# Modeling intensification in New Zealand English data

Martin Schweinberger
www.martinschweinberger.de

Habilitation project
*Acquisition, Variation, and Diachronic Development of Intensification in English*

- ▶ synchronic quantitative corpus–based study
- ▶ adjectival intensification in New Zealand English (NZE)
- ▶ based on the New Zealand component of the *International Corpus of English* (ICE-NZ)

Setting the Stage

Data and Methodology

Results

Summary

Problems & Outlook

References

# Intensification

(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*)
- Downloners
    - 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)

# Previous Research

- ► Ongoing changes are accompanied by . . .
    - ► gender and age differences (apparent time construct)
    - ► differences in the syntactic function (predicative vs attributive)
    - ► the semantic type of the modified adjective
    - ► emotional value of the modified adjective (emotional vs non–emotional)
- ► 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)
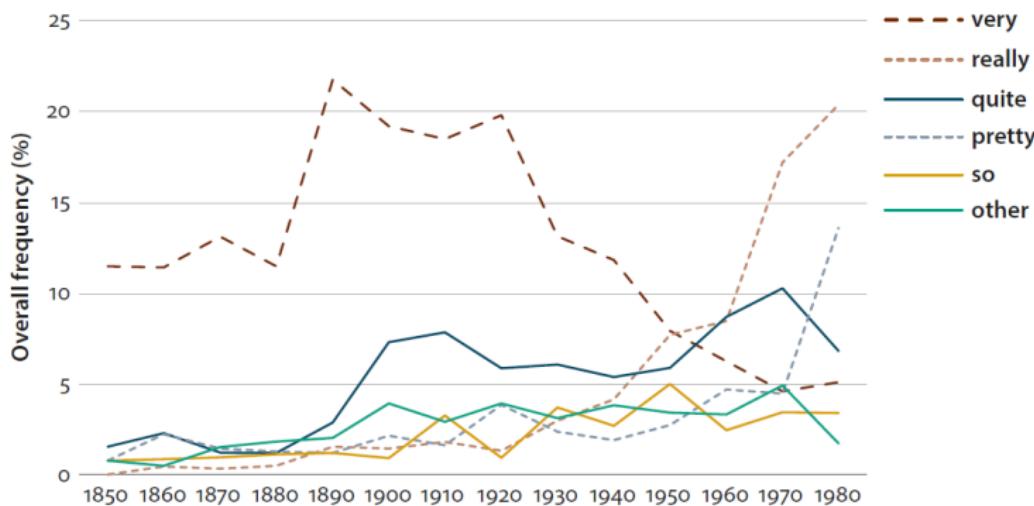


(D'Arcy 2015: 468)

# Research Question

$Q_1$

Is the NZE Intensifier system currently undergoing change?

ICE New Zealand

# ICE New Zealand

New Zealand component of the *International Corpus of English* (Bauer et al. 1999)

- ► released in 1999 (*The Victoria University of Wellington*)
- ► consists of one million words (600,000 spoken and 400,000 written)
- ► representing diverse spoken and written text types
- ► here only private dialogues (200,000 words)

DATA PROCESSING

# Data Processing

- ▶ Split spoken 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

- ▶ Determining the syntactic type of adjective (predicative vs attributive (if followed by NN* tag))
- ▶ Removal of
    - ▶ negated adjectives
    - ▶ comparative and superlative forms
    - ▶ non–intensifiable forms
      (categorical, e.g. nationalities | locations: *asian*, *Asia*)
- ▶ Sentiment Analysis
  determines the emotional value of adjectives based on the *Word-Emotion Association Lexicon* (Mohammad and Turney 2013)
- ▶ Manual cross–evaluation of automated classification
- ▶ Adding speaker information (age, sex, etc.).

Data Summary

# 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 |

RESULTS
—
VISUALIZATION

# Intensifiers across Age Cohorts

Percent of VERY of all Intensifiers
in Private Dialogue
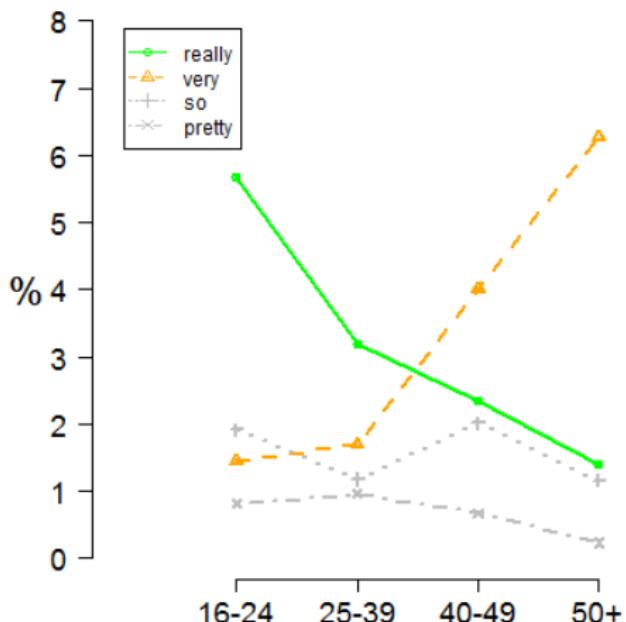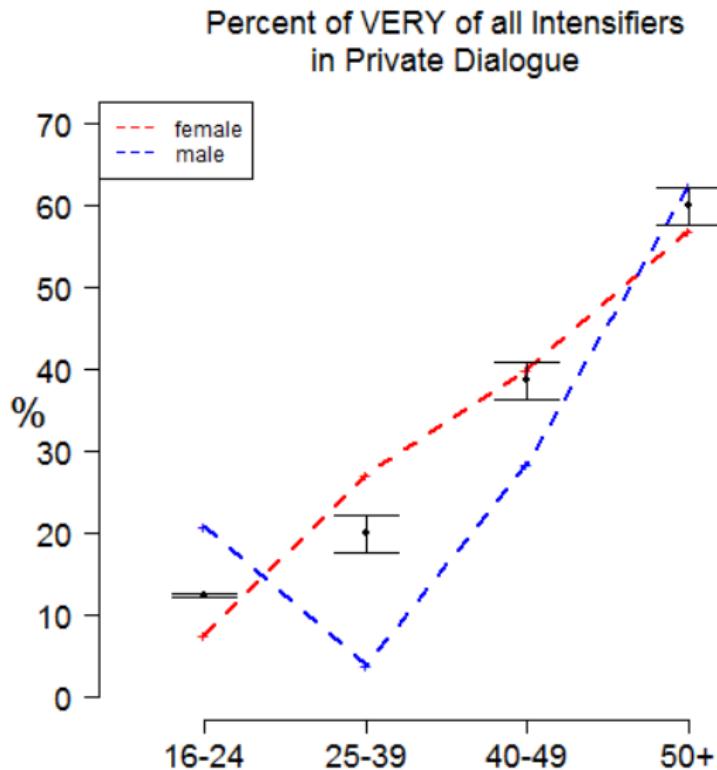
Percent of REALLY of all Intensifiers in Private Dialogue
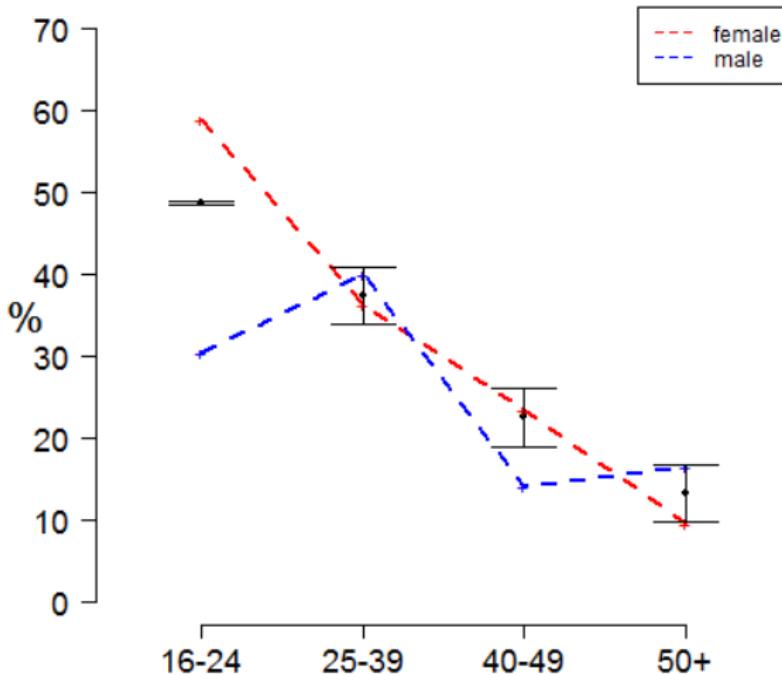
STATISTICAL ANALYSIS

# Research question

$$Q_2$$

Which factors correlate with the use of *really* (innovation)?
(age, sex, syntactic function, ...)

# Statistical Analysis

- Mixed–effects binomial logistic regression models
  - intensified adjectives only, step-wise step-up model fitting

**Dependent Variable(s)**

| **really** | nominal | occurrence of *really* vs other intensifier |
|---|---|---|

**Independent Variable(s)**

| **age** | categorical | age groups in ascending order | | |
|---|---|---|---|---|
| **sex** | nominal | male \| female | extra | linguistic |
| **eth** | nominal | pakeha \| maori | | |
| **occ** | nominal | acmp \| sml | | |
| **emo** | nominal | emotional \| nonemotional | | |
| **fun** | nominal | attributive \| predicative | intra | linguistic |
| **sem** | categorical | semantic type of adjective | | |
| **grad** | nominal | gradable \| nongradable | | |

Regression Results

# Regression Results

|  | Group(s) | Variance | SDev. | L.R. $\chi^2$ | Sig. |
|---|---|---|---|---|---|
| **Random Effect(s)** | flid | 0.83 | 0.91 | 24.01 (DF 1) | <.001*** |
| **Fixed Effect(s)** | **Estimate** | **VIF** | **OddsRatio** | **z value** | **Sig.** |
| (Intercept) | -0.69 |  | 0.5 | -1.62 | n.s. |
| age:25-39 | -0.59 | 1.12 | 0.55 | -1.49 | n.s. |
| age:40-49 | -1.46 | 1.14 | 0.23 | -3.04 | <.01 ** |
| age:50+ | -2.18 | 1.08 | 0.11 | -3.55 | <.001*** |
| grad:nograd | 1.06 | 1.01 | 2.88 | 2.7 | <.01 ** |
| sex:male | -0.84 | 1.04 | 0.43 | -2.4 | <.05 * |
| **Model statistics** |  |  |  |  | **Value** |
| Number of Groups |  |  |  |  | 115 |
| Number of cases in model |  |  |  |  | 388 |
| Observed misses |  |  |  |  | 240 |
| Observed successes |  |  |  |  | 148 |
| R2 (Nagelkerke) |  |  |  |  | 0.408 |
| C |  |  |  |  | 0.842 |
| Somers' Dxy |  |  |  |  | 0.684 |
| Prediction accuracy |  |  |  |  | 79.12% |
| **Model Likelihood Ratio Test** |  |  | L.R. $\chi^2$ | 59.43 (DF 6) | <.001*** |

# Regression Results

|  | Group(s) | Variance | SDev. | L.R. $\chi^2$ | Sig. |
|---|---|---|---|---|---|
| **Random Effect(s)** | flid | 0.83 | 0.91 | 24.01 (DF 1) | <.001*** |
| **Fixed Effect(s)** | **Estimate** | **VIF** | **OddsRatio** | **z value** | **Sig.** |
| (Intercept) | -0.69 |  | 0.5 | -1.62 | n.s. |
| age:25-39 | -0.59 | 1.12 | 0.55 | -1.49 | n.s. |
| age:40-49 | -1.46 | 1.14 | 0.23 | -3.04 | <.01 ** |
| age:50+ | -2.18 | 1.08 | 0.11 | -3.55 | <.001*** |
| grad:nograd | 1.06 | 1.01 | 2.88 | 2.7 | <.01 ** |
| sex:male | -0.84 | 1.04 | 0.43 | -2.4 | <.05 * |
| **Model statistics** |  |  |  |  | **Value** |
| Number of Groups |  |  |  |  | 115 |
| Number of cases in model |  |  |  |  | 388 |
| Observed misses |  |  |  |  | 240 |
| Observed successes |  |  |  |  | 148 |
| R2 (Nagelkerke) |  |  |  |  | 0.408 |
| C |  |  |  |  | 0.842 |
| Somers' Dxy |  |  |  |  | 0.684 |
| Prediction accuracy |  |  |  |  | 79.12% |
| **Model Likelihood Ratio Test** |  |  |  | L.R. $\chi^2$  59.43 (DF 6) | <.001*** |

# Regression Results

|  | Group(s) | Variance | SDev. | L.R. $\chi^2$ | Sig. |
|---|---|---|---|---|---|
| **Random Effect(s)** | flid | 0.83 | 0.91 | 24.01 (DF 1) | <.001*** |
| **Fixed Effect(s)** | **Estimate** | **VIF** | **OddsRatio** | **z value** | **Sig.** |
| (Intercept) | -0.69 |  | 0.5 | -1.62 | n.s. |
| age:25-39 | -0.59 | 1.12 | 0.55 | -1.49 | n.s. |
| age:40-49 | -1.46 | 1.14 | 0.23 | -3.04 | <.01 ** |
| age:50+ | -2.18 | 1.08 | 0.11 | -3.55 | <.001*** |
| grad:nograd | 1.06 | 1.01 | 2.88 | 2.7 | <.01 ** |
| sex:male | -0.84 | 1.04 | 0.43 | -2.4 | <.05 * |
| **Model statistics** |  |  |  |  | **Value** |
| Number of Groups |  |  |  |  | 115 |
| Number of cases in model |  |  |  |  | 388 |
| Observed misses |  |  |  |  | 240 |
| Observed successes |  |  |  |  | 148 |
| R2 (Nagelkerke) |  |  |  |  | 0.408 |
| C |  |  |  |  | 0.842 |
| Somers' Dxy |  |  |  |  | 0.684 |
| Prediction accuracy |  |  |  |  | 79.12% |
| **Model Likelihood Ratio Test** |  |  | L.R. $\chi^2$ | 59.43 (DF 6) | <.001*** |

# Regression Results

|                           | Group(s) | Variance | SDev.     | L.R. $\chi^2$  | Sig.      |
|---------------------------|----------|----------|-----------|----------------|-----------|
| **Random Effect(s)**      | flid     | 0.83     | 0.91      | 24.01 (DF 1)   | <.001***  |
| **Fixed Effect(s)**       | Estimate | VIF      | OddsRatio | z value        | Sig.      |
| (Intercept)               | -0.69    |          | 0.5       | -1.62          | n.s.      |
| age:25-39                 | -0.59    | 1.12     | 0.55      | -1.49          | n.s.      |
| age:40-49                 | -1.46    | 1.14     | 0.23      | -3.04          | <.01 **   |
| age:50+                   | -2.18    | 1.08     | 0.11      | -3.55          | <.001***  |
| grad:nograd               | 1.06     | 1.01     | 2.88      | 2.7            | <.01 **   |
| sex:male                  | -0.84    | 1.04     | 0.43      | -2.4           | <.05 *    |
| **Model statistics**      |          |          |           |                | Value     |
| Number of Groups          |          |          |           |                | 115       |
| Number of cases in model  |          |          |           |                | 388       |
| Observed misses           |          |          |           |                | 240       |
| Observed successes        |          |          |           |                | 148       |
| R2 (Nagelkerke)           |          |          |           |                | 0.408     |
| C                         |          |          |           |                | 0.842     |
| Somers' Dxy               |          |          |           |                | 0.684     |
| Prediction accuracy       |          |          |           |                | 79.12%    |
| **Model Likelihood Ratio Test** |    |          | L.R. $\chi^2$ | 59.43 (DF 6) | <.001***  |

# Regression Results

|  | Group(s) | Variance | SDev. | L.R. $\chi^2$ | Sig. |
|---|---|---|---|---|---|
| **Random Effect(s)** | flid | 0.83 | 0.91 | 24.01 (DF 1) | <.001*** |
| **Fixed Effect(s)** | **Estimate** | **VIF** | **OddsRatio** | **z value** | **Sig.** |
| (Intercept) | -0.69 |  | 0.5 | -1.62 | n.s. |
| age:25-39 | -0.59 | 1.12 | 0.55 | -1.49 | n.s. |
| age:40-49 | -1.46 | 1.14 | 0.23 | -3.04 | <.01 ** |
| age:50+ | -2.18 | 1.08 | 0.11 | -3.55 | <.001*** |
| grad:nograd | 1.06 | 1.01 | 2.88 | 2.7 | <.01 ** |
| sex:male | -0.84 | 1.04 | 0.43 | -2.4 | <.05 * |
| **Model statistics** |  |  |  |  | **Value** |
| Number of Groups |  |  |  |  | 115 |
| Number of cases in model |  |  |  |  | 388 |
| Observed misses |  |  |  |  | 240 |
| Observed successes |  |  |  |  | 148 |
| R2 (Nagelkerke) |  |  |  |  | 0.408 |
| C |  |  |  |  | 0.842 |
| Somers' Dxy |  |  |  |  | 0.684 |
| Prediction accuracy |  |  |  |  | 79.12% |
| **Model Likelihood Ratio Test** |  |  | L.R. $\chi^2$ | 59.43 (DF 6) | <.001*** |

Summary

- ► The NZE intensifier system is undergoing change
- ► The observed change is accompanied by extra-linguistic stratification and pragmatic constraints

But why is *really* taking over???

# Collocates

$Q_3$

Why and how is *really* replacing *very*?
Why not another variant?

$H_1$

Successful variants (*really*) associate with many and
particularly high frequency adjectives.

| | | Age | | | |
|---|---|---|---|---|---|
| **Adjective** | **Intensifier** | 16-24 | 25-39 | 40-49 | 50+ |
| difficult | really | 0 | 2 | 0 | 0 |
| good | really | 27 | 9 | 2 | 3 |
| hard | really | 5 | 3 | 0 | 1 |
| important | really | 0 | 1 | 1 | 2 |
| interesting | really | 1 | 2 | 1 | 2 |
| little | really | 0 | 0 | 0 | 0 |
| nice | really | 7 | 1 | 1 | 1 |
| strong | really | 0 | 0 | 0 | 0 |
| difficult | very | 0 | 4 | 5 | 12 |
| good | very | 5 | 15 | 20 | 23 |
| hard | very | 0 | 3 | 4 | 4 |
| important | very | 2 | 6 | 4 | 8 |
| interesting | very | 1 | 0 | 2 | 8 |
| little | very | 0 | 5 | 4 | 11 |
| nice | very | 3 | 2 | 5 | 3 |
| strong | very | 0 | 2 | 4 | 8 |
| difficult | other | 2 | 3 | 3 | 2 |
| good | other | 8 | 6 | 6 | 6 |
| hard | other | 5 | 2 | 2 | 0 |
| important | other | 1 | 0 | 2 | 2 |
| interesting | other | 1 | 1 | 0 | 1 |
| little | other | 0 | 0 | 1 | 1 |
| nice | other | 0 | 0 | 0 | 0 |
| strong | other | 0 | 3 | 0 | 0 |

increasing trend
receding trend
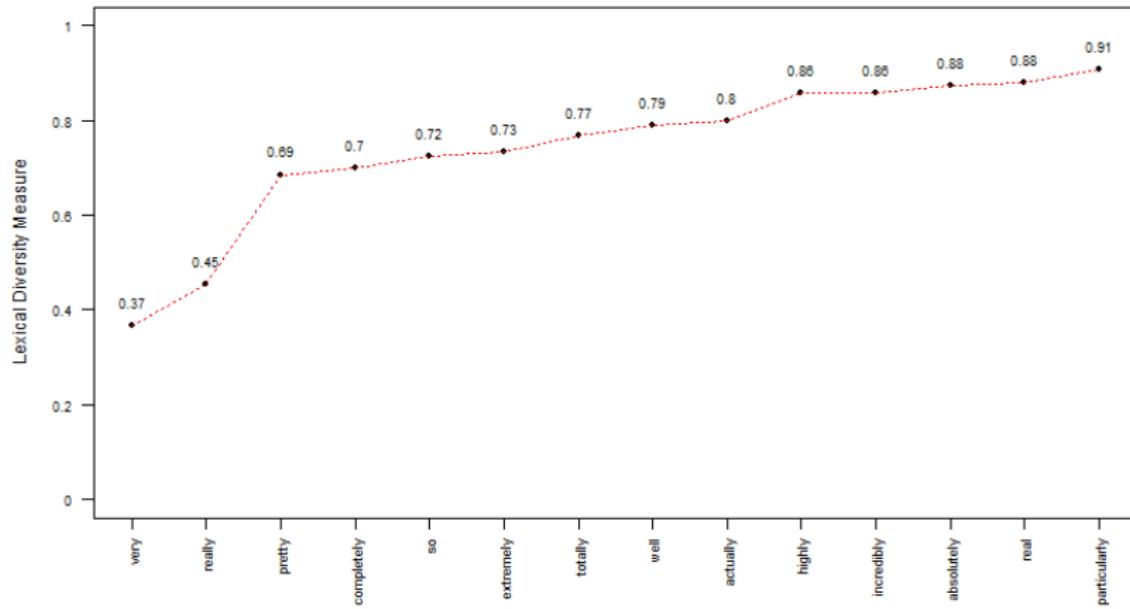
Figure: adjective types/intensifier tokens (ICE NZ)
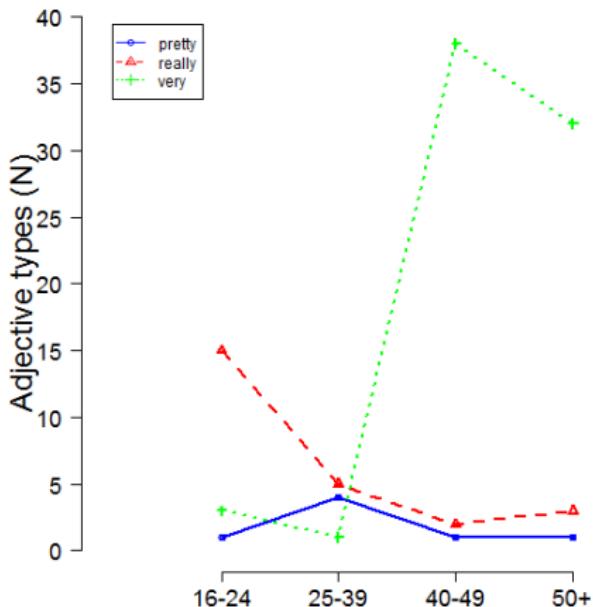
Figure: intensifier types : adjective types : age (ICE NZ)

# Covarying Collexeme Analysis (CCA)

- ▶ Member of the *collostructional analysis* family that aims at measuring the degree of attraction of lemmas in one slot of a construction to lemmas in another slot of the same construction. (Stefanowitsch and Gries 2005)

- ▶ CCA differs from most collocation statistics such that it takes syntactic structure into account and it applies Fisher-Yates exact test and is thus not based on any distributional assumptions.

Collocations by Age

| Age | Adjective | Intensifier | OddsRatio | Bonf. Corr. Sig |
|-----|-----------|-------------|-----------|------------------|
| 16-24 | really | good | 5.44 | p<.001 |
| 50+ | very | difficult | 20.07 | p<.001 |
| 50+ | very | good | 4.72 | p<.001 |
| 50+ | very | strong | 21.33 | p<.01 |

# Correspondence Analysis (CA)

▶ Exploratory method for correspondence between rows and column values/counts in distributional data (similar to factor analysis) (Baayen 2008: 128–136)

▶ Restrictions on intensifiers
Minimum of 7 intensifier tokens and intensifier type has to collocate with more than 1 adjective type

▶ Restrictions on adjectives
Minimum of 7 adjective tokens

|           | pretty | real | really | so | very |
|-----------|--------|------|--------|-----|------|
| bad       | 1      | 2    | 3      | 1   | 1    |
| clear     | 0      | 1    | 0      | 0   | 7    |
| close     | 0      | 0    | 2      | 2   | 3    |
| different | 0      | 0    | 1      | 2   | 4    |
| difficult | 2      | 0    | 2      | 2   | 21   |
| first     | 0      | 0    | 0      | 0   | 9    |

Martin Schweinberger , www.martinschweinberger.de                    37
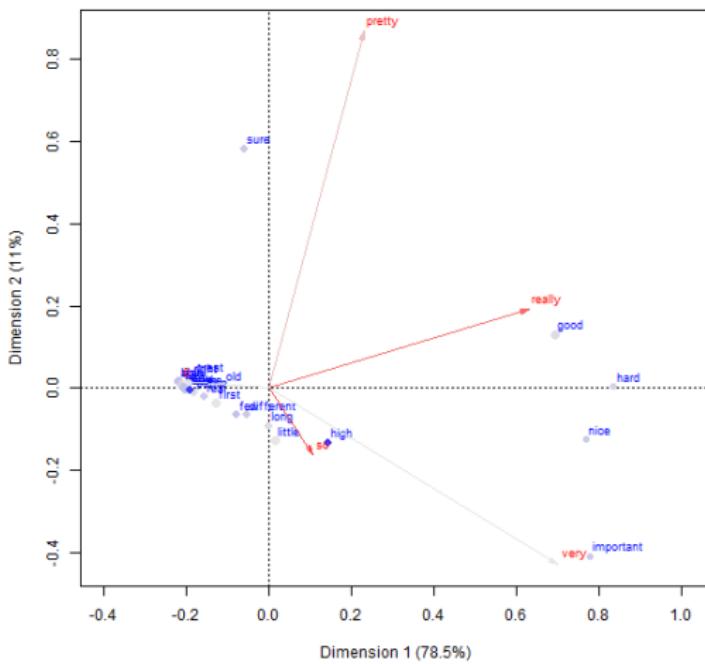
Figure: Correspondence of intensifier and adjective tokens (NZE)
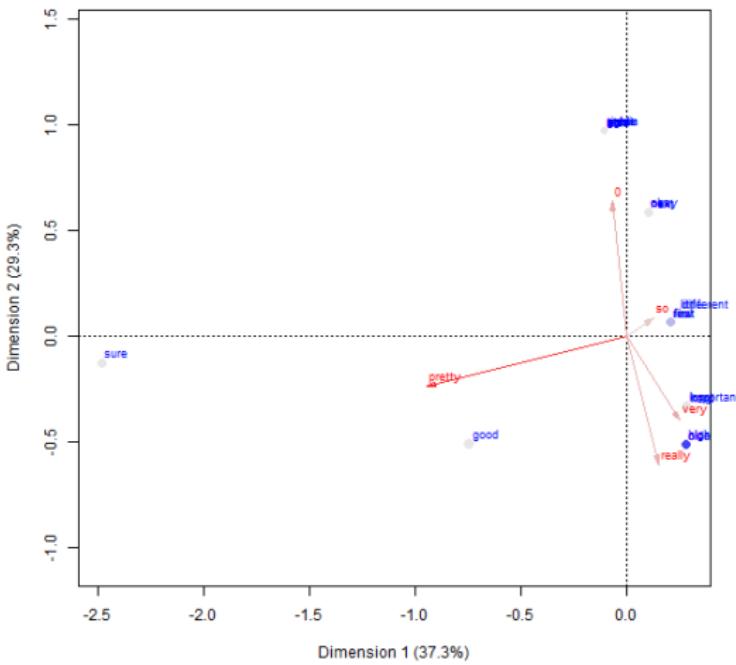
Figure: Correspondence of intensifier and adjective types (NZE)

SUMMARY

Intensifying *really*

- ▶ use declines almost linearly with age (incoming innovation)
- ▶ is dis-preferred by male speakers (female dominated change)
- ▶ is preferred by non-gradual adjectives

> *Really* is socially stratified and correlates with extra-linguistic factors (age, sex) and is subject to pragmatic constraints (gradability).

Successful intensifiers (*really*)

- ▸ increase the number of adjectives that they co-occur with
- ▸ collocate significantly with high frequency adjectives (*good* among 16-24 year olds)
- ▸ semantically similar to dominant variant (have similar collocation profiles)

Successful intensifier variants collocate with many adjectives (semantically bleached), particularly high frequency adjectives, and show collocation profiles similar to those of (previously) dominant variants.

PROBLEMS & OUTLOOK

Problems

- ▶ apply multinominal regression instead of binomial (dependent variable is categorical, really, very, other, null)
- ▶ use Wellington corpus instead of ICE (1,000,000 instead of 200,000 words) → increase of data base
- ▶ apply SVSM to adjectives to find distinct types of adjectives
- ▶ define variable context and variants better (not all adjectives take all intensifiers)!
- ▶ clean data: find true variants (*?completely* good : *very* good)

THANK YOU SO, REALLY, VERY MUCH!

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University press.

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

Bauer, L., A. Bell, D. Britain, G. Kennedy, C. Lane, M. Meyerhoff, and M. Stubbe (1999). The new zealand component of the international corpus of english (ice-nz).

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.

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.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

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

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.

Stefanowitsch, A. and S. T. Gries (2005). Covarying collexemes. *Corpus linguistics and linguistic theory 1*(1), 1–43.

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.

# MODELING INTENSIFICATION IN NEW ZEALAND ENGLISH DATA

Martin Schweinberger
www.martinschweinberger.de

Appendix

# Regression Results

|  | Group(s) | Variance | Std. Dev. | L.R.$\chi^2$(df1) | Sig. |
|---|---|---|---|---|---|
| **Random Effect(s)** | flid | 0.44 | 0.66 | 29 | p<.001*** |
| **Fixed Effect(s)** | **Estimate** | **VIF** | **OddsRatio** | **z value** | **Sig.** |
| (Intercept) | -5.04 |  | 0.01 | -14.55 | p<.001*** |
| age:25-39 | -0.57 | 1.07 | 0.57 | -2.09 | p<.05* |
| age:40-49 | -0.94 | 1.08 | 0.39 | -2.7 | p<.01** |
| age:50+ | -1.48 | 1.03 | 0.23 | -2.98 | p<.01** |
| sex:male | -0.85 | 1.01 | 0.43 | -3.46 | p<.001*** |
| fun:predicative | 0.74 | 1 | 2.09 | 4.09 | p<.001*** |
| grad:nograd | 1.88 | 1.01 | 6.52 | 6.31 | p<.001*** |
| emo:emotional | 0.79 | 1.01 | 2.21 | 4.49 | p<.001*** |
| **Model statistics** |  |  |  |  | **Value** |
| Number of Groups |  |  |  |  | 145 |
| Cases in model |  |  |  |  | 4199 |
| Observed successes |  |  |  |  | 150 |
| $R^2$ (Nagelkerke) |  |  |  |  | 0.155 |
| C |  |  |  |  | 0.844 |
| Somers' D$_{xy}$ |  |  |  |  | 0.688 |
| Prediction accuracy |  |  |  |  | 96.43% |
| **Model LL Ratio Test** |  |  | L.R.$\chi^2$(df8) | 176.67 | p<.001*** |

# Distributional Analysis)

- ▶ Network Analysis
- ▶ Words that share collocates are semantically similar
  (Stefanowitsch 2010: 368–370).
- ▶ Semantic similarity can thus be measured in collocate
  frequency.

|         | extremely | particularly | pretty | real | really | so | very |
|---------|-----------|--------------|--------|------|--------|-----|------|
| amazing | 0         | 0            | 0      | 0    | 4      | 1   | 0    |
| bad     | 0         | 0            | 1      | 2    | 3      | 1   | 1    |
| basic   | 0         | 0            | 1      | 1    | 0      | 0   | 3    |
| bright  | 0         | 0            | 0      | 0    | 0      | 0   | 4    |
| busy    | 0         | 0            | 1      | 0    | 2      | 2   | 1    |
| careful | 0         | 0            | 0      | 0    | 1      | 0   | 6    |
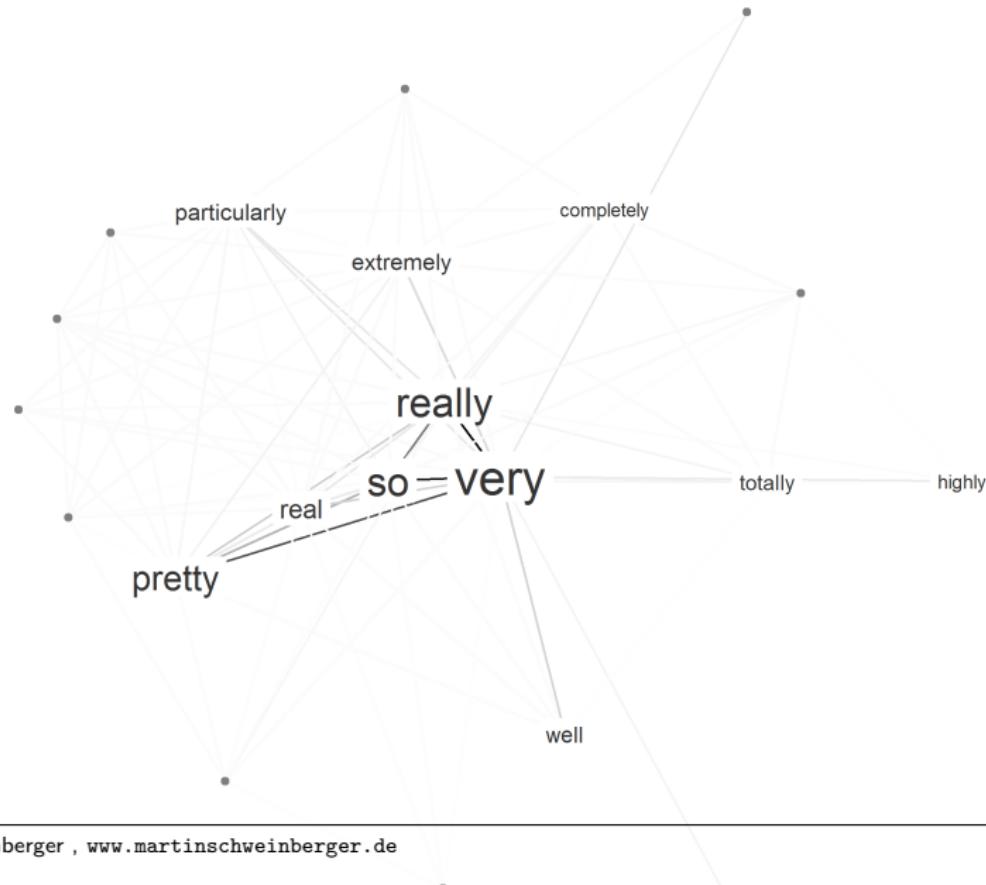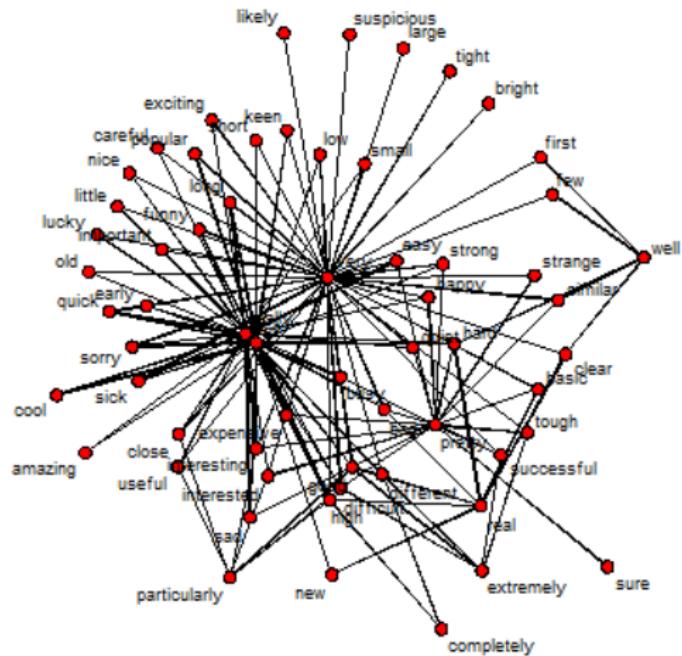
# Semantic Vector Space Models (SVSM)

- ▶ Distributional approach to semantics (cf. Levshina 2015: 323–331).
- ▶ Semantic similarity is measured as similarity of vectors of collocate frequency.
- ▶ The similarity measure is a cosine measure (the more similar the cosine value the greater the similarity in the vectors and thus the distributional collocate profile of two intensifiers.

**Cluster Dendrogram**
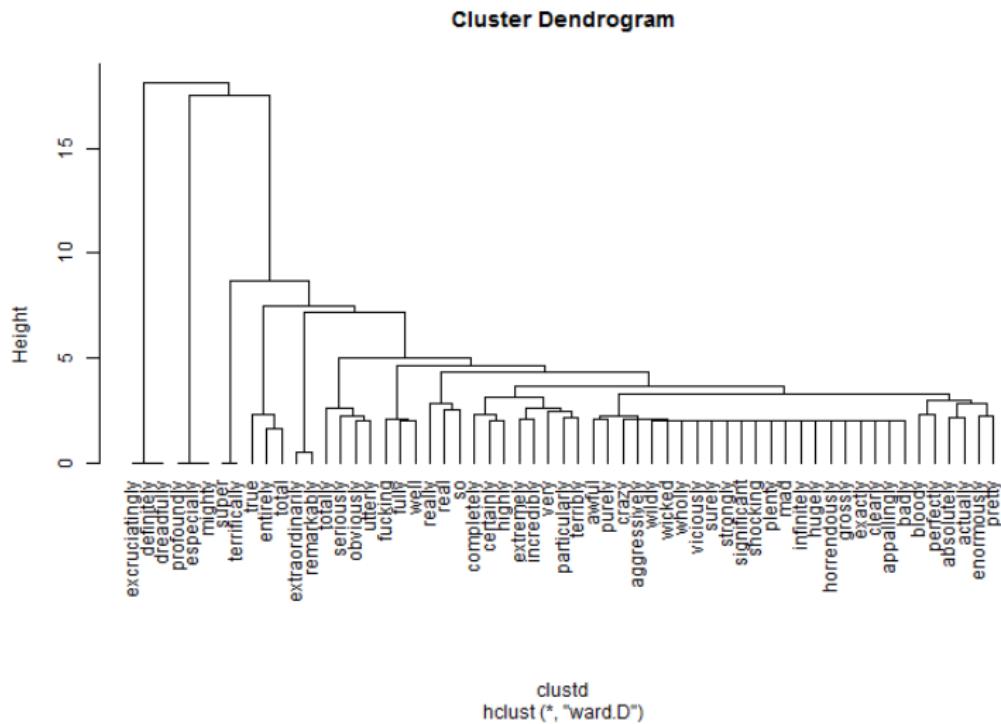


Figure: Clustering of intensifiers by adjective collocates (NZE)