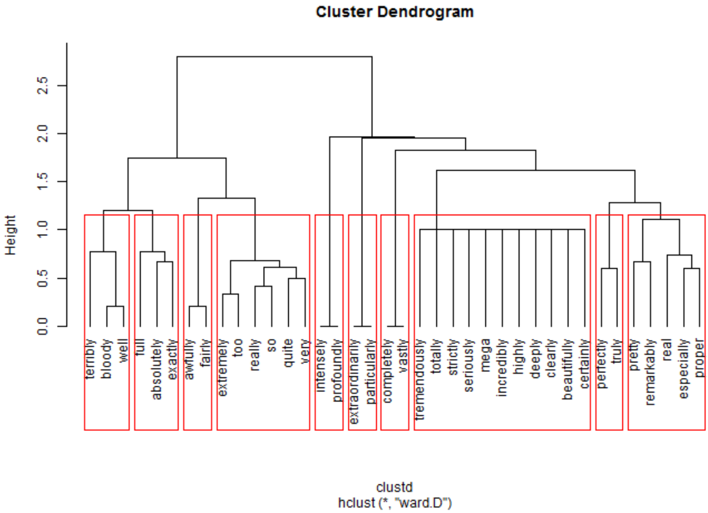
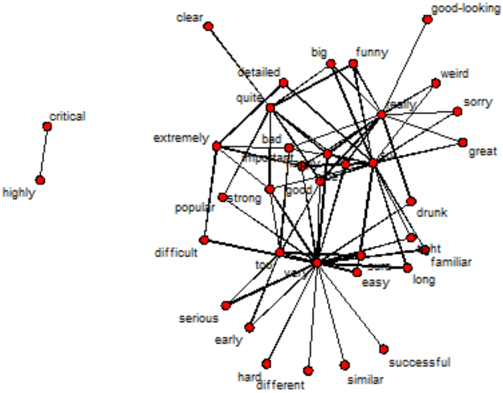


Figure: Clustering of intensifiers by adjective collocates (IrE)


good-looking
clear
big
funny
weird
detailed
quite
really
sorry
critical
extremely
bad
great
important
happy
so
nice
good
strong
highly
popular
drunk
right
difficult
familiar
too
very
sure
easy
long
serious
early
successful
hard
different
similar

Summary, Problems & Outlook

Summary

- NZE
  - network analysis: confirmed semantic similarity (→ confirms D'Arcy's (2015) claim that we are dealing with lexical replacement)
  - cluster analysis: also suggests semantic similarity (but less so than the network analysis)
- IrE
  - network analysis: substantially fewer shared collocates between *very* and *really* compared to NZE (→ no lexical replacement)
  - cluster analysis: suggests more competitors for dominance (not only *really*)

Discussion

- Reasons for different developments: higher semantic similarity of *really* and *very* in NZE compared to IrE (*really* shares collocates with *very* in NZE but less so in IrE) plus there are more competitors for dominance in IrE
- Semantic similarity - understood as similarity in collocational profiles - may be a precondition for lexical replacement
- Another factor facilitating lexical replacement could be absence of semantically similar rival variants
- Hypothesis: once *really* has become more semantically similar, it will outperform the rival variants in IrE and show a similar trajectory than the on ewe saw in NZE

Problems

- small data sets(!)
- problematic (non-)intensifiers are still present in the data (*too*, *quite*)
- no differentiation between boosters and maximizers
- disregard of (semantic) adjective types and constraints on intensifier-adjective co-occurrence (SVM for adjectives?)
- disregard of positioning (predicative : attributive)

Outlook

- Semantic Vector Space Models could enhance the classification of semantic variables (more objective means of classifying adjectives and intensifiers)
- more recent data might show that *really* has become dominant in IrE as well - check with GloWbE data
- enlarge data base (BYU corpora) to investigate semantic similarity

Thank you so, really, very much!

Bauer, L. and W. Bauer (2002). Adjective boosters in the english of young new zealanders. *Journal of English Linguistics 30*, 244–257.

Brinton, L. J. and L. K. Arnovick (2006). *The English Language: A Linguistic History*. Oxford: Oxford University Press.

D'Arcy, A. F. (2015). Stability, stasis and change – the longue durée of intensification. *Diachronica 32*(04), 449–493.

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930-1955*, Volume Studies in linguistic analysis. Basil Blackwell.

Ito, R. and S. Tagliamonte (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers. *Language in Society 32*, 257—279.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing Company.

Macaulay, R. (2006). Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change 18*, 267—283.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London & New York: Longman.

Schweinberger, M. (2017). Using intensifier-adjective bi-grams to investigate mechanisms of change. Paper presented at ICAME38. Prague, 27/5/2017.

Stefanowitsch, A. (2010). Empirical cognitive semantics: Some thoughts. In D. Glynn and K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches*, Volume 46, pp. 355–380. Walter de Gruyter.

Tagliamonte, S. (2005). So who? like how? just what?: Discourse markers in the conversations of young canadians. *Journal of Pragmatics 37*(11), 1896–1915.

Tagliamonte, S. (2006). "so cool, right?": Canadian english entering the 21st century. *The Canadian Journal of Linguistics/La revue canadienne de linguistique 51*(2), 309–331.

Tagliamonte, S. (2008). So different and pretty cool! recycling intensifiers in toronto, canada. *English Language and Linguistics 12*(2), 361–394.

Wagner, S. (2017). Amplifier–adjective 2-grams world-wide: focus on pretty. Paper presneted at ICAME 37. Charles University Prague, 27/5/2017.

Using Semantic Vector Space Models to investigate lexical replacement – a corpus based study of ongoing changes in intensifier systems

Dr. Martin Schweinberger
www.martinschweinberger.de