Computational Approaches to Textual Data

Dr. Martin Schweinberger

slides available at

www.martinschweinberger.de

The University
Of Queensland
AUSTRALIA

Aims of this talk

- Give an overview of computational approaches to analyzing textual data
- Exemplify computational approaches to analyzing textual data
- Provide information about The Language Technology and Data Analysis Laboratory (LADAL)

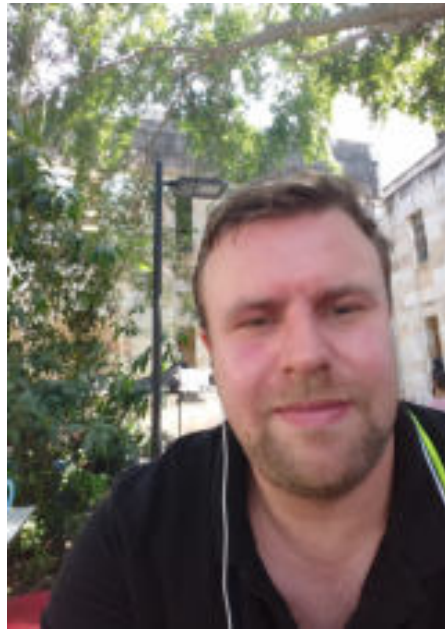
Introduction

Corpus Linguistics

Amplification in SLA

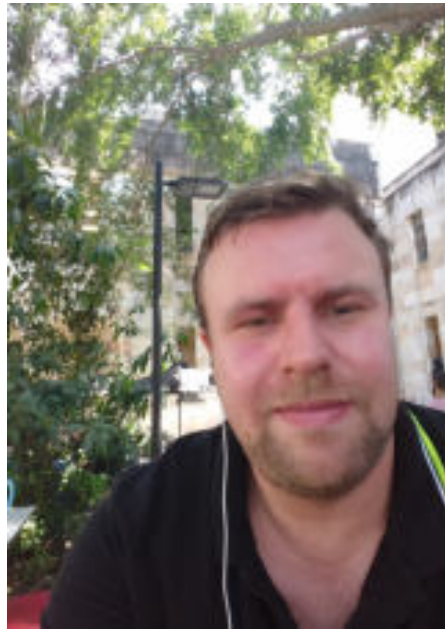
LADAL

About me



- Postdoctoral Research Fellow in Language Technologies at UQ's *School of Languages and Cultures*
- PhD in English linguistics (U Hamburg, Germany)
- Studied Philosophy, English Philology, and Psychology at U Kassel (Germany) and the National University of Ireland, Galway
- Language Technology Group at the Computer Science department of Universität Hamburg

About me



- Focus on computational approaches to language data with a specialization in statistical modeling.
- Building *The Language Technology and Data Analysis Laboratory* (LADAL).
- Consultant for issues relating to statistics, text analysis, and research design (methodology)
- Concerned with Best Practices and Quality Control in Data Analysis and Research Data Management (as a result of the Replication Crisis)

What I study

Phenomenon: Adjective Amplification

(1) And you just have to hint well then it's a very good hint (ICE-AUS:S1A-012$A)

(2) They're all really cheap <#> They're all really nice, the t-shirts in there (ICE-AUS:S1A-009$B)

(3) It was so bad (ICE-AUS:S1A-044$B)

What I study

Variationist Sociolinguistics

- Language is not homogeneous: variation is ubiquitous
  - Social factors : language use
  - Linguistic variation not random
  - Systematic correlation between certain social factors (age, gender, class, ethnicity, etc.) and language use
- Linguistic differentiation $\leftrightarrow$ social stratification

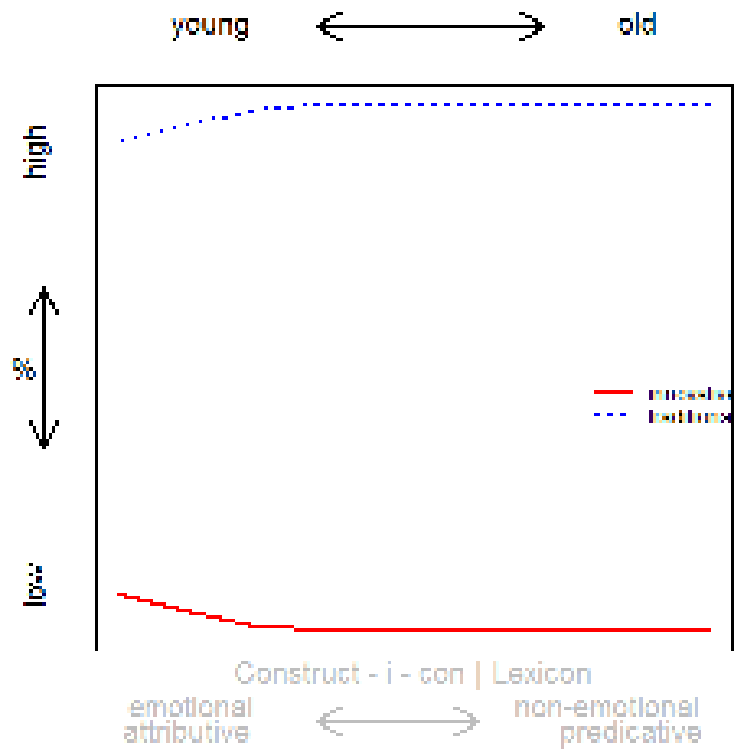
Diffusion of Innovations

young ⟷ old

upper

⟷

working



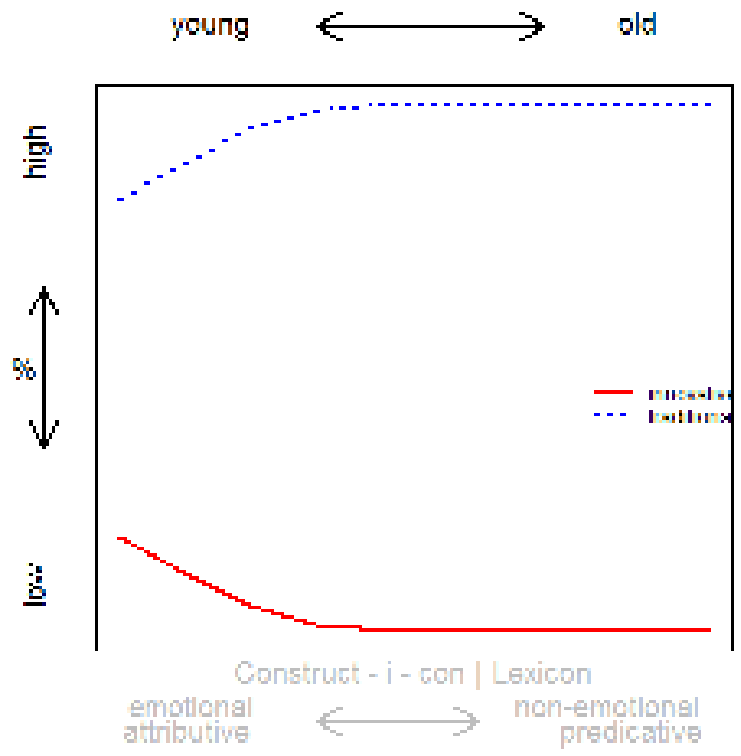
Diffusion of Innovations

young ⟷ old

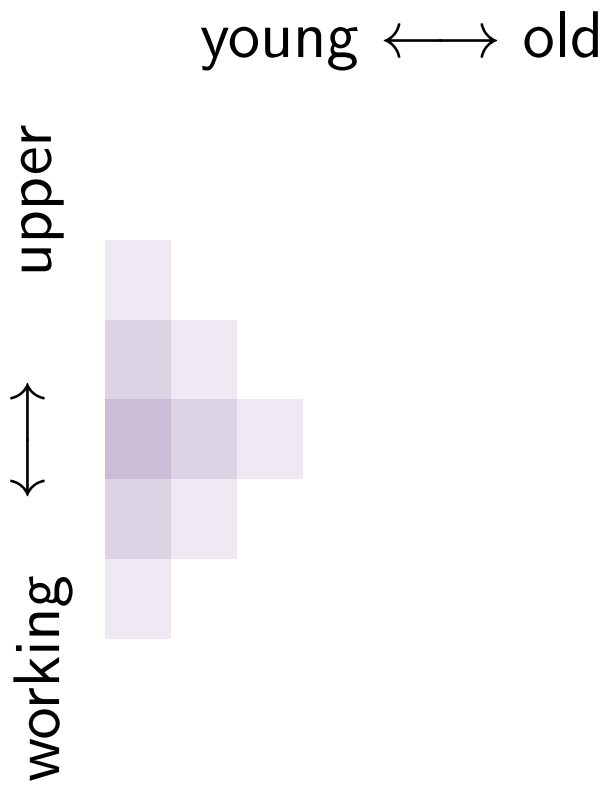
upper

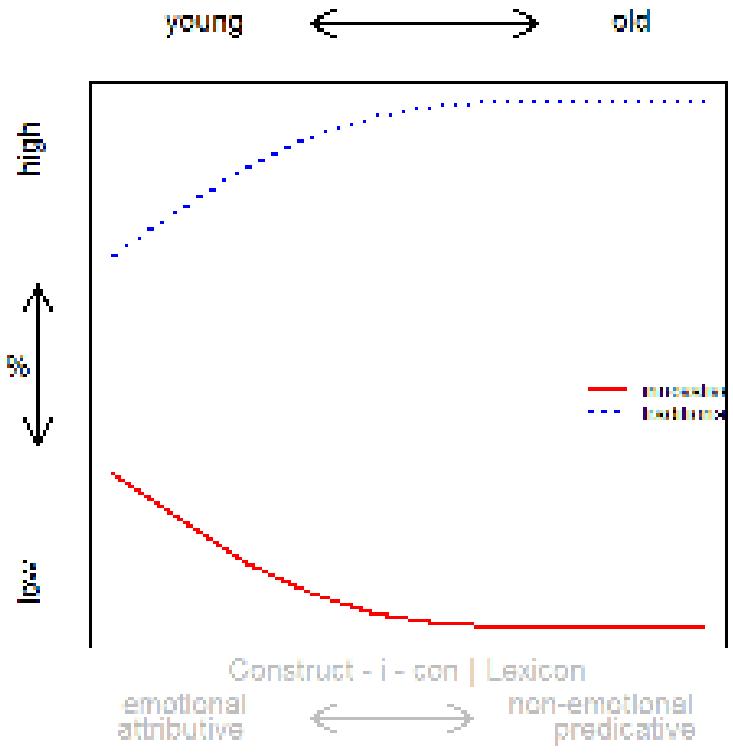
⟷

working

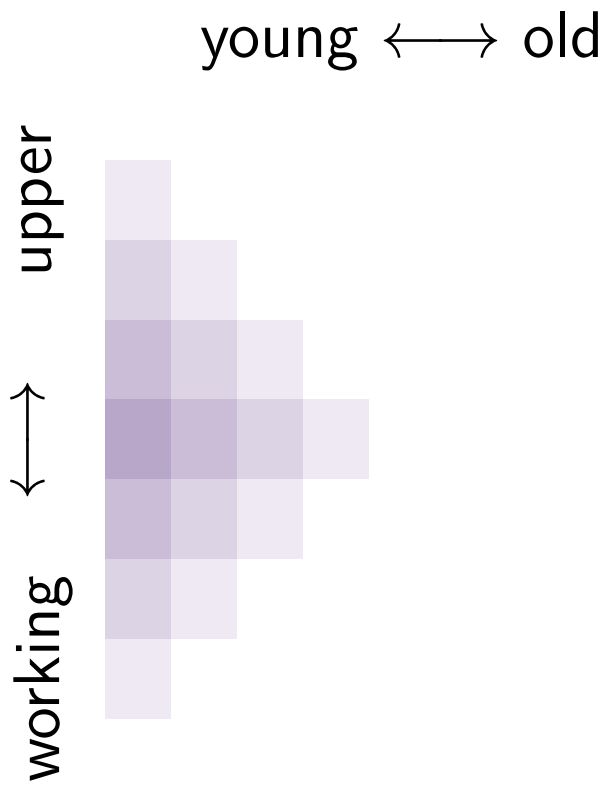


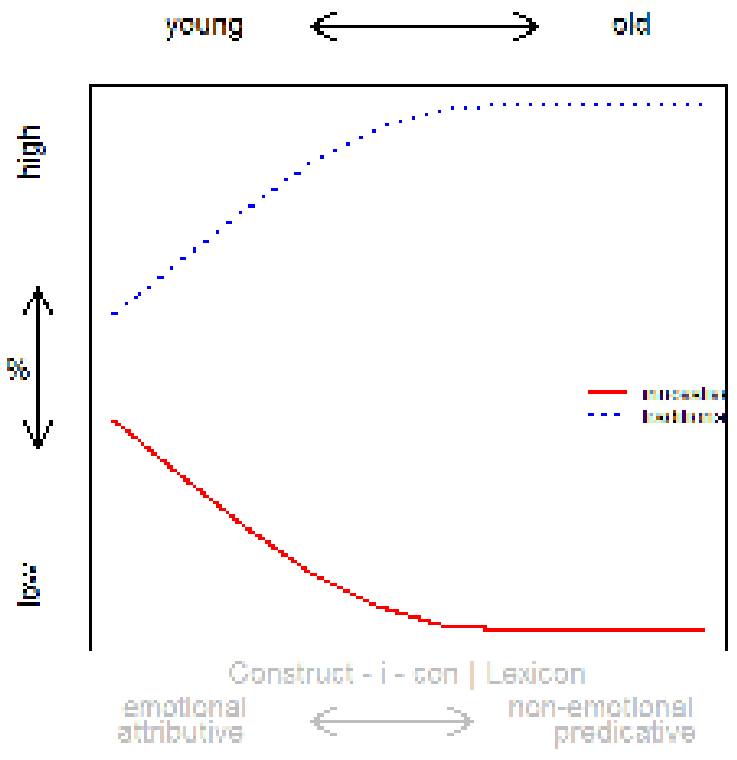
Diffusion of Innovations



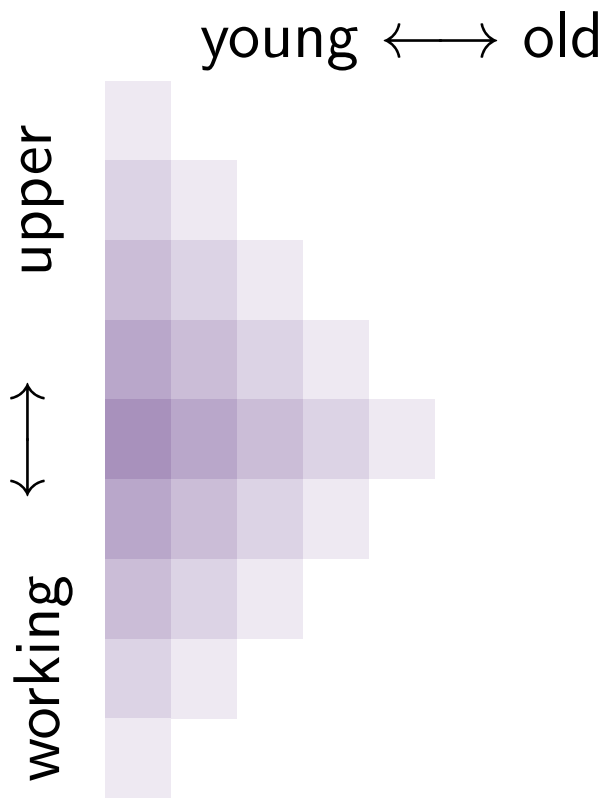


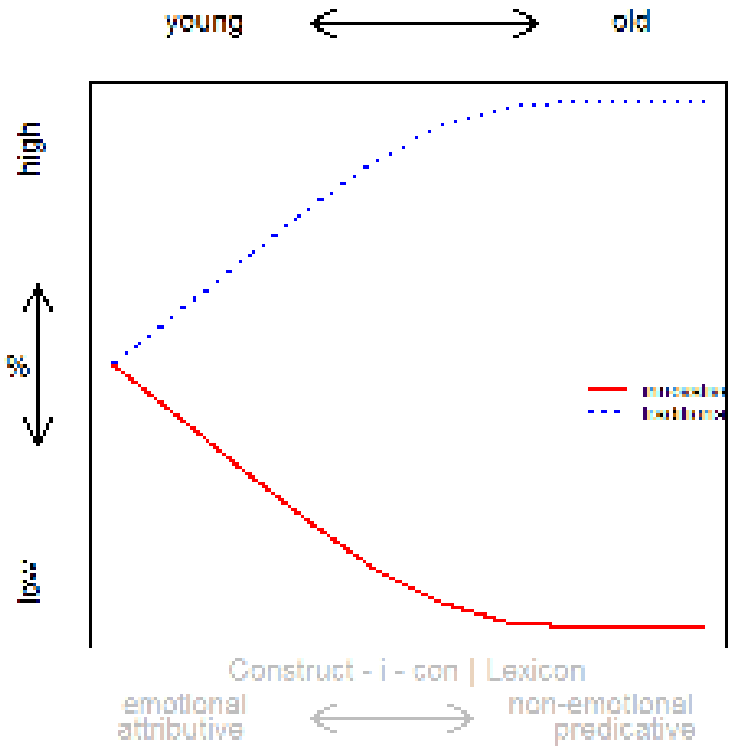
Diffusion of Innovations



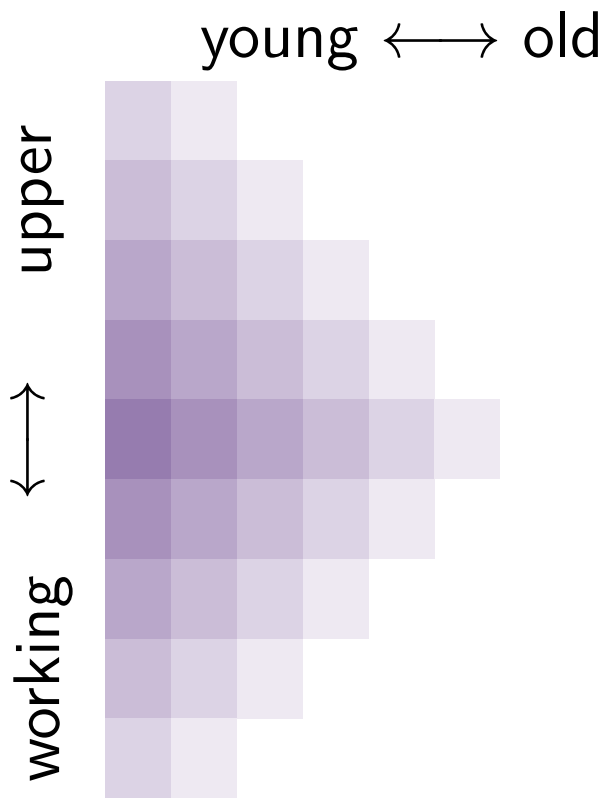


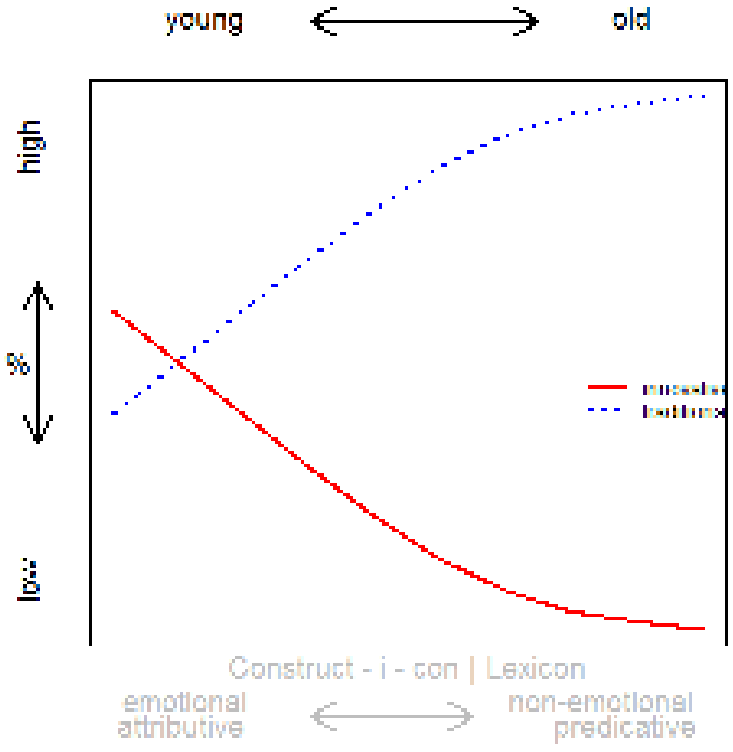
Diffusion of Innovations



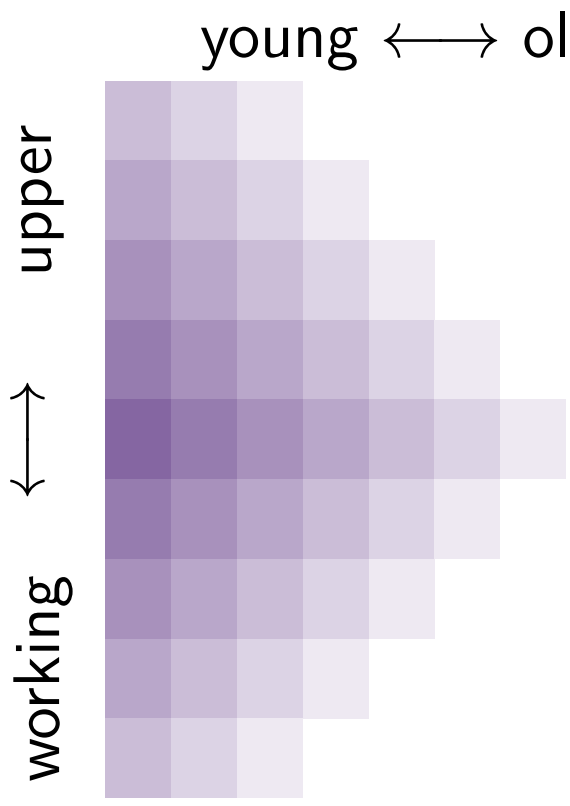


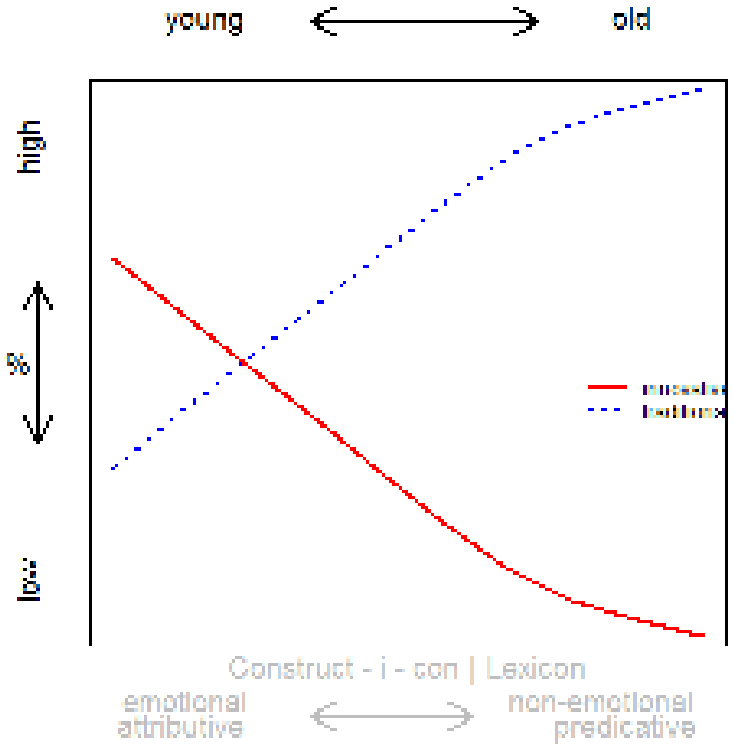
Diffusion of Innovations



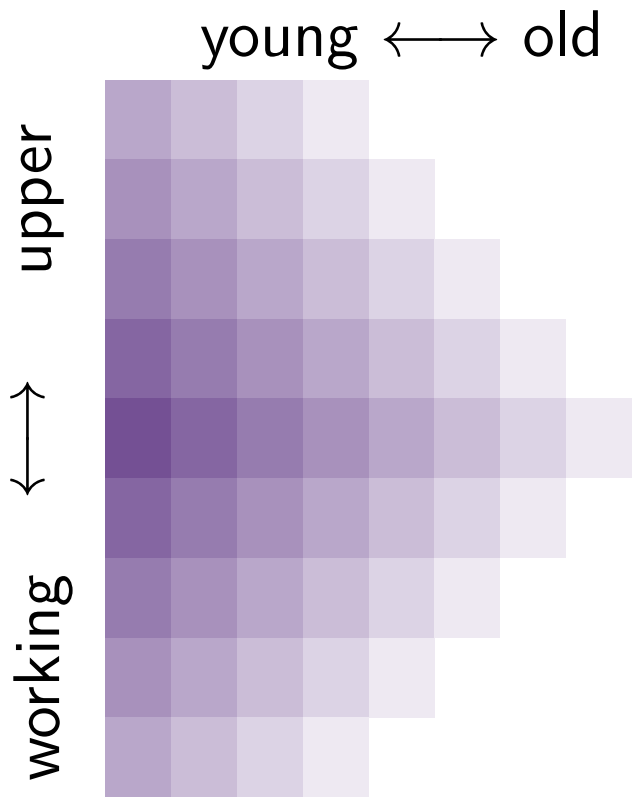


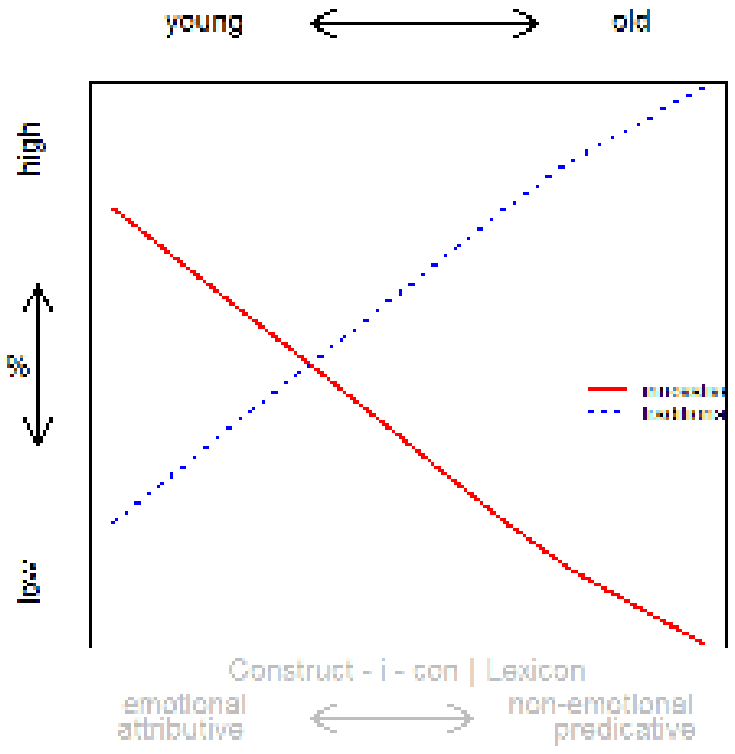
Diffusion of Innovations



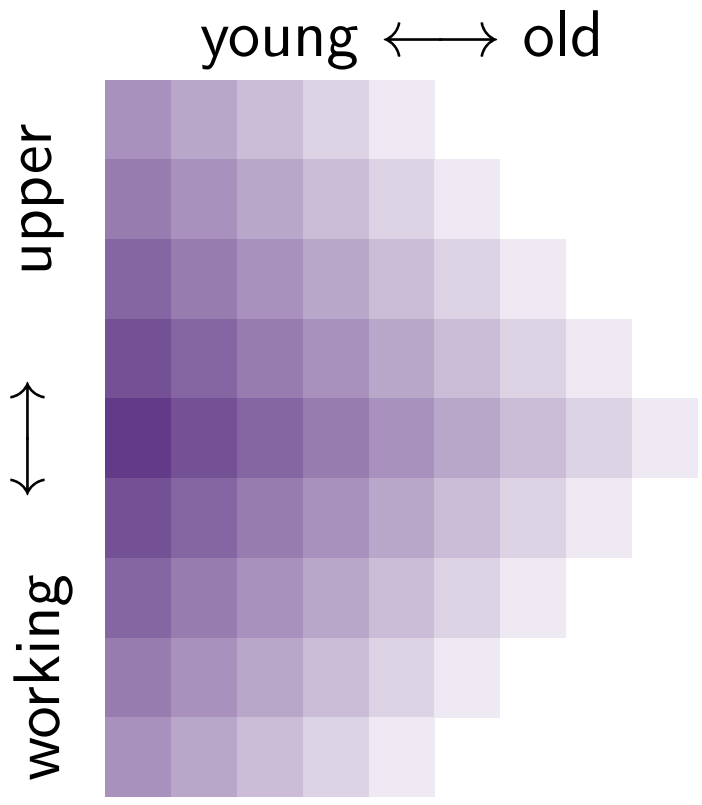


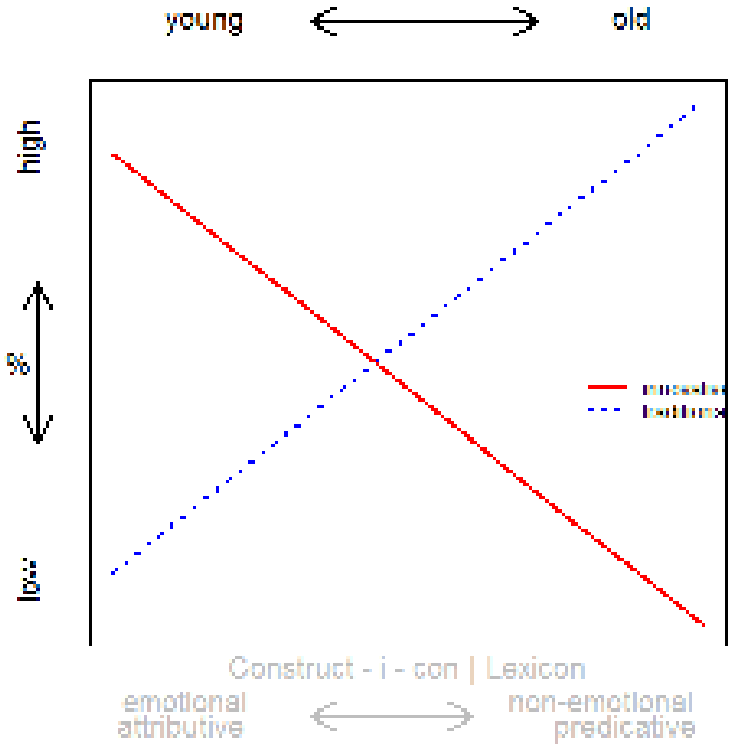
Diffusion of Innovations



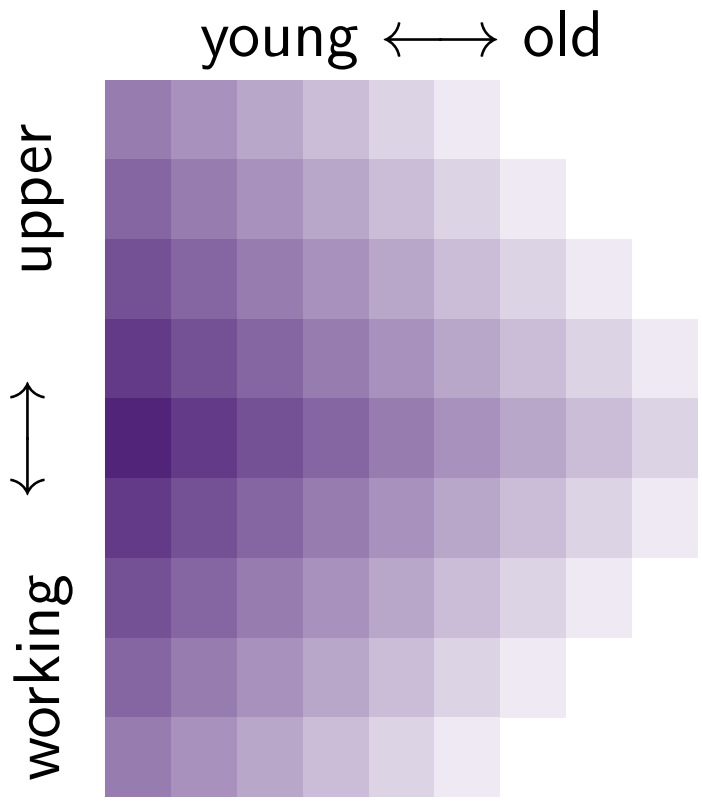


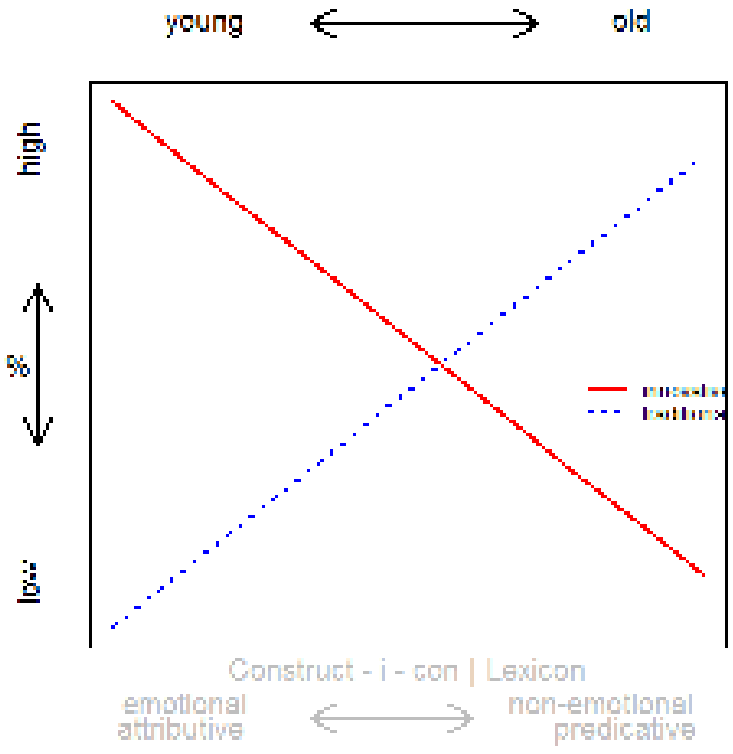
Diffusion of Innovations



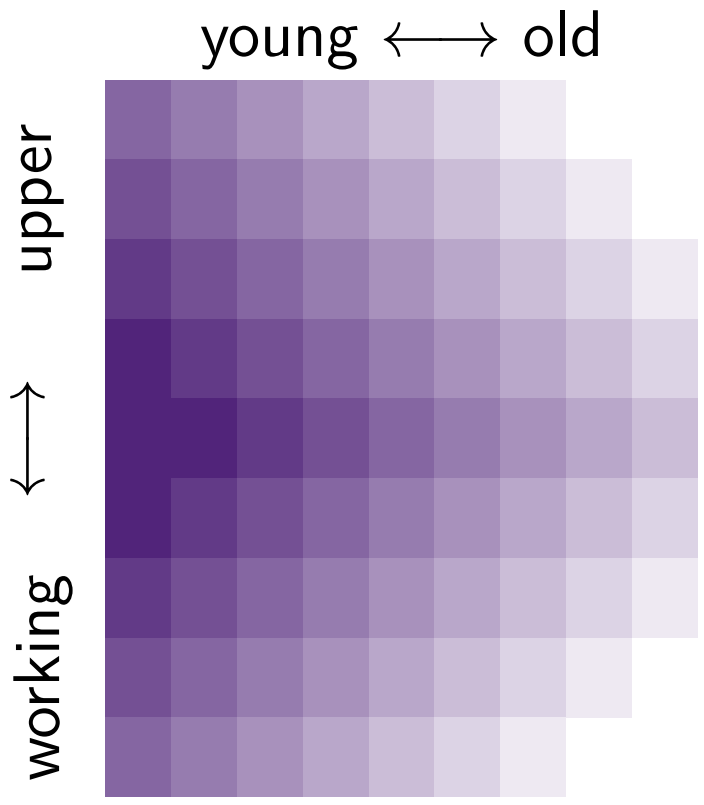


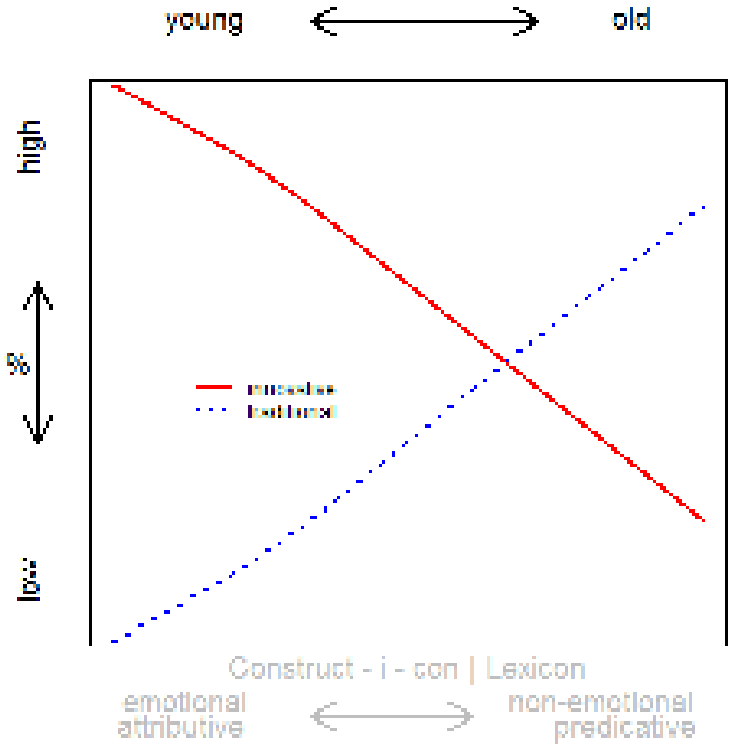
Diffusion of Innovations



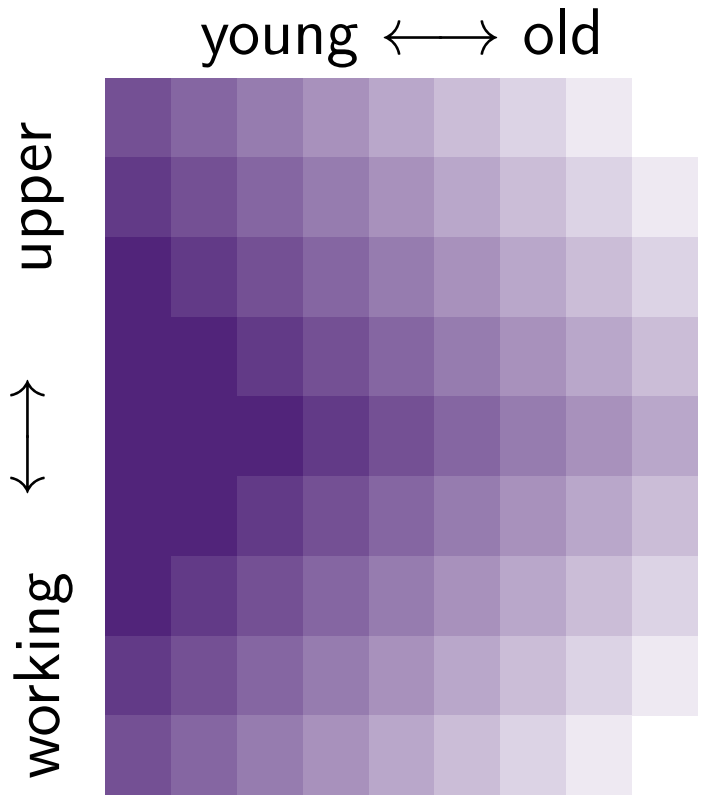


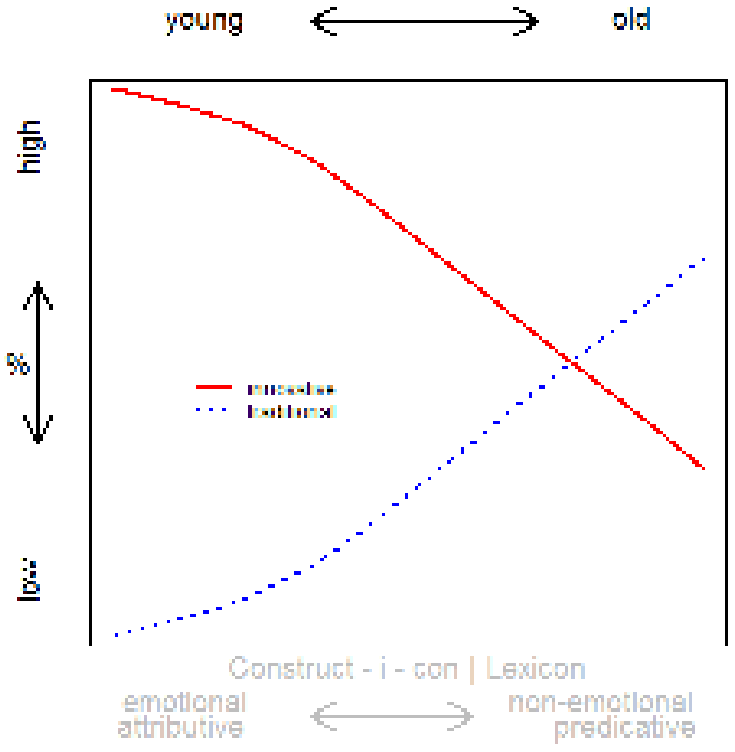
Diffusion of Innovations



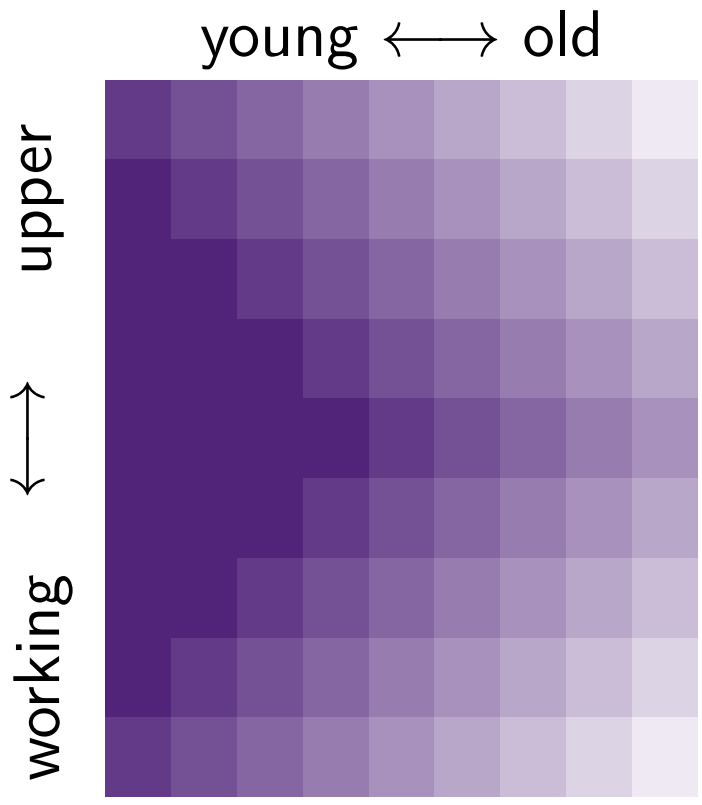


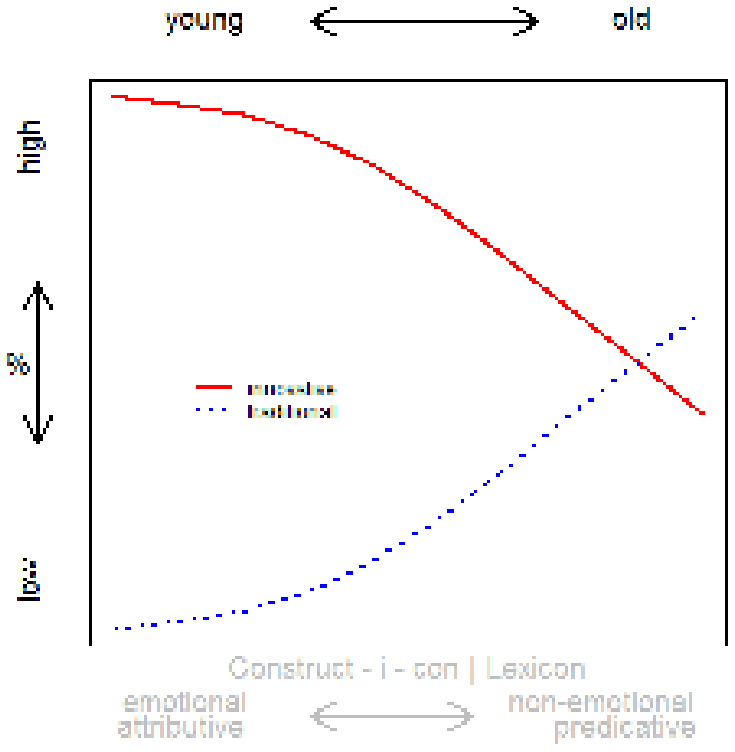
Diffusion of Innovations



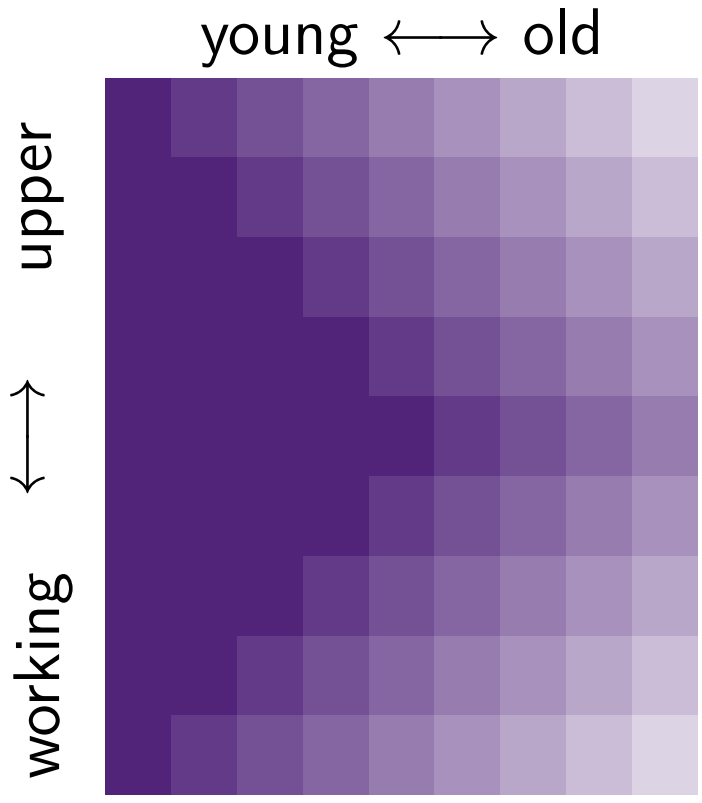


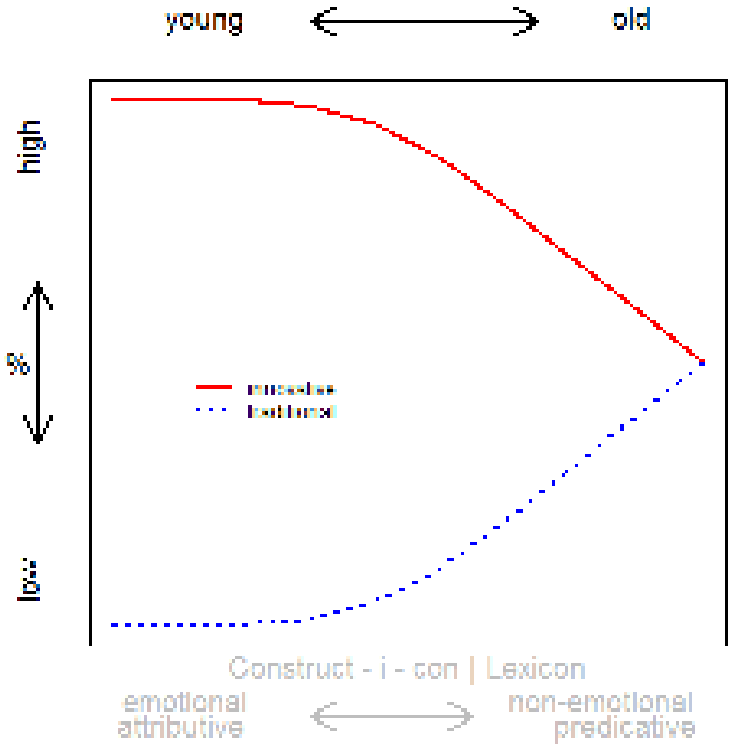
Diffusion of Innovations



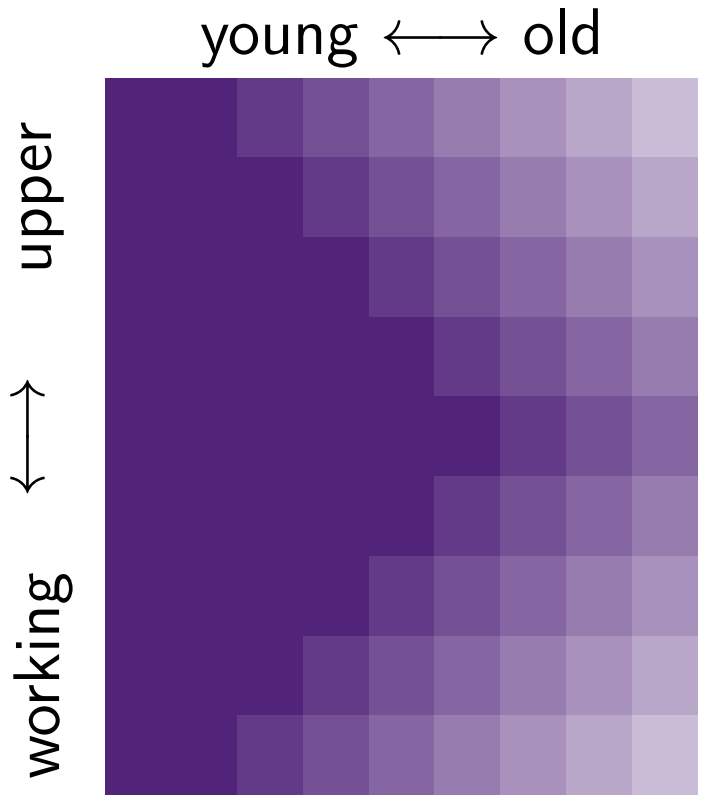


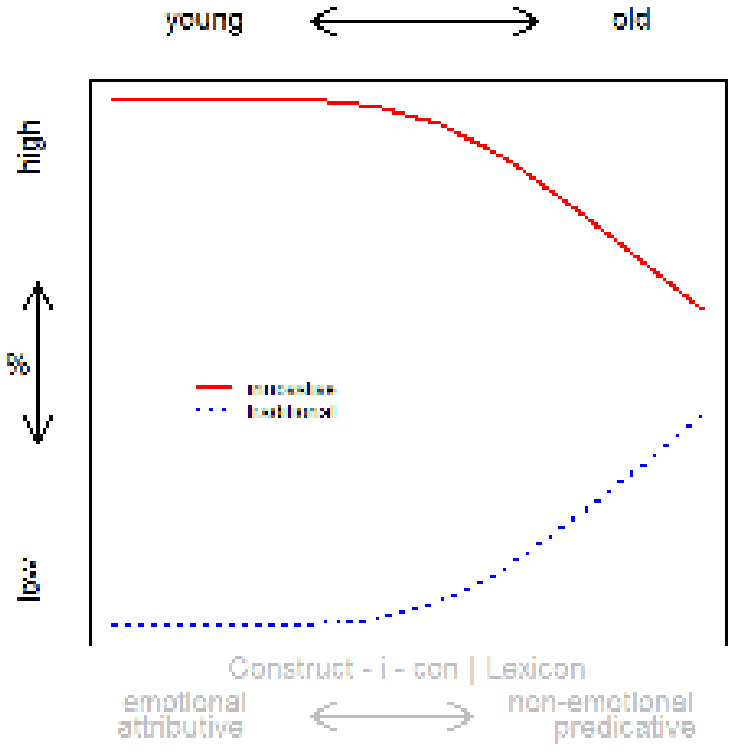
Diffusion of Innovations



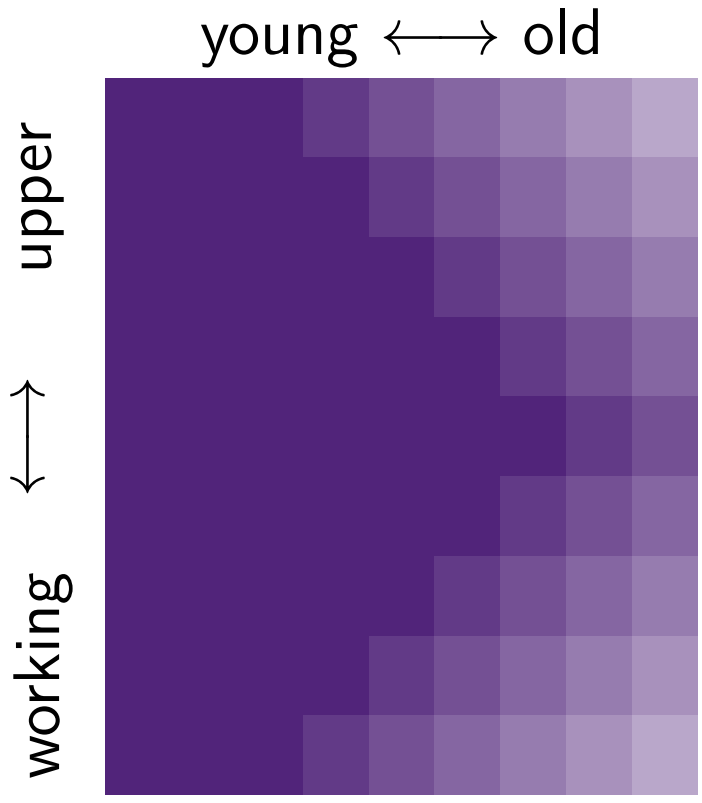


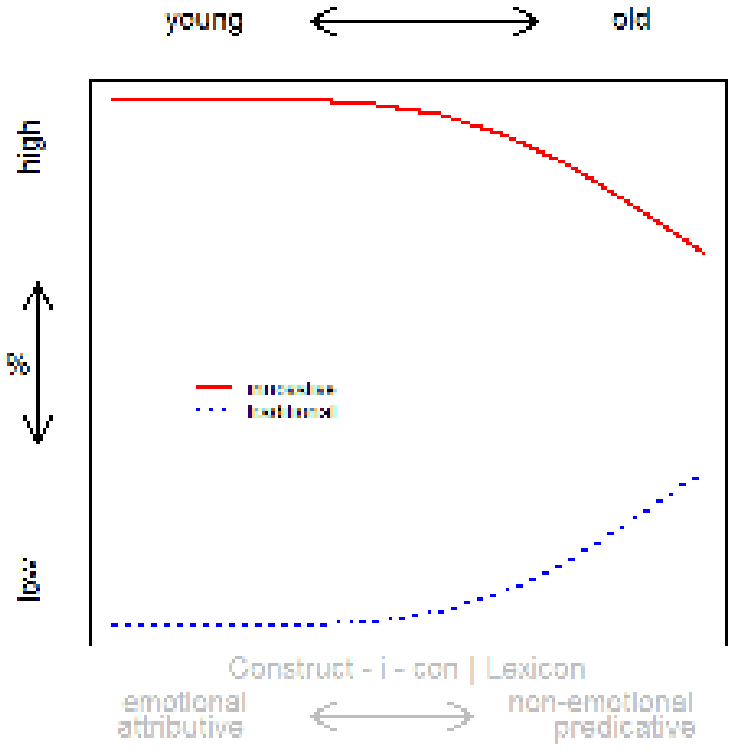
Diffusion of Innovations



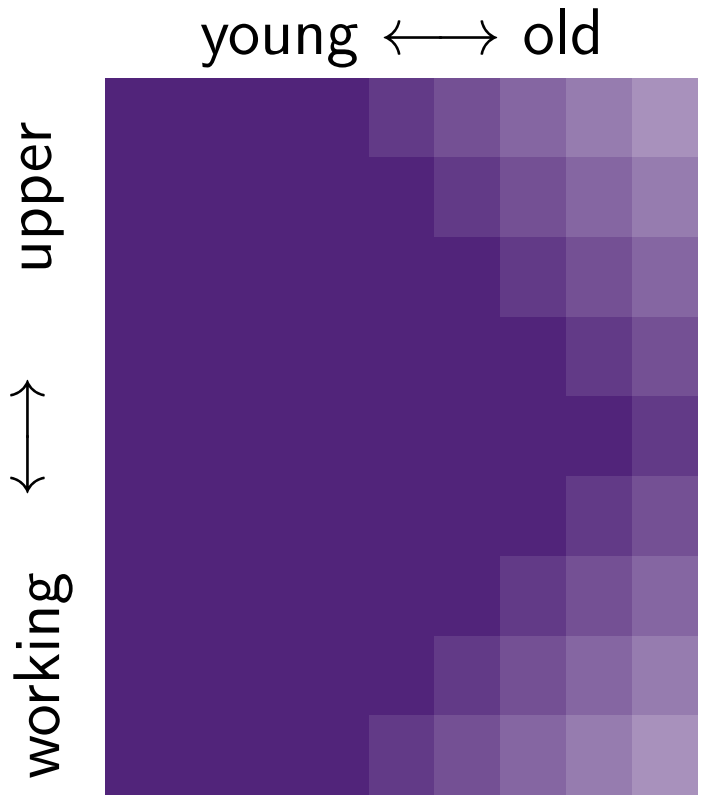


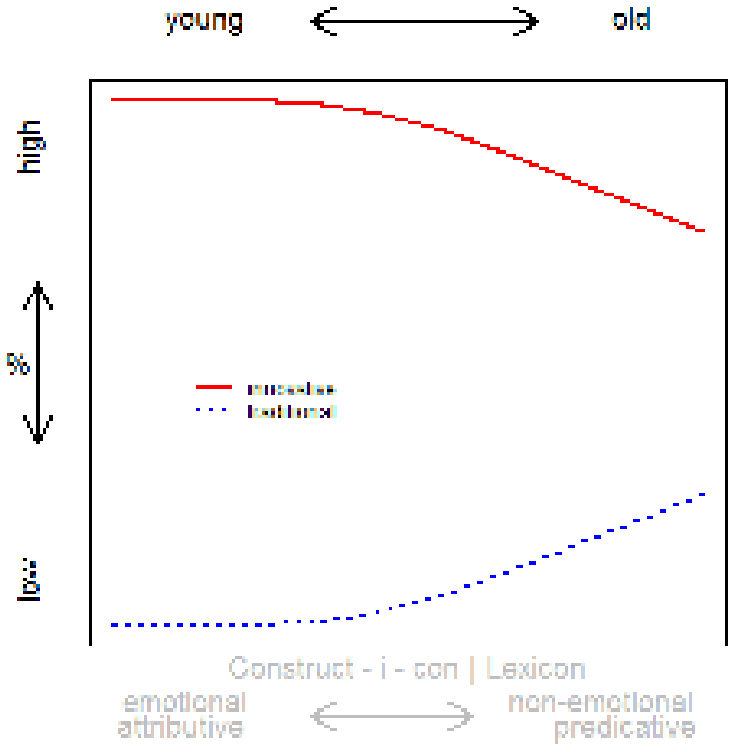
Diffusion of Innovations



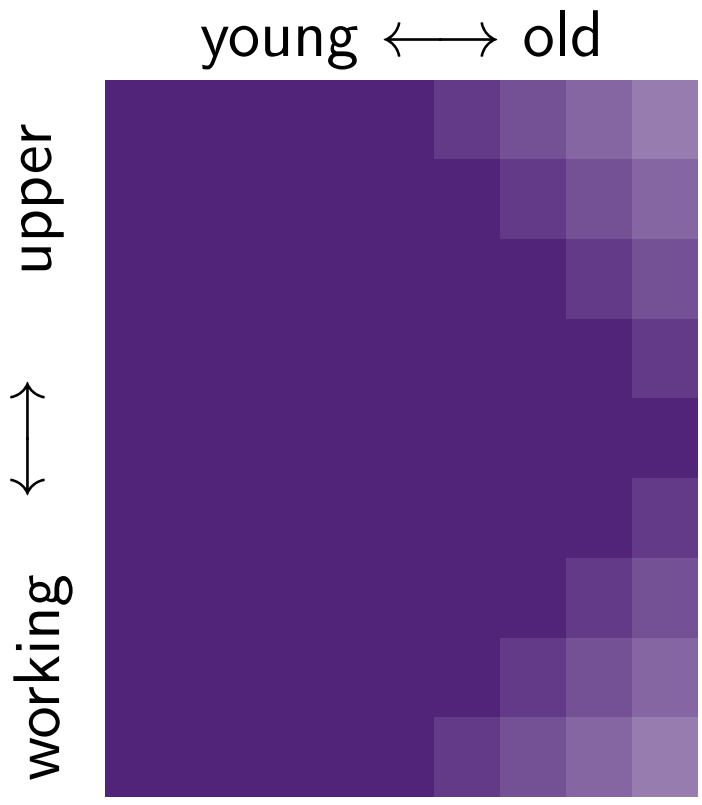


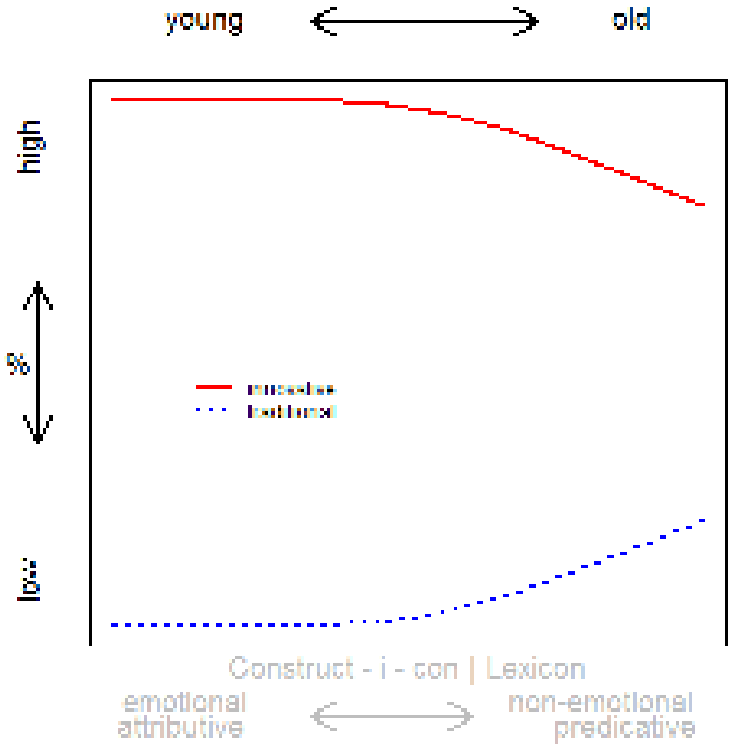
Diffusion of Innovations



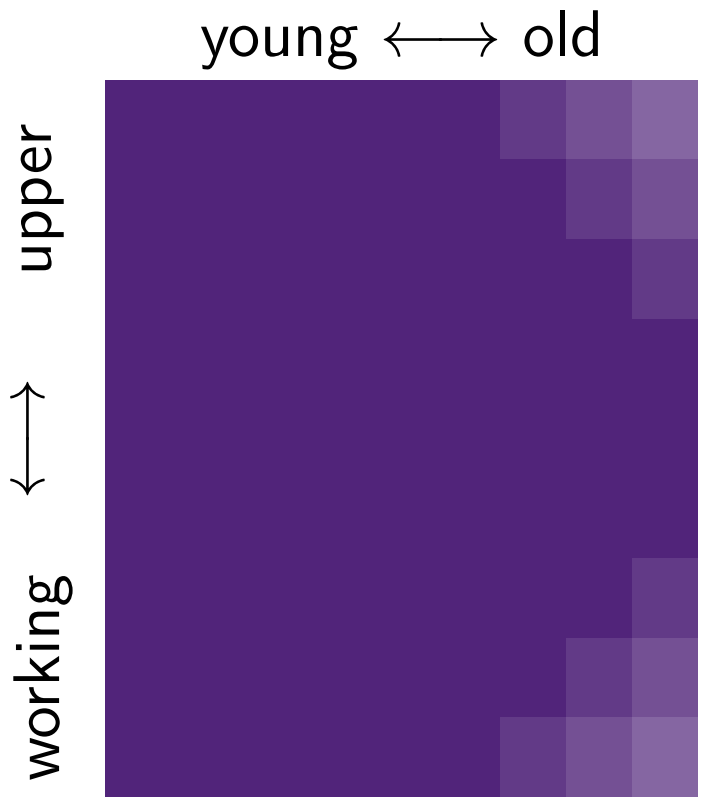


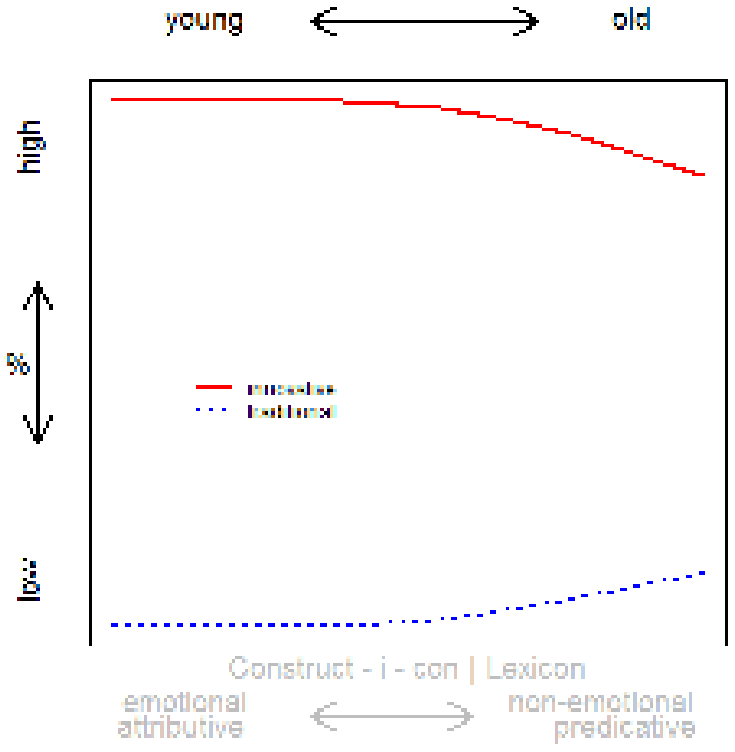
Diffusion of Innovations



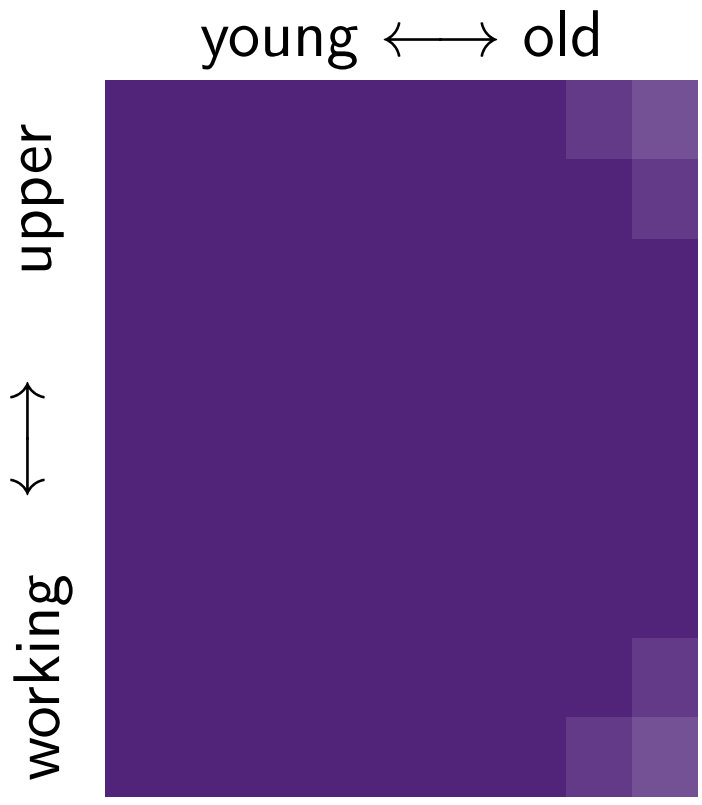


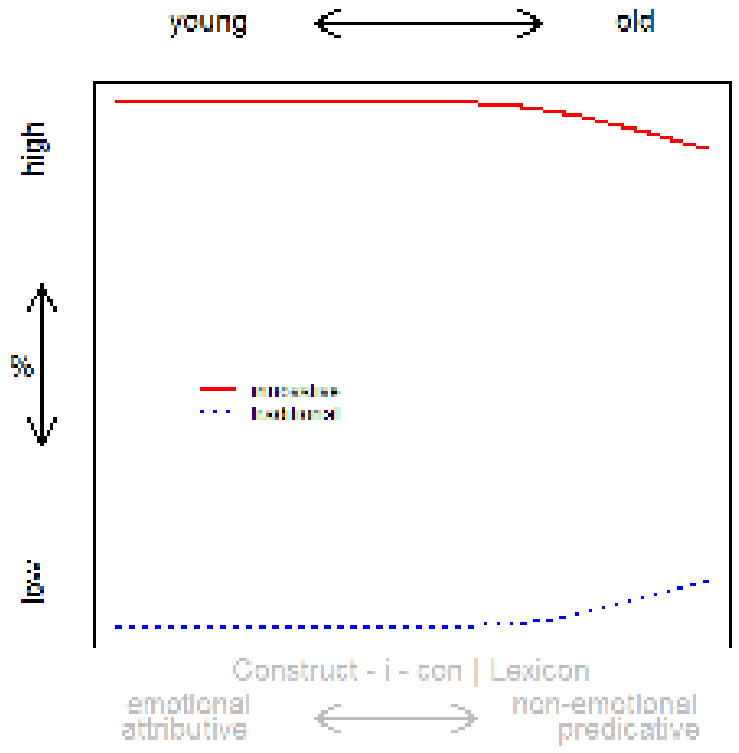
Diffusion of Innovations



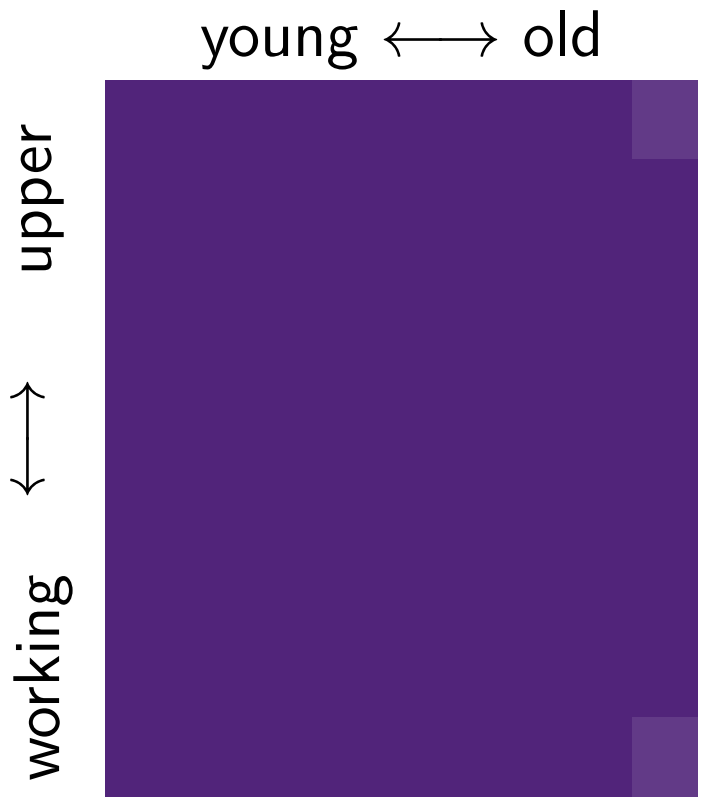


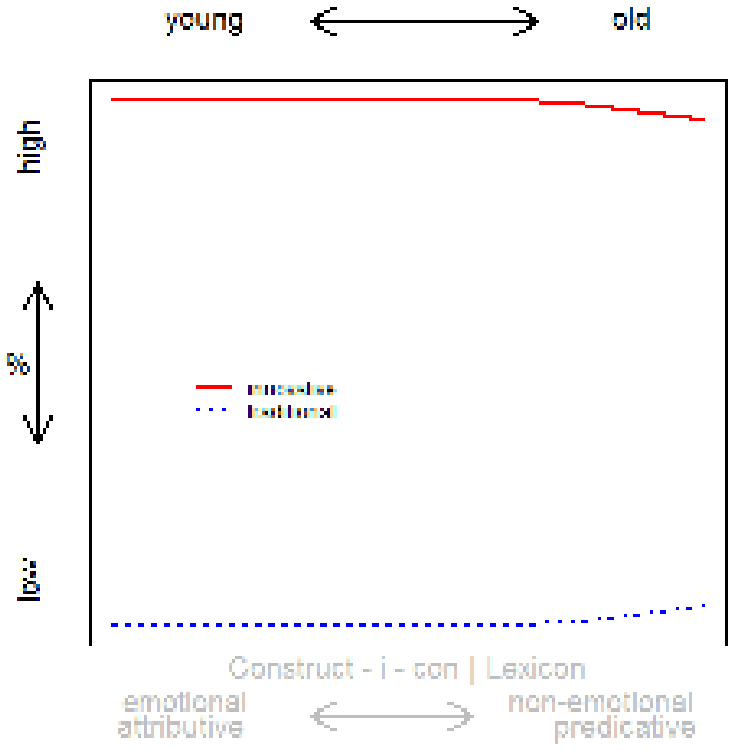
Diffusion of Innovations



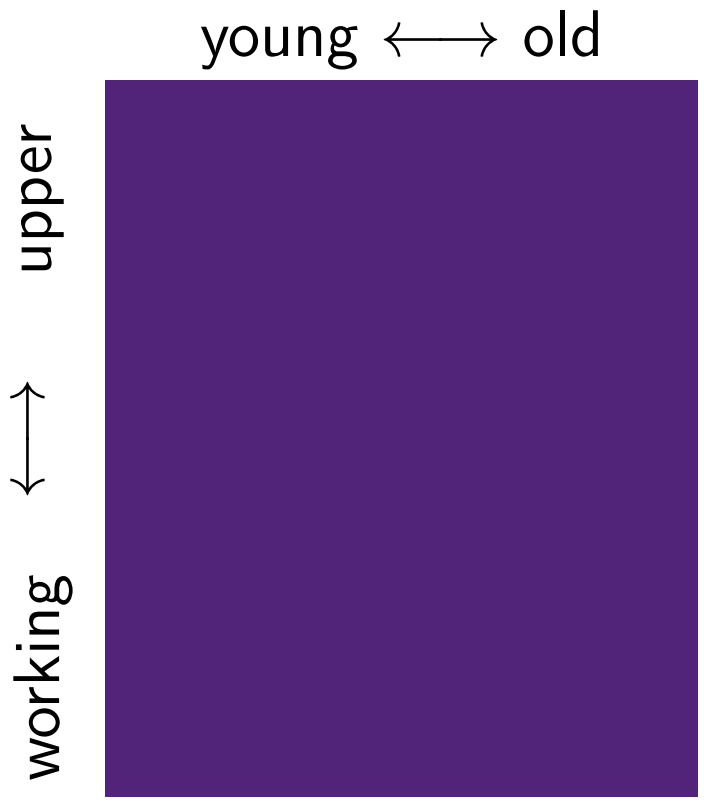


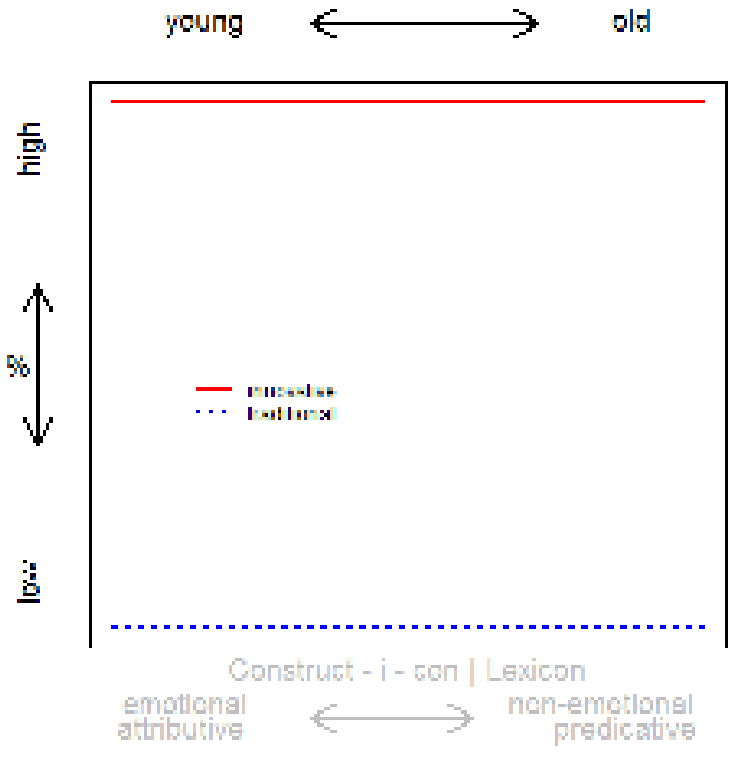
Diffusion of Innovations





Diffusion of Innovations





Adjective Amplification in Australian English

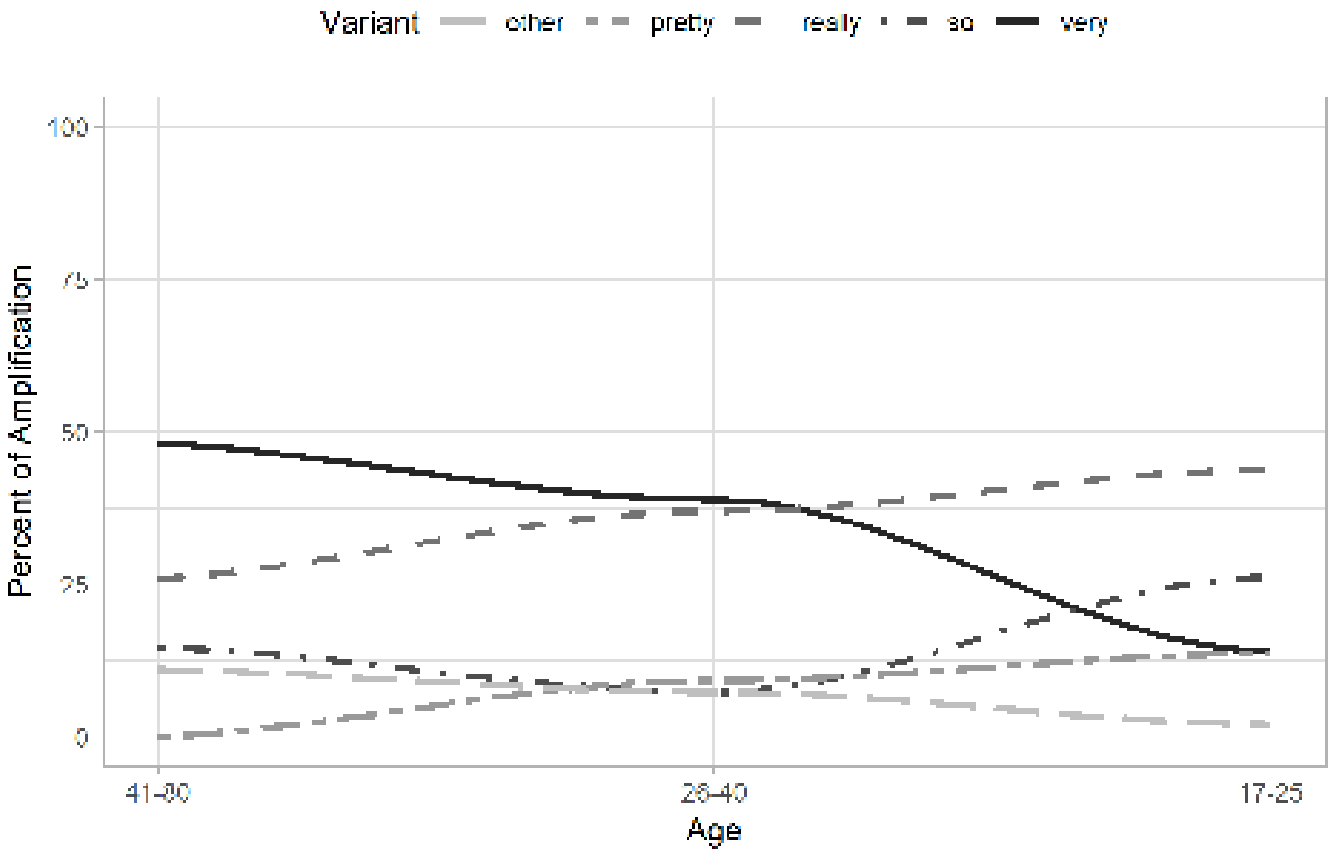


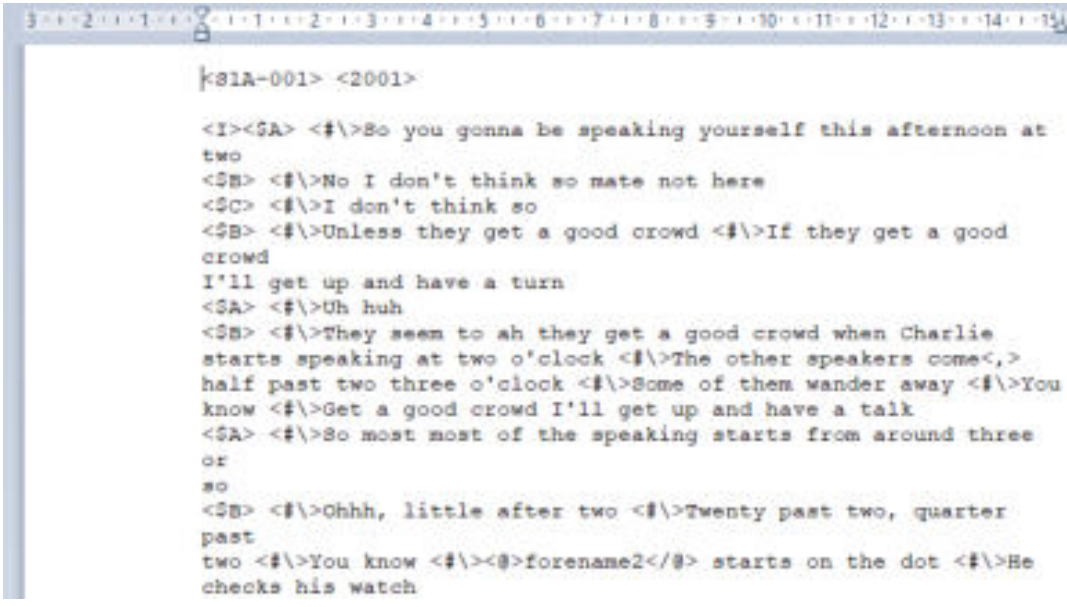
Figure 1: Adjective Amplification In AusE by age of speaker.

Corpus Linguistics

Corpus Linguistics

What do I use to investigate language change?

→ Corpora



```
<S1A-001> <2001>

<I><$A> <#\>So you gonna be speaking yourself this afternoon at
two
<$B> <#\>No I don't think so mate not here
<$C> <#\>I don't think so
<$B> <#\>Unless they get a good crowd <#\>If they get a good
crowd
I'll get up and have a turn
<$A> <#\>Uh huh
<$B> <#\>They seem to ah they get a good crowd when Charlie
starts speaking at two o'clock <#\>The other speakers come<,>
half past two three o'clock <#\>Some of them wander away <#\>You
know <#\>Get a good crowd I'll get up and have a talk
<$A> <#\>So most most of the speaking starts from around three
or
so
<$B> <#\>Ohhh, little after two <#\>Twenty past two, quarter
past
two <#\>You know <#\><@>forename2</@> starts on the dot <#\>He
checks his watch
```

- Corpora are digitized collections of texts
- I use transcriptions of spoken conversations that are accompanied by socio-demographic information about the speakers (age, gender, education level, socio-economic status, etc.)

Advantages

- Cheap and relatively easy to analyze
- Allow other researchers to check what I have done (full transparency & allows replication)

Corpus Linguistics

What can you do with corpora?
→ e.g. data-driven classification

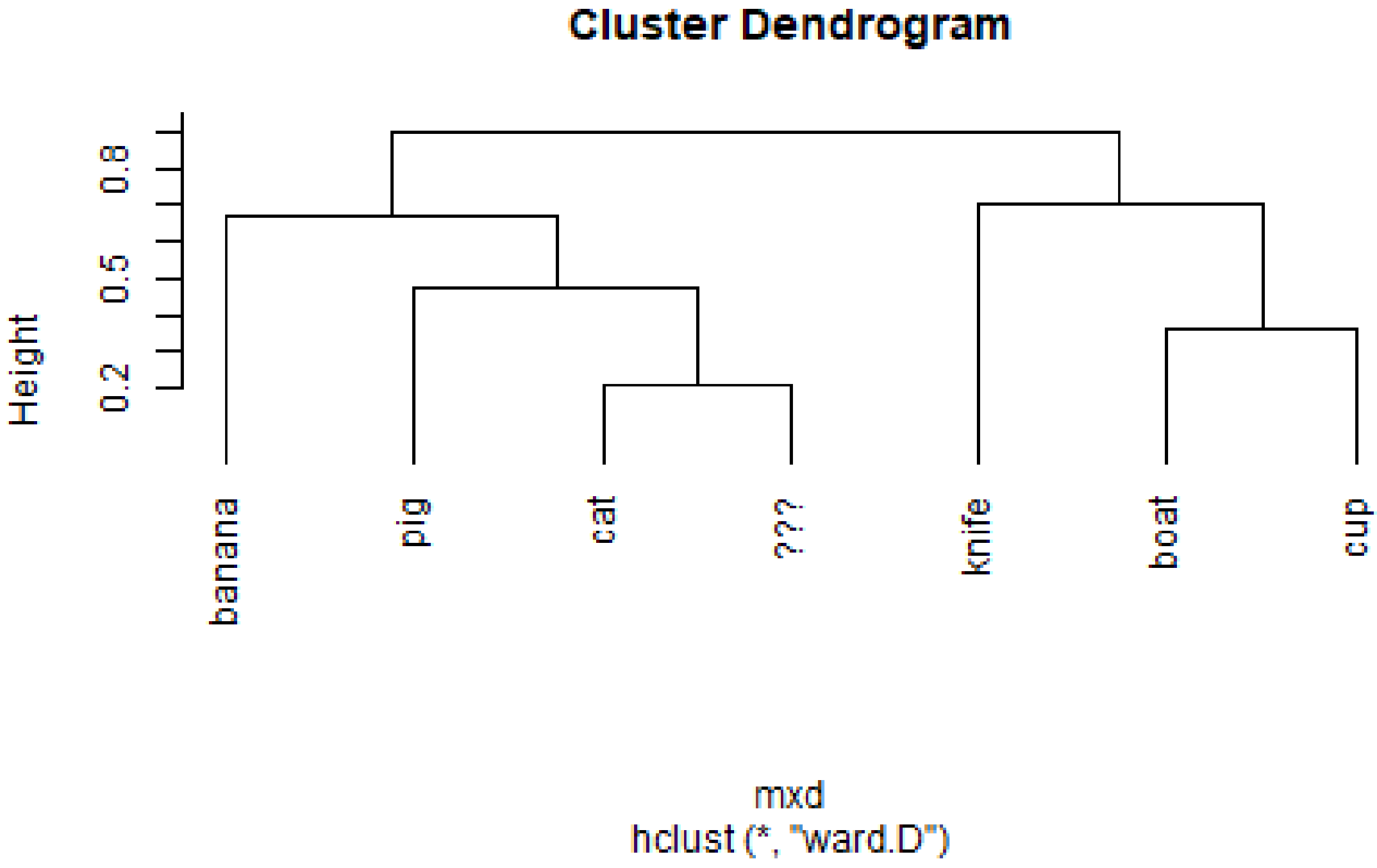
	get	**see**	**use**	**hear**	**eat**	**kill**
knife	31	16	69	0	2	0
cat	36	38	4	4	6	20
???	66	58	9	34	28	12
boat	46	21	17	4	0	0
cup	59	6	5	1	1	0
pig	4	15	3	1	7	21
banana	7	2	2	0	12	0

Table 1: Co-occurrences of selected nouns and verbs in the British National Corpus.

How data-driven classification works

	knife	**cat**	**???**	**boat**	**cup**	**pig**
cat	.62					
???	.60	.21				
boat	.48	.33	.32			
cup	.76	.59	.58	.36		
pig	.72	.36	.45	.64	.88	
banana	.71	.57	.47	.60	.72	.64

Table 2: Distance matrix based on co-occurrences of selected nouns and verbs in the British National Corpus.


Cluster Dendrogram
Height
0.2
0.5
0.8
banana
pig
cat
???
knife
boat
cup
mxd
hclust (*, "ward.D")

Q_1

How can this help in the analysis
of adjective amplification?

$\rightarrow$ Are there meaningful clusters of amplifiers?

Are all adjective amplifiers the same?

(4) That's very good!

(5) That's really good!

(6) ?That's completely good!

(7) ?That's absolutely good!

(8) That's absolutely amazing!

(9) ?That's very amazing!

Data Processing

- Part-of-speech tagged every word
- Extracted all adjectives
- Identified adj. preceded by an amplifier
- Determined the type of amplifier
- Tabulated co-occurrences of amplifiers and adjectives
- Visualized results

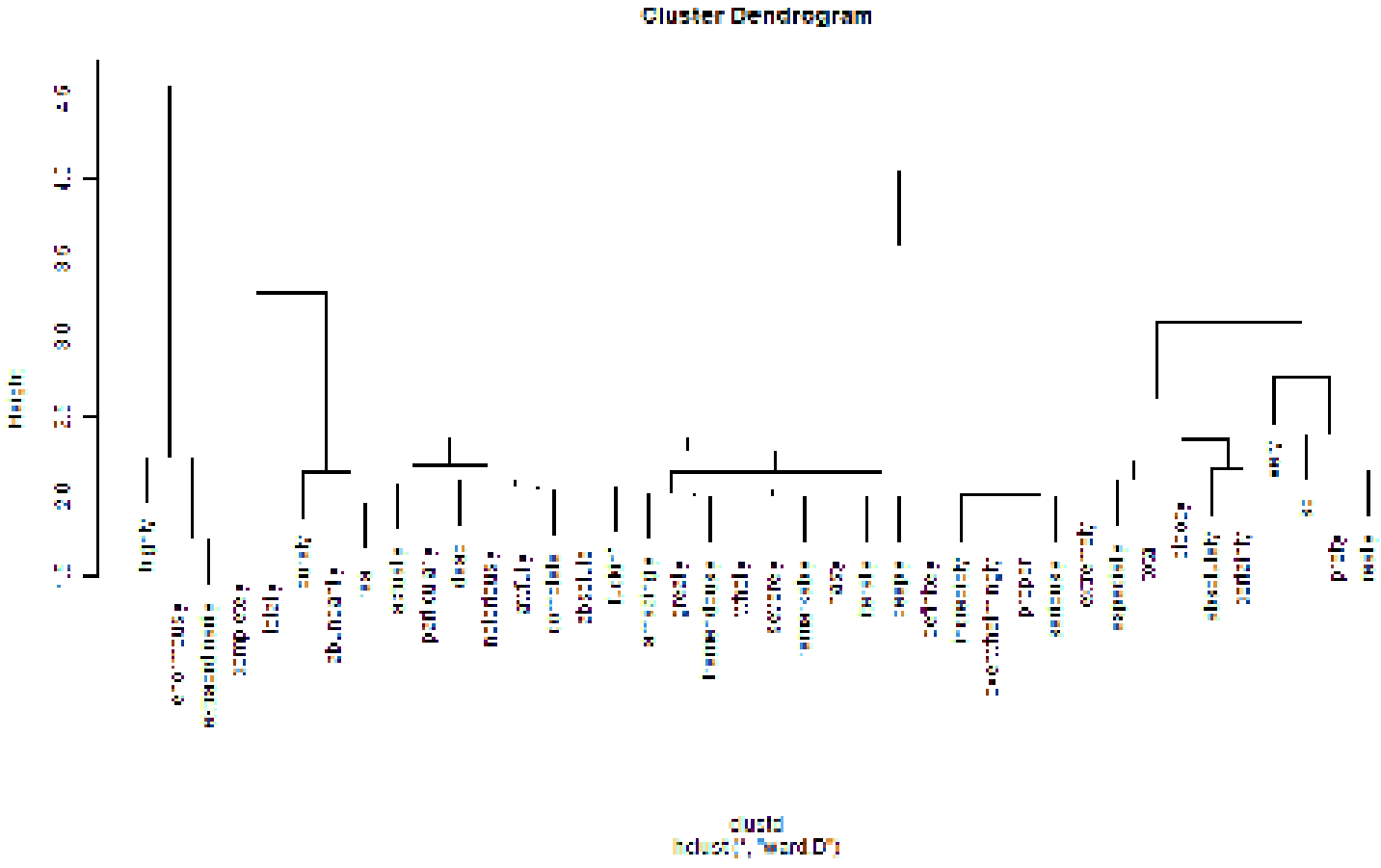


Figure 2: Rooted dendrogram showing the clustering of amplifiers in Australian English based on the semantic vector space model.

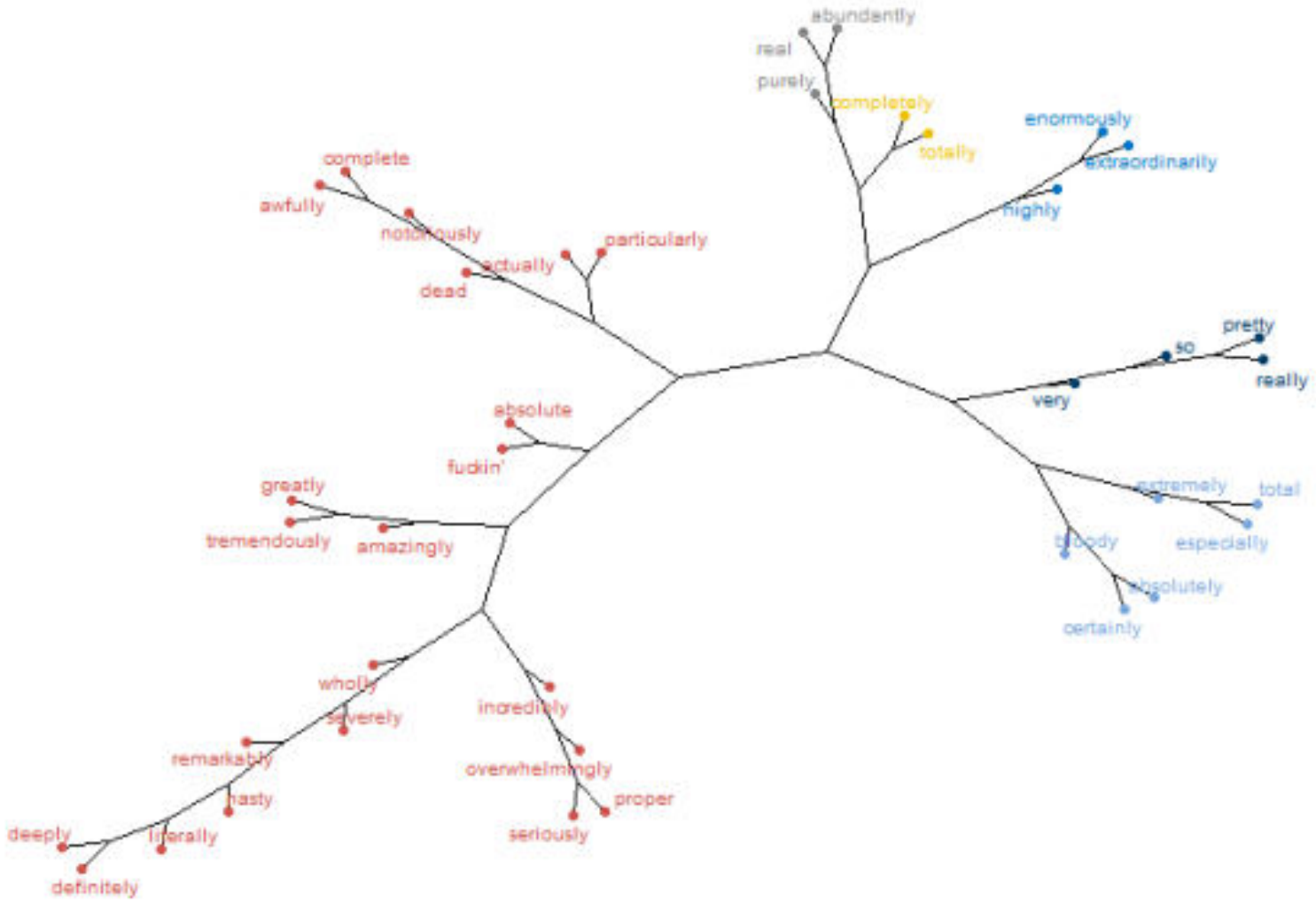


Figure 3: Unrooted dendrogram showing the clustering of amplifiers in Australian English based on the semantic vector space model.

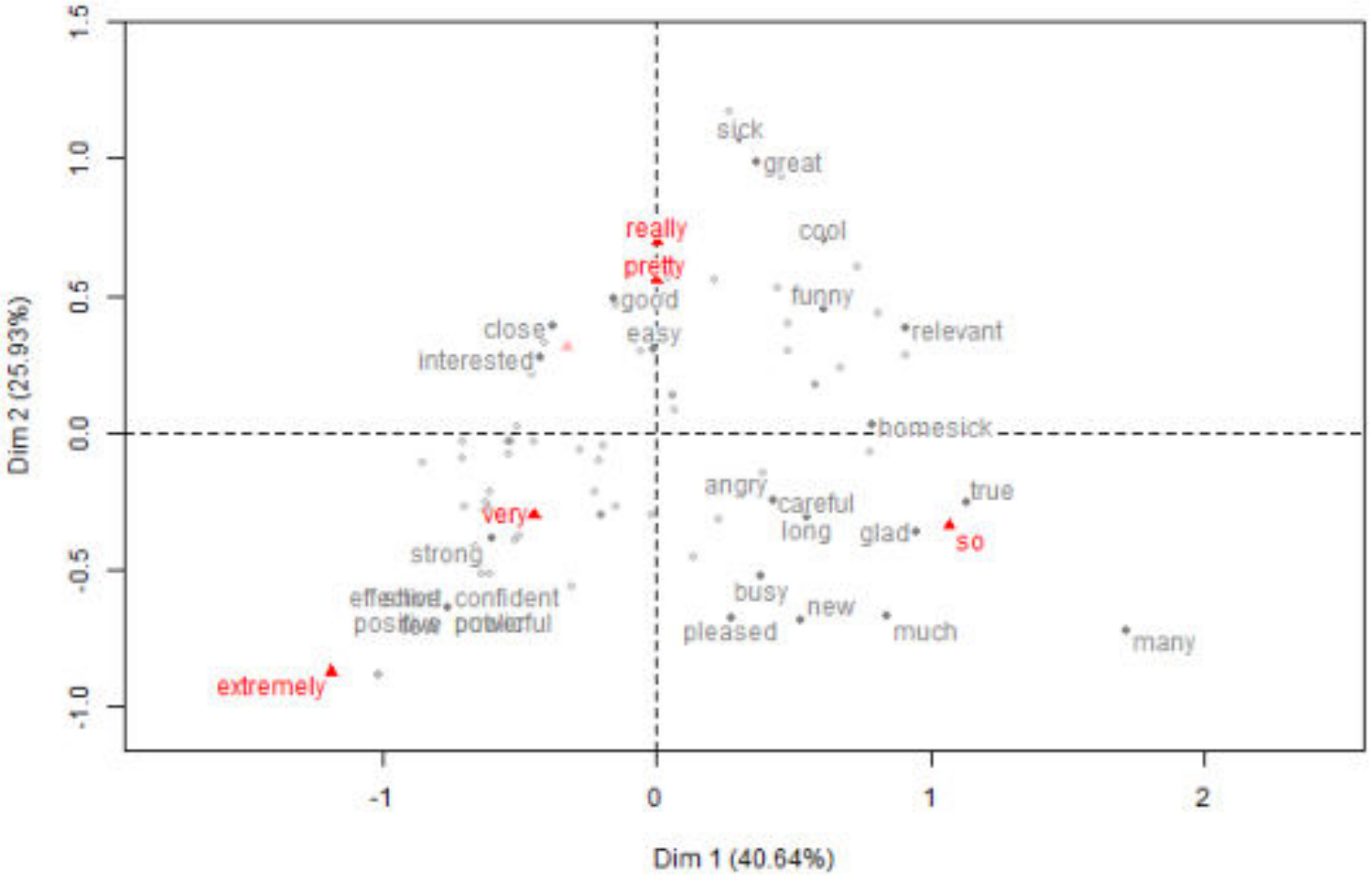


Figure 4: Results of a correspondence analysis based on amplifiers and their co-occurrences with adjectives in Australian English.

Q_2

What is this good for?

How and where do learners of English (NNS) differ from native speakers (NS) with respect to adjective amplification?

Amplification in Second Language Acquisition (Language Learning and Teaching)

Data

- *International Corpus of Learners of English* (ICLE)
  - 2.5 mil. words representing argumentative writing by intermediate to advanced Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, and Swedish learners of English
- *Louvain Corpus of Native English Essays* (LOCNESS)
  - 290.000 words of argumentative essays by American and British university students and British A-level students
  - LOCNESS was specifically designed to allow meaningful comparisons between the learner data represented in the ICLE.

- Processing as described above.

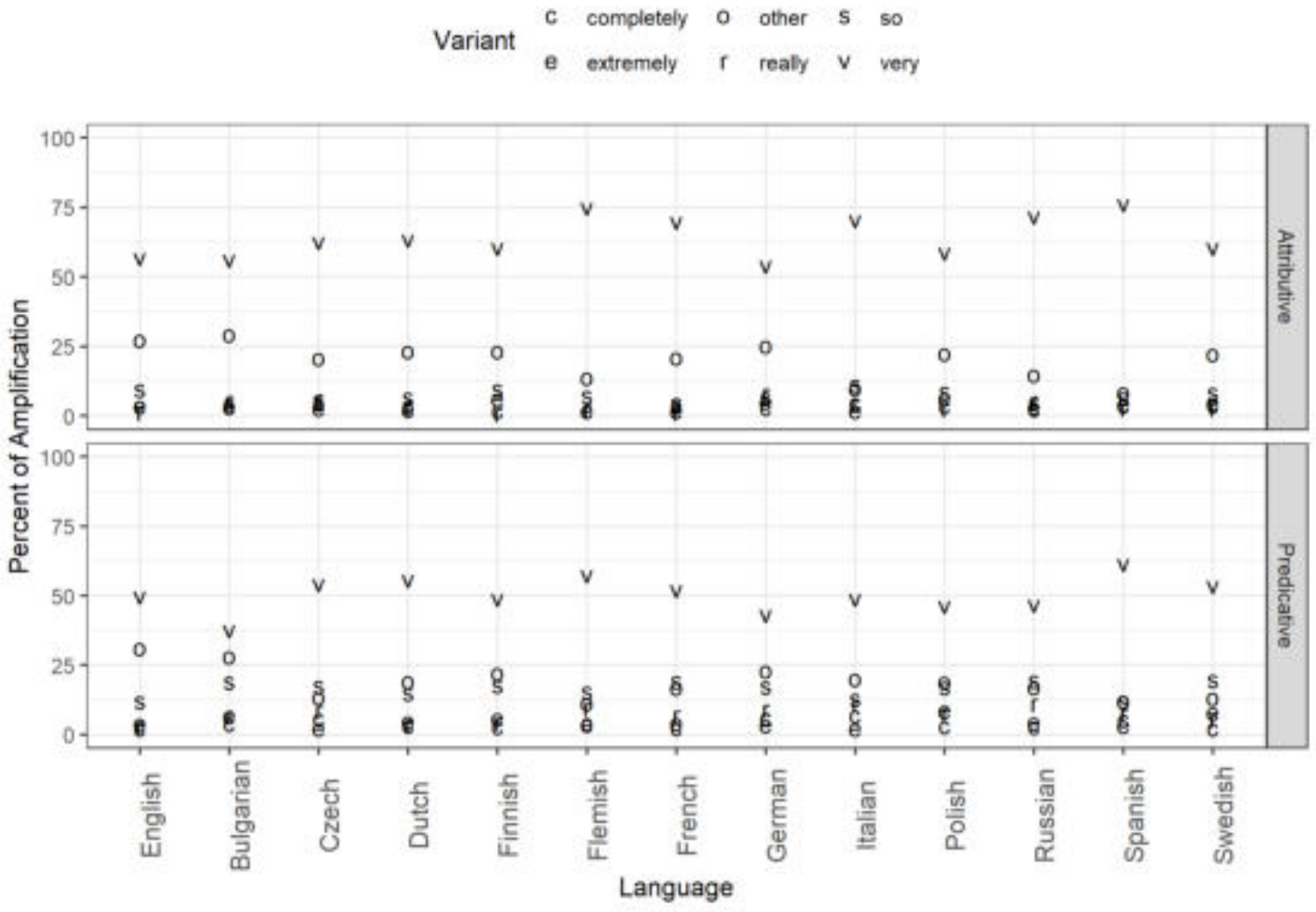


Figure 5: Percentages of amplifiers by L1-background.

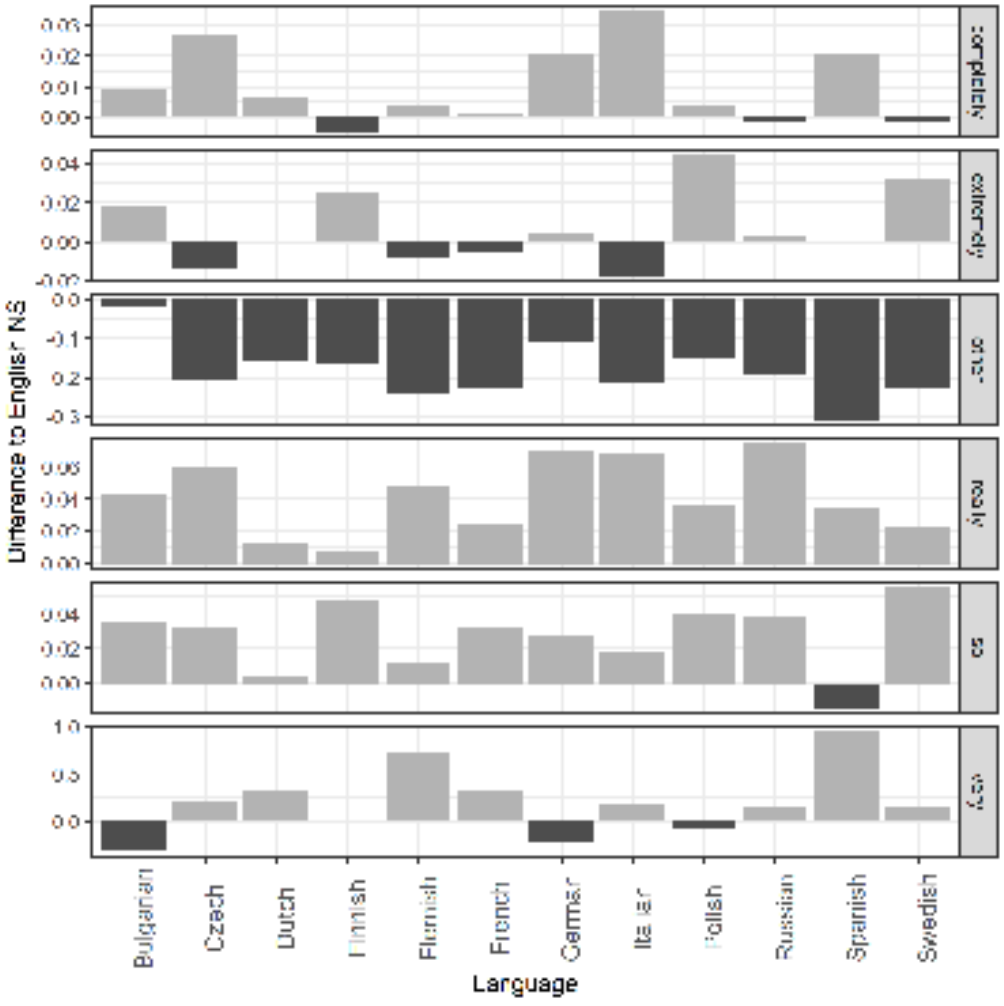


Figure 6: Bar graphs showing the difference to expected frequencies of amplifier types based on NS use.

The Language Technology and Data Analysis Laboratory (LADAL)

LADAL

What is LADAL?

- HASS eResearch support infrastructure for digital HASS at the UQ School of Languages and Cultures
- Targeted at humanities researchers
- Offers pathways into new research possibilities
  - Specialist computing lab for language-based computational and experimental work (the Computational and Experimental Workshop)
  - Online virtual lab (the LADAL website `https://slcladal.github.io/index.html`

LADAL

Services

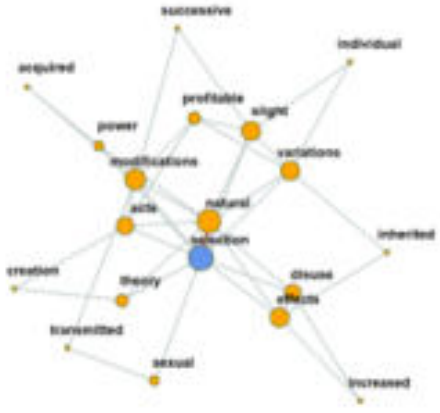
- Specialized training/support (workshops) on digital research methods and technologies
- Information and self-guided study materials
- Hands-on practical tutorials on topics relating to digital tools, computational methods for data extraction and processing, data visualization, and statistical analyses (learning to "code")
- Face-to-face consultations

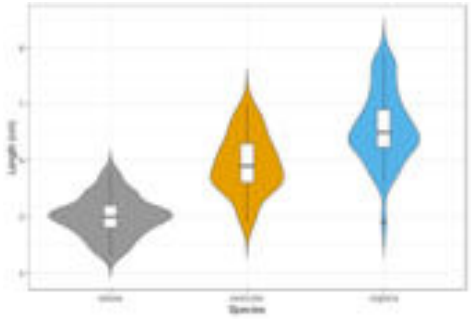
LADAL

```
# load function for concordancing
source("https://slcladal.github.io/rscripts/ConcR_2.5_LoadedFiles.R")
# start concordancing
darwinorganism <- ConcR(darwin, "organism[s]{0,1}", 50)
# inspect data
darwinorganism[1:5, 2:ncol(darwinorganism)]
```

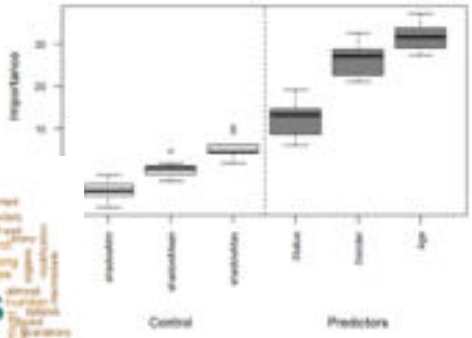
Aims of LADAL?

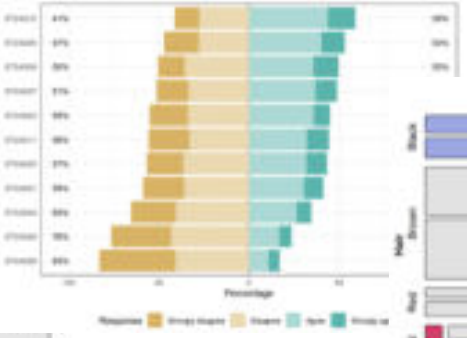
- Development of skills in
  - Digital tools and data management
  - Computational methods and (basic) programming skills
  - Data extraction / transformation / processing / analysis
  - Data visualization (including geo-spatial mapping and interactive web apps)
  - NLP applications (text analysis) and various statistical procedures (including classification and machine learning)

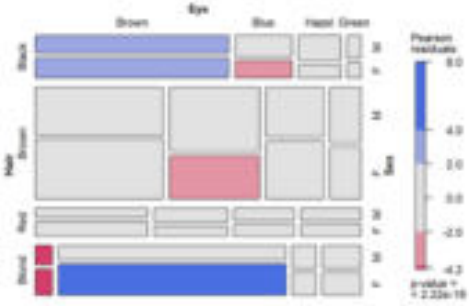

successive
individual
acquired
profitable
slight
power
variations
modifications
natural
acts
selection
inherited
creation
theory
disuse
effects
transmitted
sexual
increased

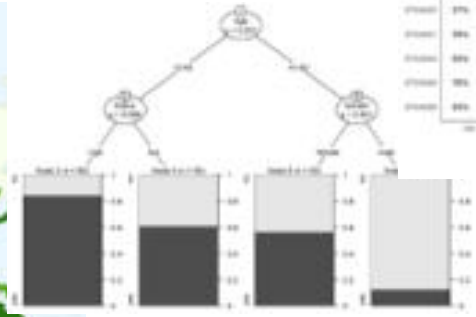



species


Importance
Control
Predictors




Eye
Brown
Blue
Hazel
Green
Hair
Black
Brown
Red
Blond



Thank you so, really, very much!

Computational Approaches to Textual Data

Dr. Martin Schweinberger

slides available at

www.martinschweinberger.de

The University
Of Queensland
AUSTRALIA