Universität Zürich[UZH]

Institut für Computerlinguistik

Text Crunching Center (TCC): Data-driven methods for social and medical science, linguistics and digital humanities

PD Dr. Gerold Schneider

Dr. Tilia Ellendorff

Institut für Computerlinguistik

gschneid@cl.uzh.ch

https://www.cl.uzh.ch/de/TCC.html

30.3.2022

Contents

Our Center
Our Experiences

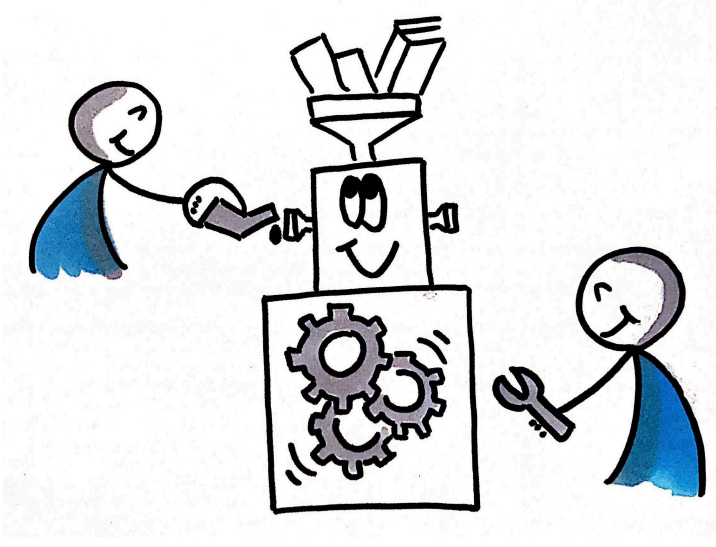
Our center: Text Crunching Center (TCC)

We offer our expertise in the following areas:

- Text Analytics
- Text Mining
- Sentiment Detection
- Digital Humanities
- Machine Translation
- ...

We offer consulting and support in

- Digitalisation
- Processing of text, including multilingual and historical texts
- Advice on tools, software and best practices
- Help with project applications and common projects
- Ready-made solutions



Our center: Text Crunching Center (TCC)

Institut für Computerlinguistik

We are a young platform (2020) and a small team:

- 2 postdocs: Gerold Schneider & Tilia Ellendorff
- several student helpers

Our customers:

- UZH internal (~ 50 % of our customers)
- Academic world-wide (~ 25 % of our customers)
- Private Industry (~ 25 % of our customers, ~ 50% of our revenue)

Our main tools:

- Python & many libraries
- R & many libraries
- GUI tools like LightSide and gephi

Our added value:

- We are more efficient than e.g. a doctoral student
- Applied: we use state-of-the-art methods with a focus on performance and applicability
- We have many small, precisely defined projects & services (e.g. OCR of XXX pages)
- Close collaboration with other platforms, particularly UZH LiRI, and all linguistics institutes

Our data: Often unstructured, e.g. clinical texts

Textual data is also called unstructured and often unlabelled data.

Besides supervised methods, unsupervised, data-driven methods are particularly suitable to explore these:

- Often no clear “gold standard” category to predict. Diagnosis?
- Many of our customers have no hypotheses → exploratory research
- *Befund vs. Befinden*: the patients’ individual experiences are far more than binary sentiment or the diagnosis
- New patterns and correlations can be detected, and assumed ones confirmed.
- Help the patient: coping strategies, find people with similar concerns

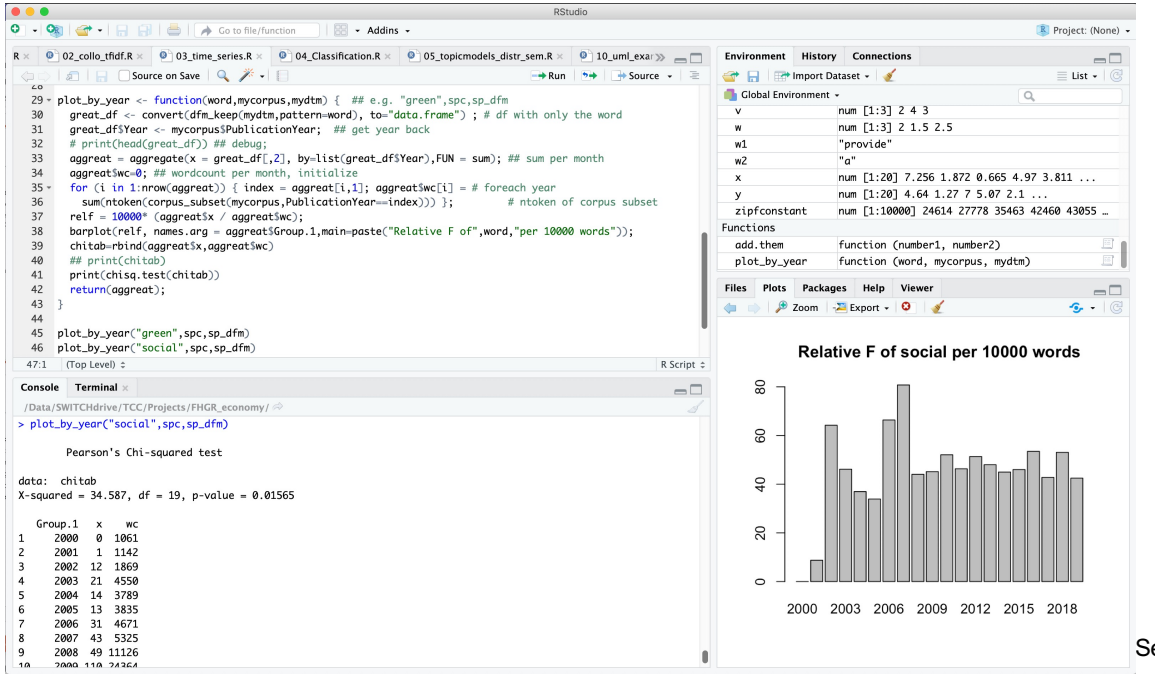
Our data: Often unstructured, e.g. clinical texts

- Ample amounts of rich texts have not been exploited much yet:

"Every day across the country, doctors are seeing patients and carefully documenting their conditions, social determinants of health, medical histories and more into electronic health records (EHRs). These documentation-heavy workflows produce rich data stores with the potential to radically improve patient care. The bulk of this data is not in discrete fields, but rather free text clinical notes. Traditional healthcare analytics depends predominantly on discrete data fields and occasionally regular expressions for free text data, missing a wealth of clinical data." (Katie Morris Claveau, https://towardsdatascience.com/clinical-natural-language-processing-5c7b3d17e137)

Our experiences

- More exploratory, unsupervised research than we had expected
- Many customers do not have hypotheses → exploratory
- Coaching is more demanded than ready-made solutions: customers want to be part of the process, be able to use the methods afterwards, knowledge transfer
  - R Notebooks
  - Jupyter Notebooks
  - Additional tools
  - Showcase with customers' data

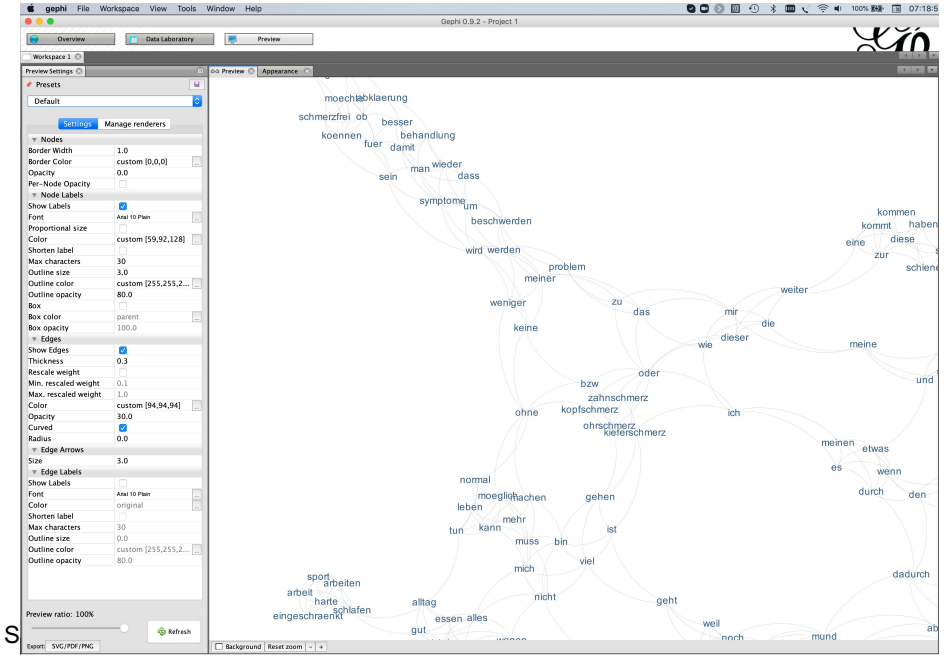


Our experiences

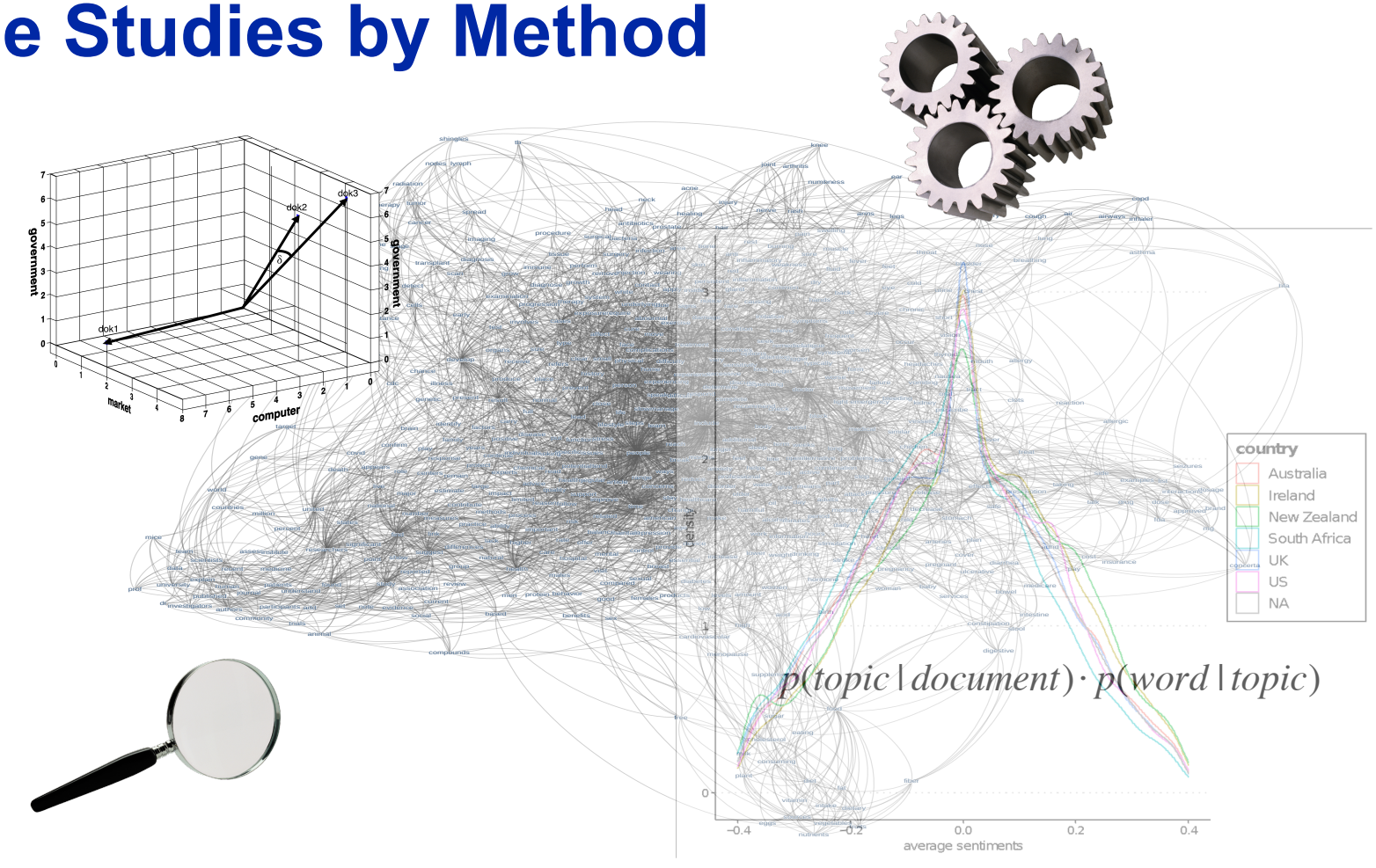
- Also teaching of workshops and creating online teaching material is demanded
- High demand for common project applications (Innosuisse, DSI Initiative, SNF)
- Challenges in exploratory / data mining research: expectations are often too high or too low

  Wow – exciting !! ←→ Yes, but this is not doing the magic we expected

- The background and previous knowledge and skills of customers are very varied
- Technical coaching for non-technical people can be challenging
- After initial talks, it is hard to draw the line / assess if the potential customers will accept the offer

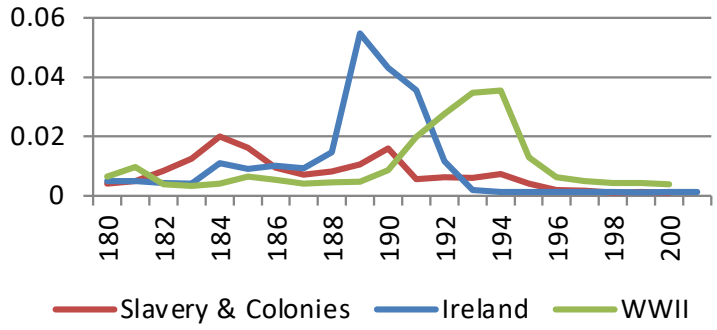


Case Studies by Method

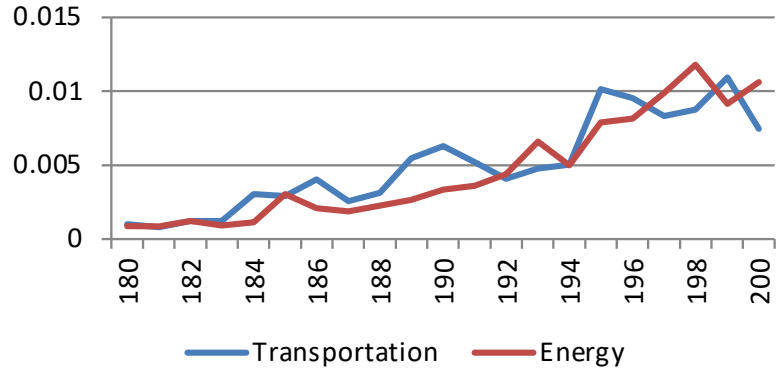


Topic Modelling: British Parliamentary Debates (Hansard Corpus, 1803-2015)

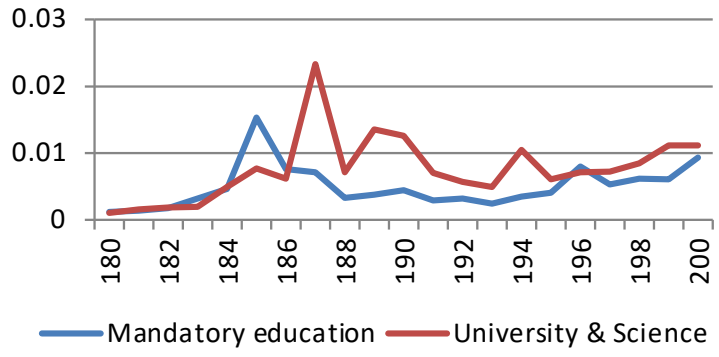
Historical facts



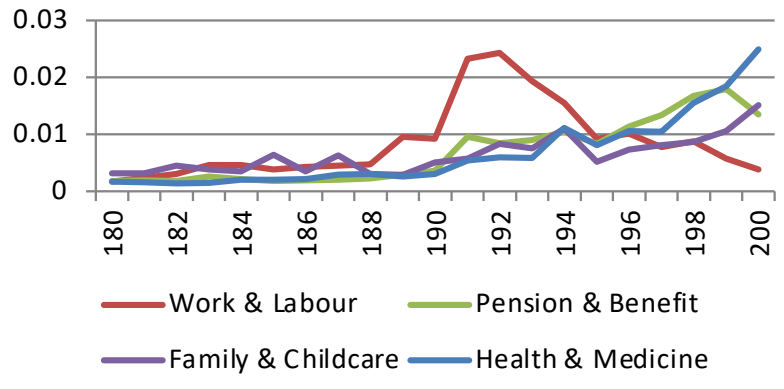
Democratisation through technical innovations



Democratisation through education



Democratisation through social protections



Topic Change and democratisation (indeces)

Pearson correlations between topic change and democratisation and economic indeces

	topic 0	topic 1	topic 6	topic 12	topic 15	topic 33	topic 50	topic 64	topic 78
Pearson with DEMOC	0.22	-0.13	-0.41	-0.87	-0.85	-0.72	0.37	0.34	-0.49
Pearson with POLITY	0.21	0.05	-0.36	-0.88	-0.87	-0.77	0.36	0.41	-0.40
Pearson with GDP	0.32	-0.31	-0.19	-0.54	-0.58	-0.55	-0.11	0.27	-0.62
Pearson with EDUC	0.03	-0.42	-0.36	-0.66	-0.70	-0.73	-0.07	-0.02	-0.78
Key-words	university	ireland	election	court	government	bank	war	school	colony
	student	secretary	vote	case	treaty	debt	Germany	education	slave
	education	chief	constituency	judge	France	money	World	child	governor
	research	county	candidate	offense	Spain	interest	peace	teacher	island
	science	dublin	representation	justice	russia	exchequer	nation	parent	canada

Pearson correlations between topic change and democratisation and economic indeces

	topic 20	topic 30	topic 38	topic 39	topic 44	topic 46	topic 51	topic 59	topic 72	topic 74	topic 92
Pearson with DEMOC	0.60	0.74	0.85	0.88	0.76	0.32	0.88	0.72	0.73	0.71	0.91
Pearson with POLITY	0.52	0.67	0.77	0.85	0.67	0.38	0.80	0.67	0.65	0.63	0.87
Pearson with GDP	**0.96**	0.32	0.85	0.73	**0.98**	0.60	0.88	0.56	0.99	0.85	0.65
Pearson with EDUC	**0.94**	0.36	0.93	0.80	**0.96**	0.52	0.93	0.64	0.98	0.81	0.76
Key-words	European	defence	pension	railway	health	police	water	trade	information	child	company
	community	air	benefit	transport	service	crime	oil	union	report	woman	business
	European	aircraft	people	service	hospital	officer	energy	worker	document	family	private
	union	force	insurance	line	medical	force	gas	strike	survey	young	profit
	policy	equipment	allowance	rail	patient	home	electricity	industrial	record	marriage	firm

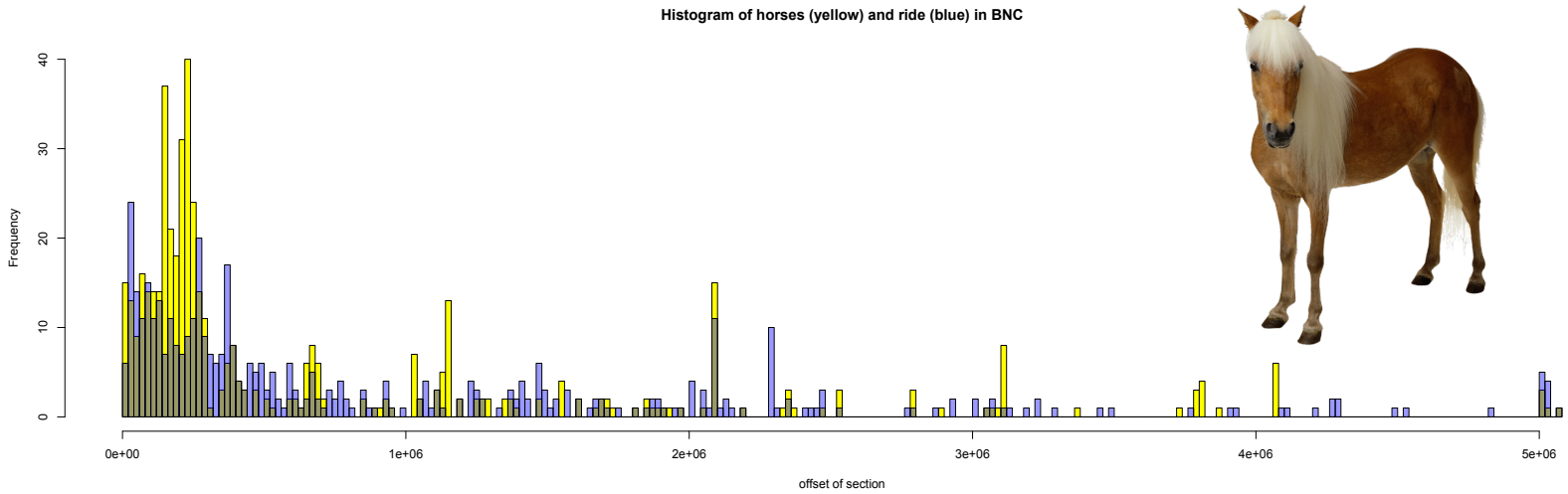
Schneider, Gerold and Maud Reveilhac (accepted for publication). “Colloquialisation, Compression and Democratisation in Parliamentary Debates”. In *Exploring Language and Society with Big Data: Parliamentary discourse across time and space*, ed. Minna Korhonen, Haidee Kotze and Jukka Tyrkkö. *Studies in Corpus Linguistics* Series. Amsterdam: Benjamins.

Conceptual Maps: Method

Kernel Density Estimation calculates semantic distances between words

Idea: semantically similar words often co-occur in the discourse → distributional semantics

E.g. *horse* and *ride* in the BNC



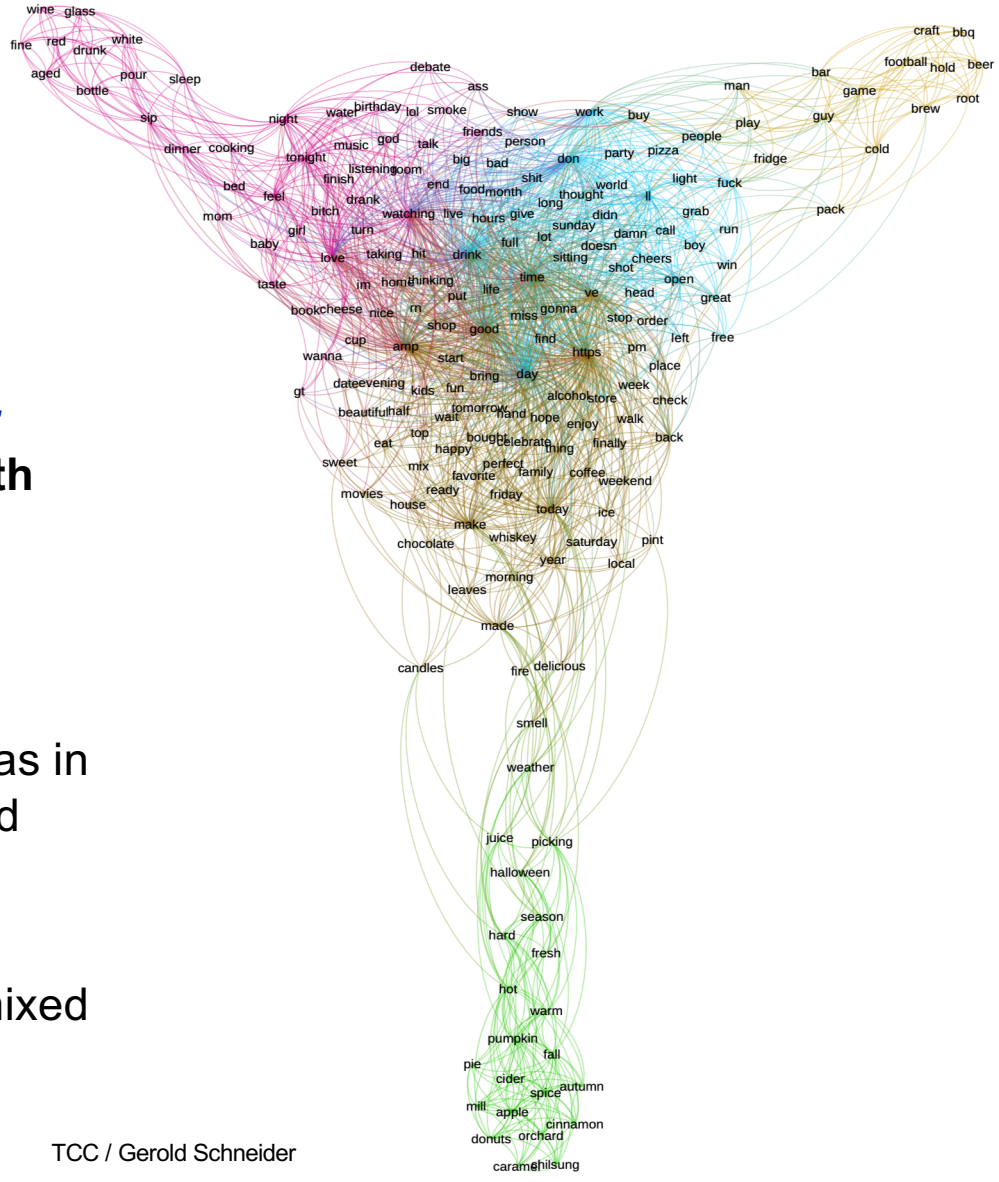
Kernel Density estimation is a popular smoothing method (Zucchini 2003)

Similar words are plotted close to each other on the arising concept map

We use textplot (http://dclure.org/tutorials/textplot-refresh/).) & gephi

Conceptual maps

Tweets on *beer, cider, wine:*
Kernel Density Estimation with textplot and gephi



Network of associations.

Other networks are also possible, as in classical network analysis: selected concepts, NER, topics (from topic modelling)

Meta- and content words can be mixed

Conceptual Maps with DIPEx CMI (Intensive Care Unit)

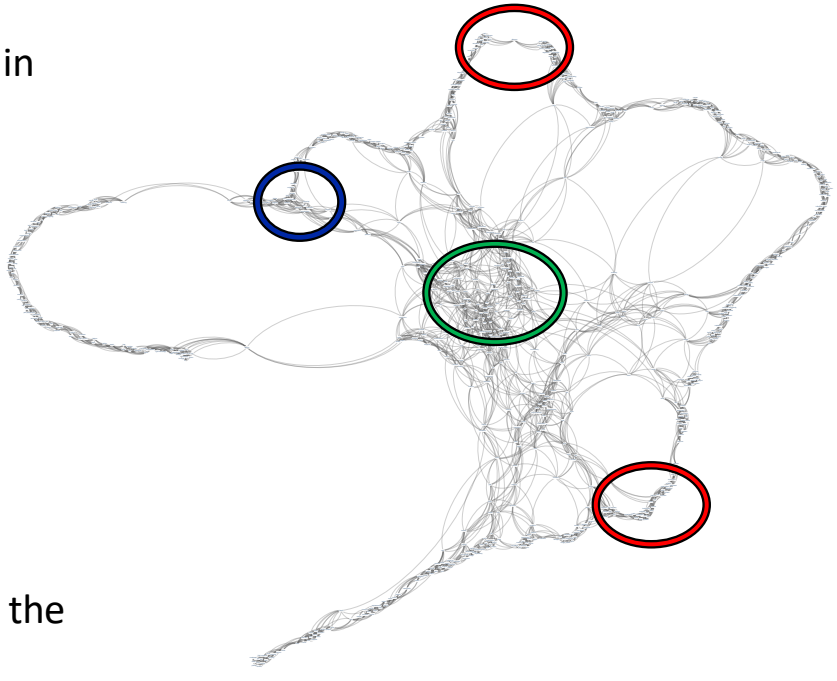
DIPEx = DB of International Patient Experiences

DIPEx CMI: Patient Interviews about their experiences in intensive care

only 8000 words, way too small for distributional semantics ...
but can we explore the space between words and concepts, plot the plots?

The conceptual map (500 w) shows

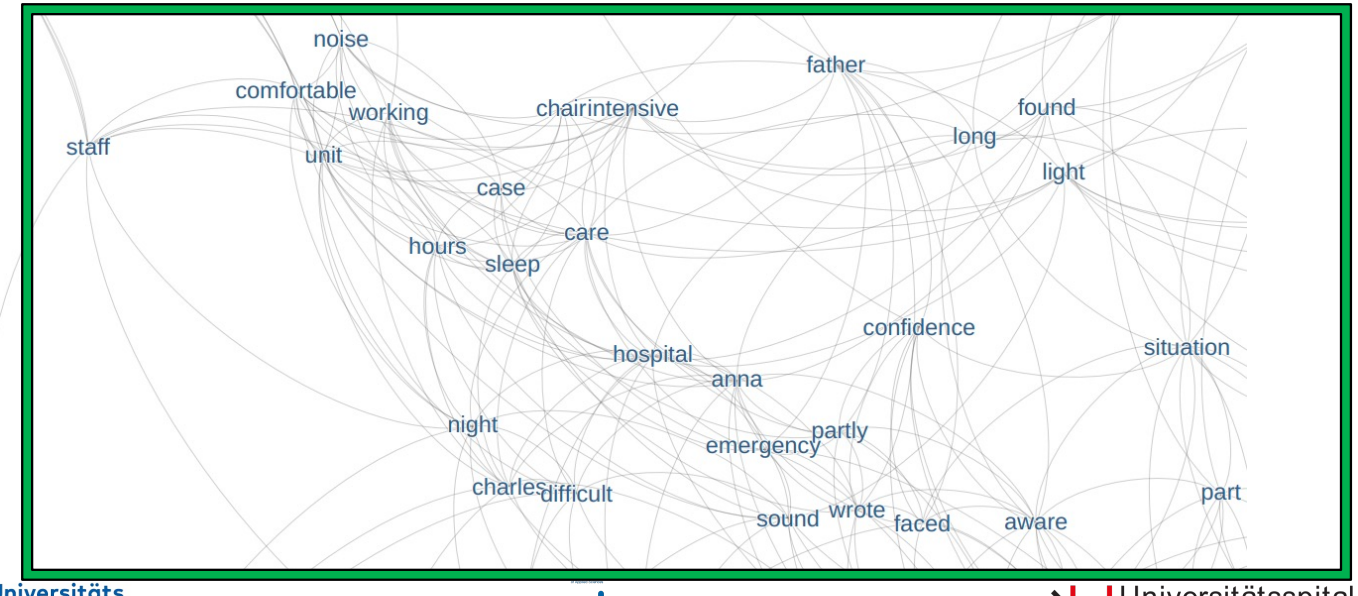
- Common core (green) in the centre
- Shared issues (blue) in the periphery
- Individual adventures (coma, hallucinations, red) in the edges



Conceptual Maps with DIPEx CMI & MS

The conceptual map (500 w) shows

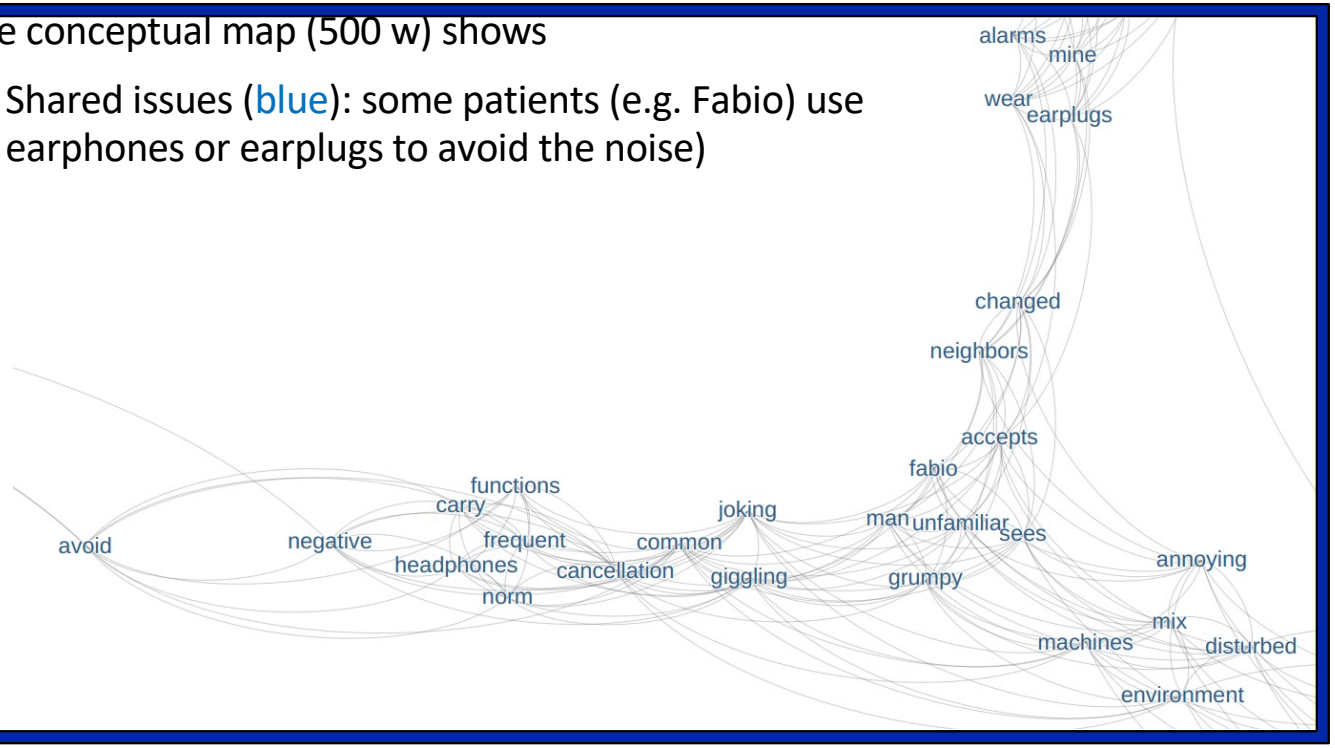
- Common semantic core (green): most patient (e.g. Charles or Anna) feel comfortable, but find it difficult to sleep, partly because of the noise



1.2 NLP Approaches to Analysis:
Teaser 3: Conceptual Maps with DIPEx CMI & MS

The conceptual map (500 w) shows

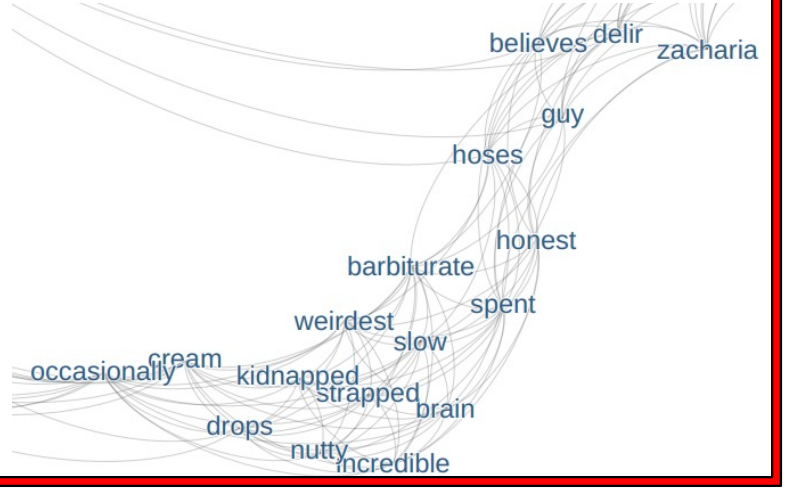
- Shared issues (blue): some patients (e.g. Fabio) use earphones or earplugs to avoid the noise)



1.2 NLP Approaches to Analysis:
Teaser 3: Conceptual Maps with DIPEx CMI & MS

The conceptual map (500 w) shows

- Individual adventures (coma, hallucinations, red):
  Zacharia, in delir, believes that he is kidnapped.
  He finds that incredibly nutty, himself.



1.2 NLP Approaches to Analysis:
Teaser 3: Conceptual Maps with DIPEx CMI & MS

From texts to maps … and back

"One of them was very funny, he had a curly beard, curly hair and was very, very sensitive, I noticed that … my friends who later came to visit me in the intensive care unit, they almost fell off their chairs when I first thanked them for being there and singing the song … my wife and children said stay here, stay here. And then I just prayed to God that He would give me another chance, that I would be allowed to return again"

Distributional Semantics (word2vec) with DIPEx Multiple Sclerosis patient interviews

```
> closest_to(training_ms20,"leben")
          word similarity to "leben"
1         leben             1.0000000
2  gelassenheit             0.5548906
3   beziehungen             0.5310467
4      ausmacht             0.5218526
5      negative             0.5194965
6     krankheit             0.5168033
7        dingen             0.5158901
8    lebenswert             0.5157932
9        lebens             0.5121614
10    positiven             0.5043837
```

Kommentar: die Ebene an Reife ist beindruckend: Gelassenheit, leben mit der Krankheit. Sich konzentrieren auf was das Leben ausmacht: Beziehungen zu Menschen, positive Dinge als lebenswert annehmen.

```
> closest_to(training_ms20,"koerper")
         word similarity to "koerper"
1      koerper               1.0000000
2       physik               0.5999420
3     koerpers               0.5681156
4         wahr               0.5043965
5      sondern               0.4641749
6  bewusstsein               0.4570825
7     staerken               0.4353815
8        power               0.4327700
9    benoetigt               0.4313358
10  intensiver               0.4302129
```

Kommentar: gegen die Physik kommt niemand an. Stärker wahrnehmen, intensiver erleben, das Bewusstsein stärken.

Document Classification

Detect pre-annotated, often binary classes. Even subtle semantic differences can often be detected; but depends heavily on data.

Teaser: US Speeches (< 2013)
Prediction of Party Affiliation (Republican / Democrat) based CORPS II Corpus: 8 mio words, 3618 Speeches (Guerini et al. 2013). We use speakers that have at least 10 speeches (and only American ones).

Logistic regression achieves 95-98% accuracy

Actual vs. Predicted	actual dem	actual rep
predicted dem	**1787**	36
predicted rep	131	**1294**

# Reden	Name
889	Bill Clinton
427	George W. Bush
388	Ronald Reagan
356	Dick Cheney
347	Barack Obama
316	John F. Kennedy
107	Michelle Obama
102	Margaret Thatcher
93	Laura Bush
61	Richard M. Nixon
53	Al Gore
51	Alan Keyes
43	Joe Biden
36	Condoleezza Rice
26	John Kerry
18	Hillary Rodham Clinton
13	Lynne Anne Vincent Cheney
13	Howard Dean
10	John Edwards

Die typisch republikanischsten Merkmale sind:

Merkmal	F-score
've	0.6455
're	0.6443
nation	0.6443
it_'s	0.6336
men	0.6333
–	0.6312
i_'m	0.6286
'm	0.6286
you_all	0.6273
freedom	0.6261
we_'re	0.6254
well	0.6224
<PERIOD>_he	0.6219
<PERIOD>_and	0.6203
great	0.6192
's	0.6177
one	0.6159
government	0.6158
america	0.6153
military	0.6147

Typisch **un**republikanischste Merkmale (Auswahl):

Merkmal	F-score
nra	0.0014
equal_pay	0.0014
of_climate	0.0014
racial_<COMMA>	0.0014
insurance_program	0.0014
high-wage	0.0014
our_steel	0.0014
without_health	0.0014
in_clean	0.0013
together_across	0.0013
campaign_finance	0.0013
to_hillary	0.0013
service_program	0.0013
fugitives	0.0013
stalkers	0.0013
our_planet	0.0013
financial_system	0.0013
after_high	0.0013
student_loans	0.0013
toxic_waste	0.0013
<PERIOD>_hillary	0.0013
from_welfare	0.0013
national_service	0.0013
more_police	0.0013

Language Use and Cognitive Aging
Method: quadratic regression

Healthy older adults VS younger adults

Research with Laura Luo & Mike Martin

Data:

Eight-minute conflict conversations

Unsupervised conversations while videotaped.

364 couples aged 19 to 82 (M = 48.24, SD = 18.33)

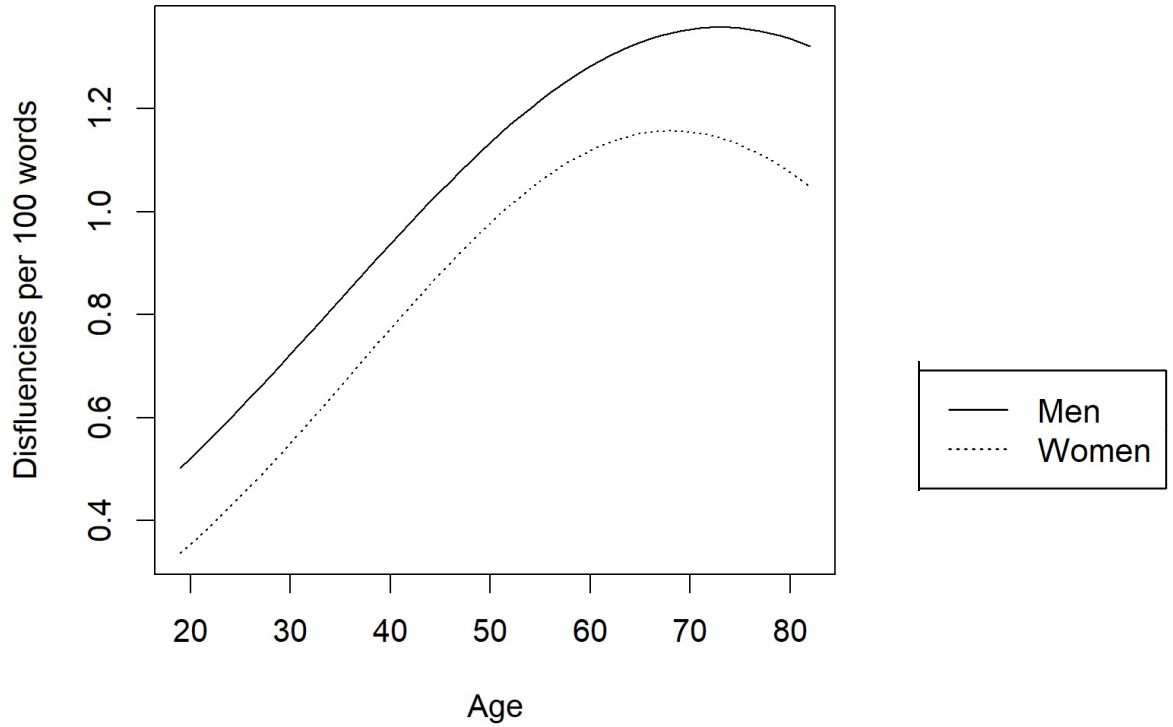
URPP Dynamics of Healthy Aging

Results:

- More unique words
- More uncommon words
- Simpler grammatical structures
- More utterance of disfluencies
- Increased crystalized intelligence, e.g., vocabulary & world knowledge
- Decreased fluid intelligence, e.g., working memory & fluency

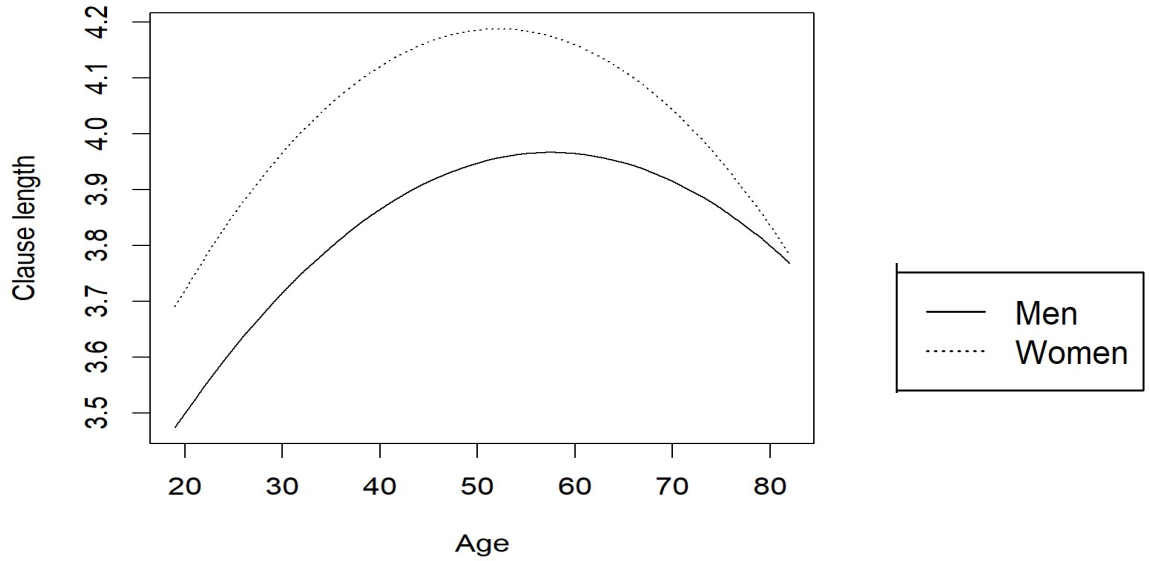
Stylistics: Language & Age: Disfluencies by quadratic regression

- The utterance of disfluencies increased until old age (Age_{women} = 68; Age_{men} = 73), and declined afterwards.
- The model fitted better than the model with only linear age effects ($\Delta\chi^2_{(2)} = 6.20, p < .05$).



Stylistics: Language & Age: Grammatical Complexity: Clause Length

- Clause length increased until midlife (Age_{women} = 52; Age_{men} = 58) and then declined.
- The model fitted better than the model with only linear age effects ($\Delta\chi^2_{(2)}$ = 16.99, p < .001).



Syntactic Complexity

Subordinating and Co-ordinating clauses

Subordinations are frequent in **hypotactic** style, which is typical for argumentation and elaboration. Co-ordinations are frequent in **paratactic** style, which is a characteristic of narration

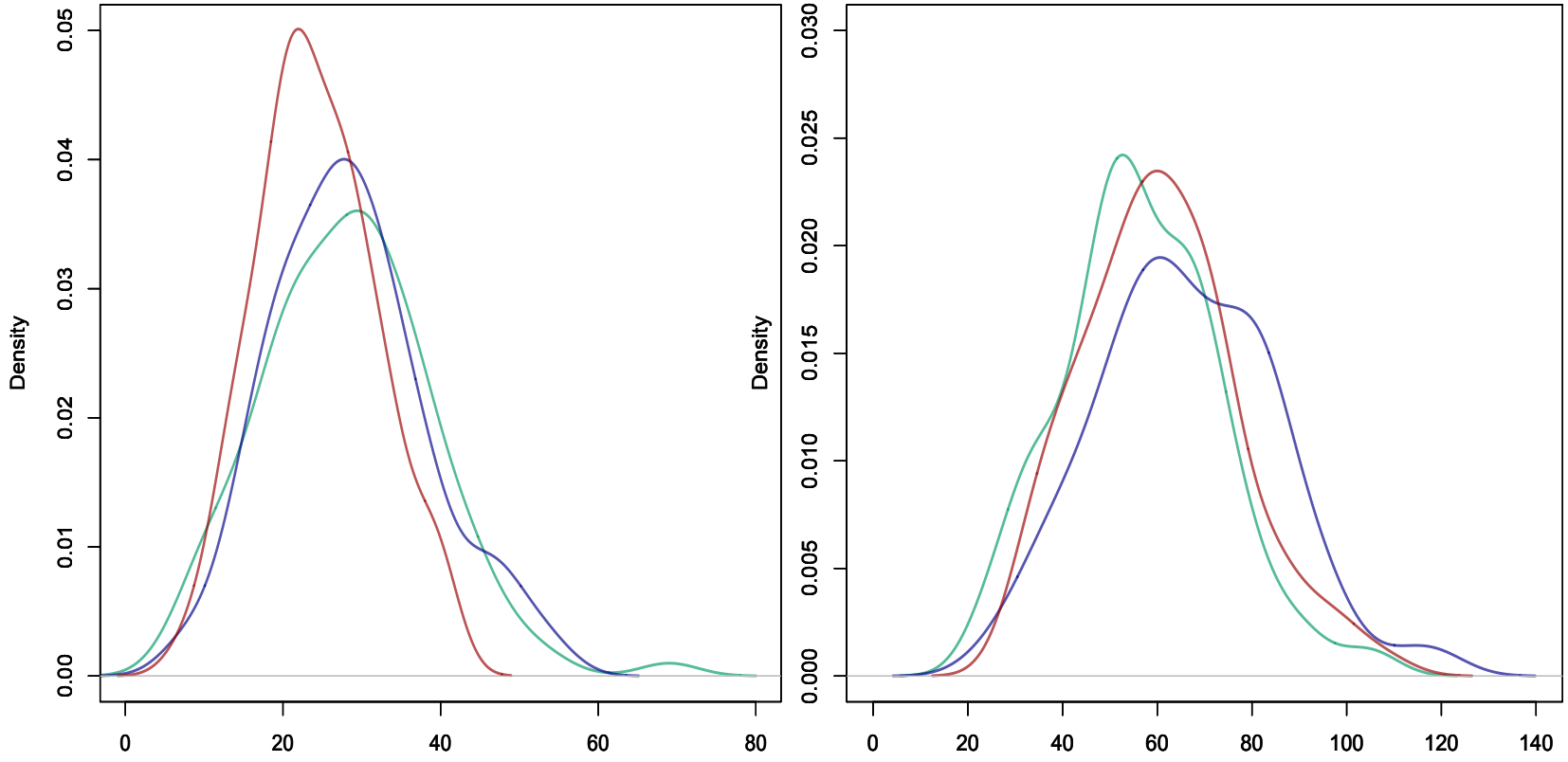
Group	Age	Ø Subordinations	∑Subord**	Subord/1K W
1	20-35	28.26	3109	20.60
2	40-55	28.45	3272	21.02
3	65-80	**24.29**	**2551**	**19.06**

Group	Age	Ø Conjunctions	∑Conj***	Conj/1K W
1	20-35	56.54	6220	41.22
2	40-55	66.07	**7598**	**48.81**
3	65-80	60.65	6368	**47.58**

Syntactic Complexity

Subordinating clauses (⬇️) & co-ordinating clauses (⬇️)

Per-couple, absolute counts (20-35, 40-55, 65-80)

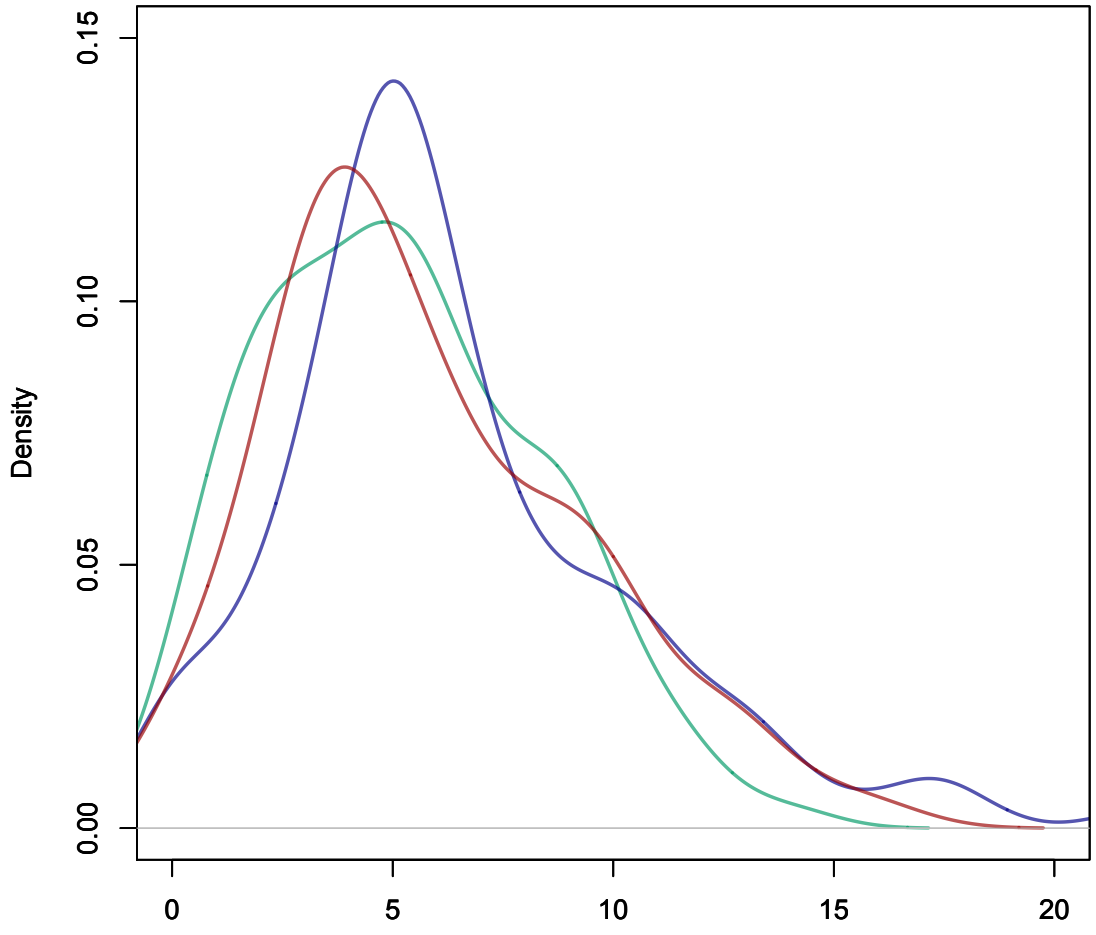


Syntactic Complexity

Relative clauses

(other subord.)

no significant
or marked
differences



Cognitive NLP Models: Method

A good model is parsimonious and obtains a good prediction of human cognitive behaviour. We use e.g. surprisal.

$$Bigram\ surprisal = log\frac{1}{p(w_1)} + \log\frac{1}{p(w_2|\ w_1)}$$

"the forward transitional probability $P(w_k|w_{k-1})$ is a simple form of surprisal" (Demberg and Keller, 2008)

- **Surprisal = *Surprise of word in its context*:** allows us to measure the competition between the idiom and syntax principle (Sinclair 1991)
- **Language dominated by the idiom principle** has low surprisal, many chunks, is easy to process & remember, but contains little information.
- **Language which makes maximal use of syntactic creativity** can compress a lot of information into few words, but makes it hard for readers or listeners to follow: surprisal is very high, the continuation of the utterance is hard to predict
- **Shannon's noisy channel** easily breaks down when redundancy is too low.
  - Spoken: misunderstanding, uncertainty
  - Written: longer reading times, backtracking
- Levy & Jaeger (2007): successful communication needs to strike a balance between the two: surprisal should stay constant across the entire text. This is the principle of uniform information density (UID). "UID can be seen as minimizing comprehension difficulty" (Levy & Jaeger 2007: 850).

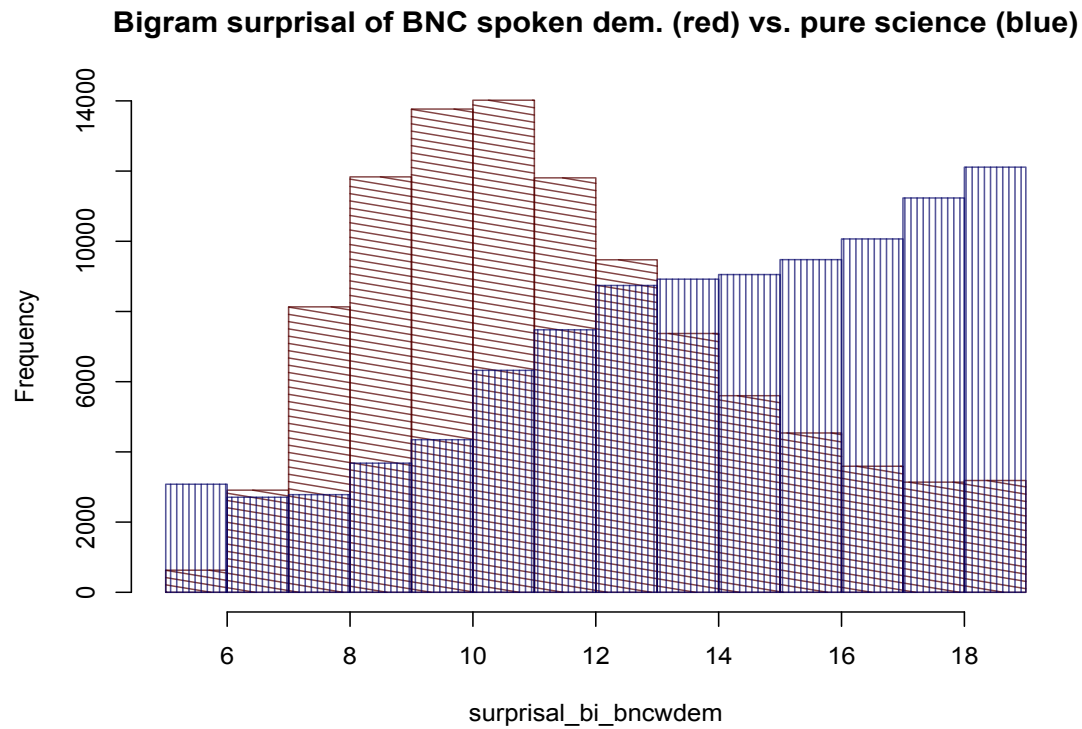
Surprisal: Example (from GECO)

$$Bigram\ surprisal = log\frac{1}{p(w_1)} + \log\frac{1}{p(w_2|\ w_1)}$$

ID	Word	Reading Time (ms)	Surprisal
115	I	512	11.263337
116	was	279	16.195808
117	trying	225	12.843803
118	to	231	11.66212
119	make	184	15.584163
120	up	470	17.04785
121	my	0	15.787277
122	mind	277	19.215173
123	what	214	15.162532
124	to	130	11.381907
125	do	0	17.664406
126	when	468	13.040814
127	I	0	17.910504
128	ran	0	19.851737
129	across	3723	24.424627
130	John	555	26
131	Cavendish	717	26
	.		
132	I	0	12.345475
133	had	676	15.32029
134	seen	250	21.537101
135	very	0	15.455384
136	little	115	17.182405
137	of	236	14.541067
138	him	412	15.457112
139	for	0	14.028664
140	some	423	16.702161
141	years	303	26

Surprisal across Genres

- UID holds well for spoken language, but compressed genres like scientific writing or news (Biber & Conrad 2009) show a skew towards a high level

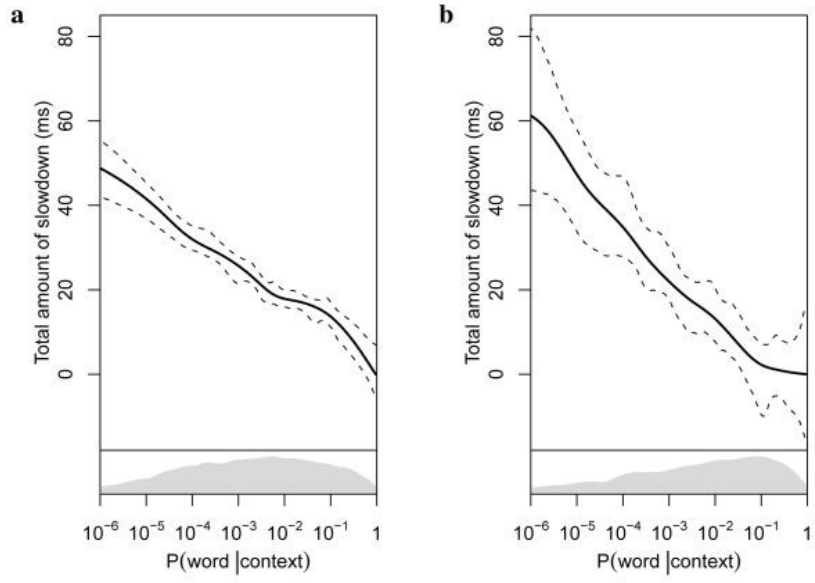


Surprisal: Previous Research

$$Bigram\ surprisal = log\frac{1}{p(w_1)} + \log\frac{1}{p(w_2|\ w_1)}$$

"the forward transitional probability $P(w_k|w_{k-1})$ is a simple form of surprisal" (Demberg and Keller, 2008)

- Eye movement experiments have shown that surprisal correlates to reading times (Demberg and Keller, 2008).
- Correlations to EEG activity has also been investigated (Frank et al. 2013)
- Smith & Levy (2013):
  The effect of word predictability on reading time is **logarithmic**, across 6 orders of magnitude
- But many details are unclear and operationalisations are contested
- Sparseness is a problem
- Some features have not been tested, e.g. collocation, semantics of the discourse



Surprisal & Other features

List of features:

- Word length
- Surprisal
- POS tag
- Punctuation
- Word similarity *sim* (similarity to previous noun, from distributional semantics)
- Surprise in the discourse *mydistance*: log of distance to last previous occurrence of same noun

We use linear regression to predict reading times. As individual flucutations are quit high, we predict means and medians across the readers. Individual differences are unsystematic, except:

- Fast readers have a much better model fit, and higher correlation to surprisal
- L2 readers have much lower model fit, and show difficulties in specific areas

Linear regression: results

ACCURACY OF PREDICTION	√(O-E)^2=	Ø(Z-score)=	relative offness=
For means (medians)	typical error in ms	typical error/sd	typical error/mean
Length (L)	79.93	0.5580	0.4002
Surprisal (S)	95.99	0.6700	0.4806
L+S	79.09	0.5521	0.3960
L+S+tags	78.17	0.5457	0.3914
L+punctuation	78.51	0.5480	0.3931
L+S+punct.	78.24	0.5461	0.3918
L+S+tags+punct.	77.42	0.5404	0.3876
L+S+tags+punct.+sim+mydist	77.22	0.5390	0.3867
Medians ← L+S+tags+punct.+sim+mydist	64.78	0.5345	0.4700

Mean Z-score between individual readers is 1.24

Linear regression: feature weights

```
> gecofitBEST = lm(ppMedians ~ LENGTH + SURPRISAL + PUNCTUATION + pos +
 DEPlen + sim + dep_rel + mydistance, data = eyegecoms)

> drop1(gecofitBEST, test = "F")

Single term deletions

Model:
ppMedians ~ LENGTH + SURPRISAL + PUNCTUATION + pos + DEPlen +
    sim + dep_rel + mydistance
            Df Sum of Sq       RSS    AIC   F value    Pr(>F)
<none>                   93529657 113522
LENGTH       1  27131217 120660874 116765 3678.2326 < 2.2e-16 ***
SURPRISAL    1    174819  93704476 113544   23.7006 1.139e-06 ***
PUNCTUATION  1    988930  94518587 113654  134.0712 < 2.2e-16 ***
pos         12   1024380  94554037 113637   11.5731 < 2.2e-16 ***
DEPlen       1       595  93530252 113520    0.0806    0.7764
sim          1    121718  93651375 113537   16.5016 4.890e-05 ***
dep_rel     41   1405841  94935498 113630    4.6486 < 2.2e-16 ***
mydistance   1    961266  94490923 113651  130.3207 < 2.2e-16 ***
---
```

These results suggest the following approximate ranking of feature weights:

LENGTH > PUNCTUATION >= mydistance > SURPRISAL >= sim >= POS > Dependency Relation

L1 vs L2 feature weights

Language Learners are generally **harder to predict**, and **surprisal is a worse** predictor → Lack of routinization
Goal: Use eye-tracking reading data in language therapy & diagnostics, to predict reading difficulty, and find areas of difficulty for L2 learners

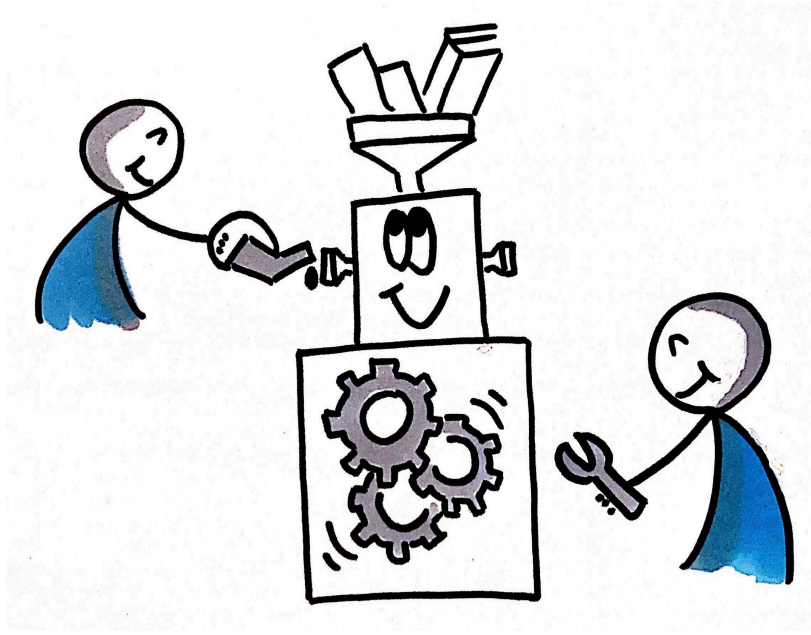
```
> summary(gecofit1s)  ## L1
               Df    Sum Sq   Mean Sq F value Pr(>F)
LENGTH          1 107561422 107561422  6678.1 <2e-16 ***
SURPRISAL       1   3377679   3377679   209.7 <2e-16 ***
PUNCTUATION     1   4674232   4674232   290.2 <2e-16 ***
Residuals   12745 205279403     16107
```

```
> summary(gecofit2s) ## L2
               Df    Sum Sq   Mean Sq F value   Pr(>F)
LENGTH          1 138631778 138631778 3663.56  < 2e-16 ***
SURPRISAL       1   1679436   1679436   44.38 2.81e-11 ***
PUNCTUATION     1   2818953   2818953   74.50  < 2e-16 ***
Residuals   12745 482280309     37841
```

Conclusions

- Our center:
  - Staff
  - Customers
  - Experiences

- Shown several case studies:
  - Topic Modelling (Democracy)
  - Conceptual Maps (Food, Patient Experience)
  - Distributional Semantics (MS)
  - Document classification (Politics)
  - Stylometry & Stylistics (Language & Age)
  - Language Models (Reading Times)



Q & A

Thank you for your attention!

Discussion / Question & Answers

(Reserve slides in the following)