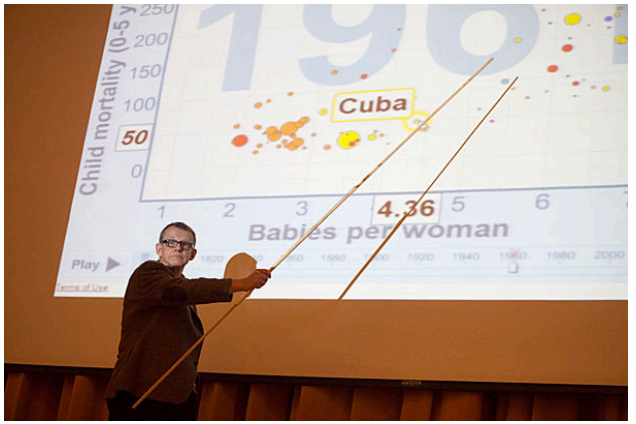
Analyzing linguistic change with motion charts

Martin Hilpert

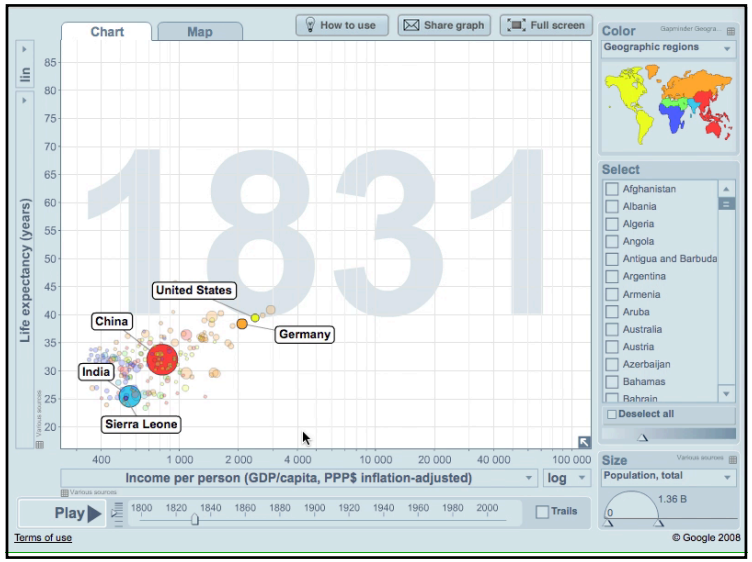
UNIVERSITÉ DE NEUCHÂTEL

1



Hans Rosling

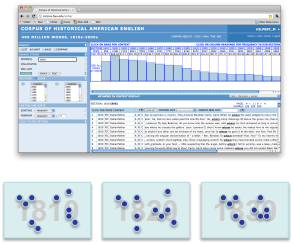
2

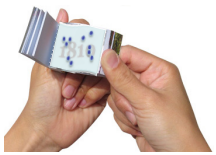


3

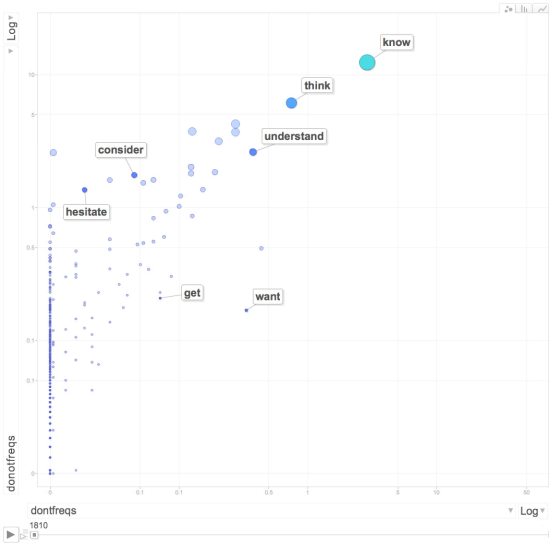
What about linguistic data?

- Use a corpus that represents identical kinds of text across multiple periods of time
- Select a phenomenon and create a visualization for each corpus period.
- View the visualizations in sequence.





4



5

Plan for today

- Two brief case studies
  1. noun-participle compounds (*whisky-soaked*)
  2. many a NOUN (*many a surprise*)
- Making your own motion charts

6

1. Noun-participle compounds

7

Huddleston and Pullum (2002)

- *drug-related*, *home-made*, *safety-tested*, *taxpayer-funded*
- "These compounds generally correspond to syntactic passives with a PP: *related to drugs*, *made at home*, *tested for safety*, *funded by taxpayers*, etc." (2002: 1659)

8

Bauer et al. (2013)

- The first element "cannot receive an object interpretation"
  - doctor-recommended (SUBJECT)
  - arsenic-exposed (PREPOSITIONAL OBJECT)
  - *lunch-eaten (OBJECT)

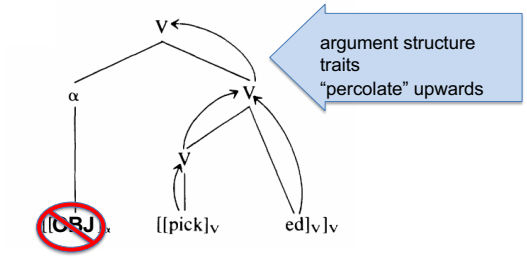
9

Lieber (1983)

- The past participle changes the argument structure of a transitive verb.
- The past participle 'expels' the direct object:
  - passive:
    - The object of a transitive verb cannot be part of the verb phrase.
    - *The strawberries [were picked by hand]$_{VP}$*
  - noun-participle compounding:
    - The object of a transitive verb cannot be part of the compound.
    - the [hand-picked]$_{A}$ strawberries

10

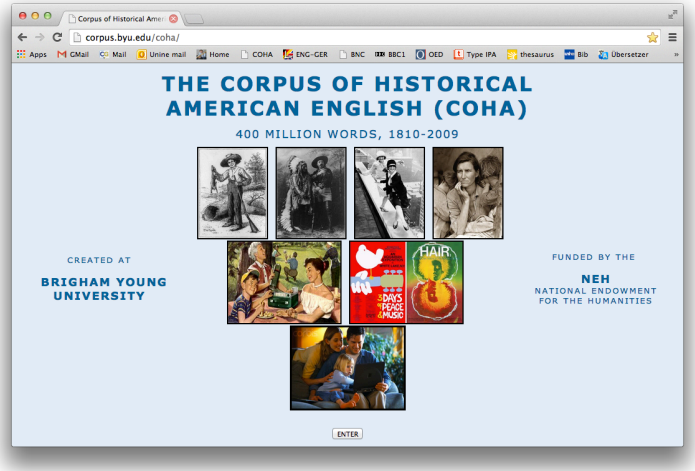
Lieber (1983)



11

Questions:
How does noun-participle compounding develop? Are those developments related to changes in the English passive?

12



13

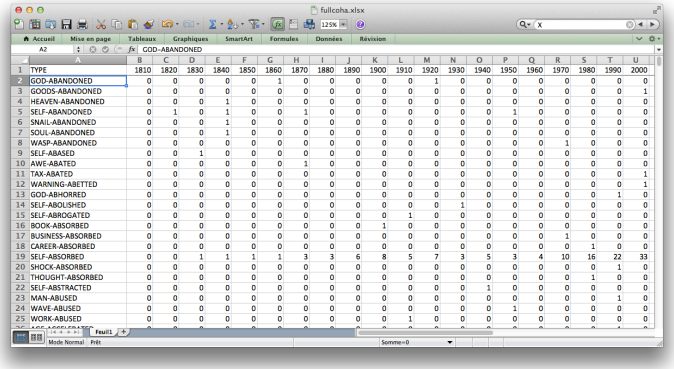
data

- examples from COHA
  - words that were tagged as adjectives
  - words that contained a hyphen
  - words that ended in a letter combination found in past participles
    - government-funded, weather-beaten, sorrow-bent, Florida-born, U.S.-built, etc.

14

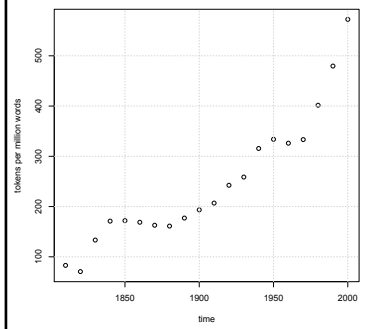
150,000 tokens

31,000 types



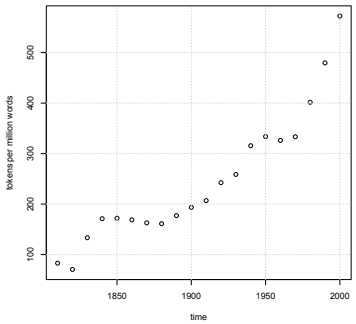
15

increase in token frequency

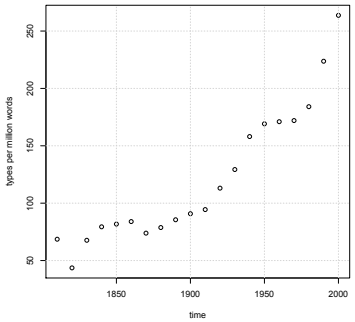


16

increase in token frequency



increase in type frequency

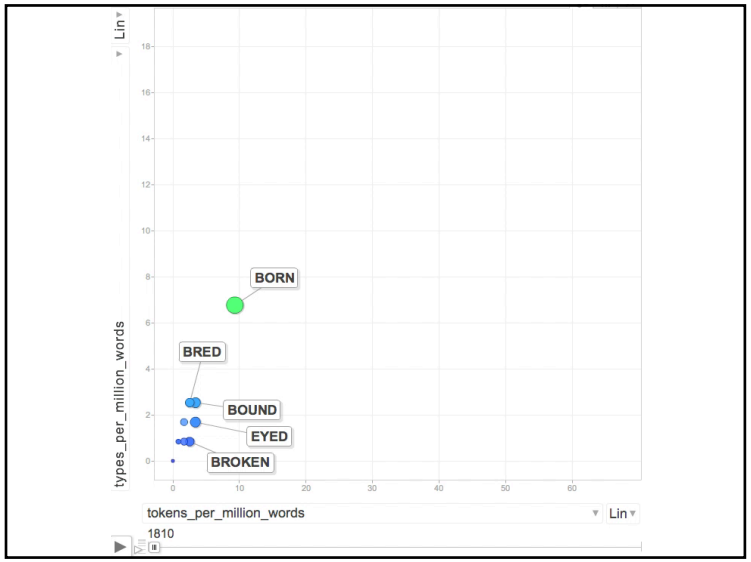


17

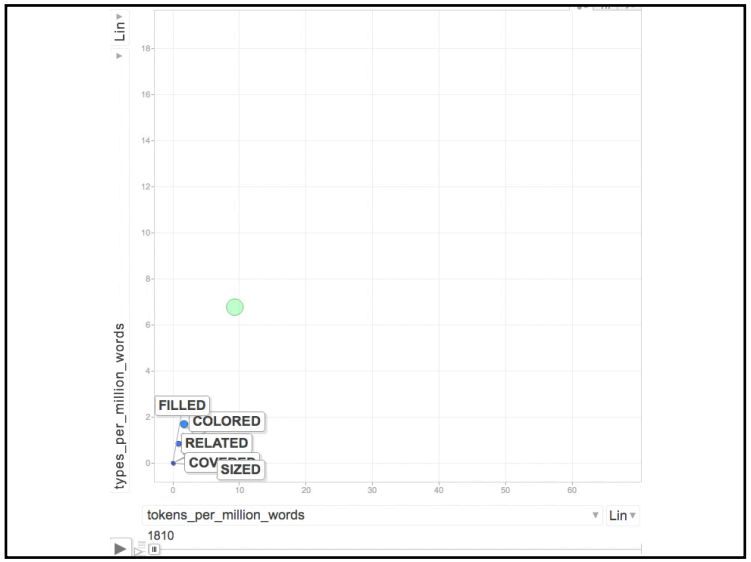
change in the participles

- How does the increase in token and type frequencies come about?
- Which are the participles that carry the increase?
- Participle families of different sizes:
  - wetted: dew, gall, snow, tear
  - yellowed: age, fear, opium, silt, smoke, sun, time, tobacco
  - coated: aluminum, bearskin, beech, blood, candy, caramel, carbon, cement, ... (129 members)

18



19



20

summarizing the developments

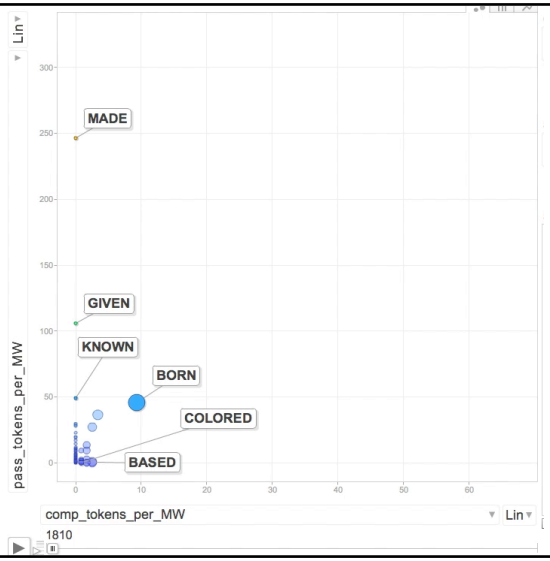
- 1810s
  - Oregon-born, Harvard-bred, context-bound, doe-eyed
- 1900s
  - honey-colored, arrow-shaped, chocolate-covered
- 2000s
  - Houston-based, work-related, toddler-sized

21

How does this compare to the passive?

- The passive with *be* in COHA:
  - [be] [v?n*]
  - 5280 types, approx. 3M tokens
- Overlapping participle types were identified:
  - government-*funded*
  - It was *funded* by the government.
- Do the overlapping types show similar frequency developments?

22



23

interpretation

- The two constructions change independently.
- The participle types that stand out most in the history of noun-participle compounding do not correspond to passive sentences.
  - *The company is based in Houston.*
  - *The problem is related to drug abuse.*
  - *The car is sized just right.*
- >> Both constructions inherit characteristics from the participle, but beyond that, speakers treat them as **two separate constructions**.

24

2. many a NOUN

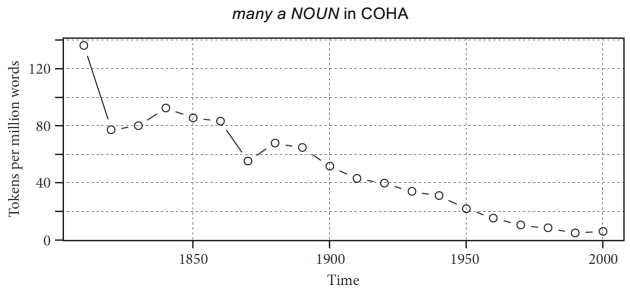
25

many a NOUN

- Time nouns
  - Many a day will pass until this construction is properly understood.
  - I've thought that many a time myself.
- Human beings
  - College education was at a premium, with many a father resisting education for a daughter.
  - Many a labour voter is not happy with the outcome.
- Totally random nouns
  - During my time in Australia I enjoyed many a sausage roll for brekkie.

26

frequencies fall, but productivity stays high



27

Question:
What happens to the semantic spectrum of *many a NOUN*? Why can speakers still say things like *many a sausage roll* ?

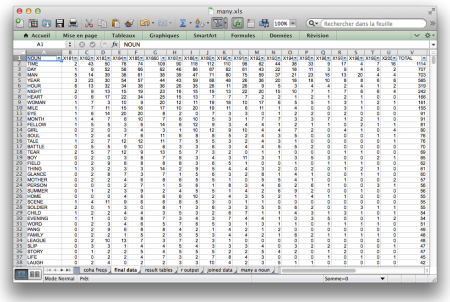
28

data

- examples from COHA
  - all sequences of *many*, *a* / *an*, and a noun
  - focus on the 230 most frequent types (63% of the data)

15,000 tokens

3,140 types



29

analysis

- For the 230 most frequent types, a **semantic vector space** was constructed on the basis of synchronic corpus data.
- What are the relative similarities in a group of words?

30

analysis

- For the 230 most frequent types, a **semantic vector space** was constructed on the basis of synchronic corpus data.
- What are the relative similarities in a group of words?
- Words that occur with the same collocates are judged to be similar.
- Given a word such as CHURCH, what are the lexical words that co-occur with it in a window of 4L and 4R?

31

BNC collocation frequencies (four-left, four-right window)

terms

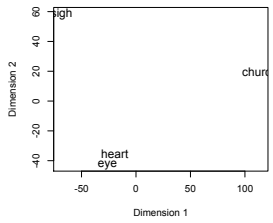
collocates

	church	heart	eye	sigh
abbey	30	0	1	0
above	13	8	26	0
activities	21	2	0	0
always	28	28	23	5
half	8	0	10	0
long	28	11	7	48
family	47	10	4	1
gave	9	13	12	109
christ	141	8	1	0
walk	15	2	1	0

32

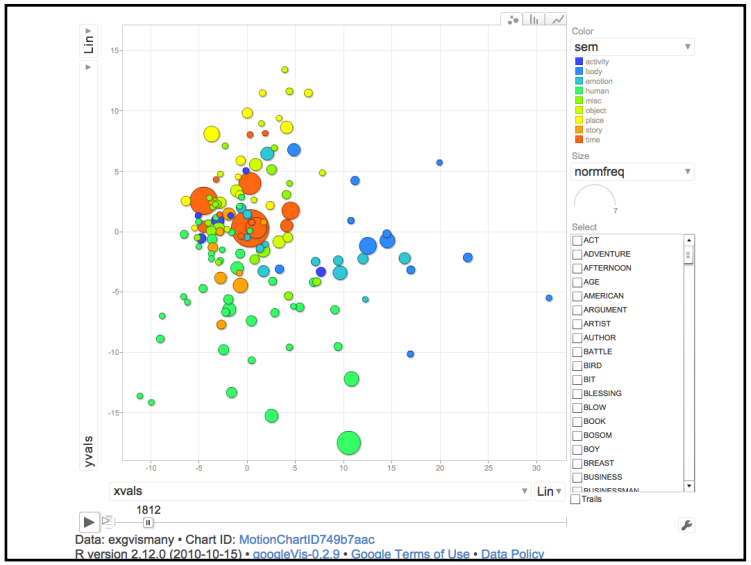
from collocate frequencies to a map

	church	heart	eye	sigh
abbey	30	0	1	0
above	13	8	26	0
activities	21	2	0	0
always	28	28	23	5
half	8	0	10	0
long	28	11	7	48
family	47	10	4	1
gave	9	13	12	109
christ	141	8	1	0
walk	15	2	1	0

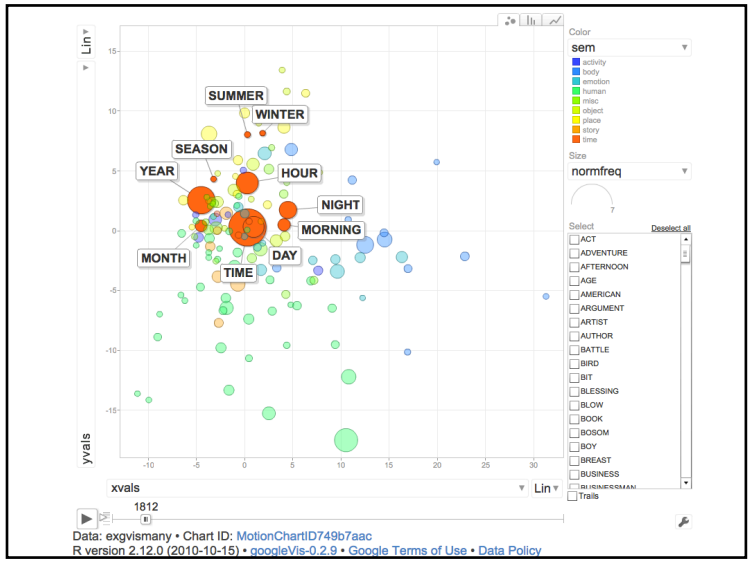



```
          church      heart        eye
heart 144.50606
eye   153.54153   23.60085
sigh  186.34645 106.44717 110.46266
```

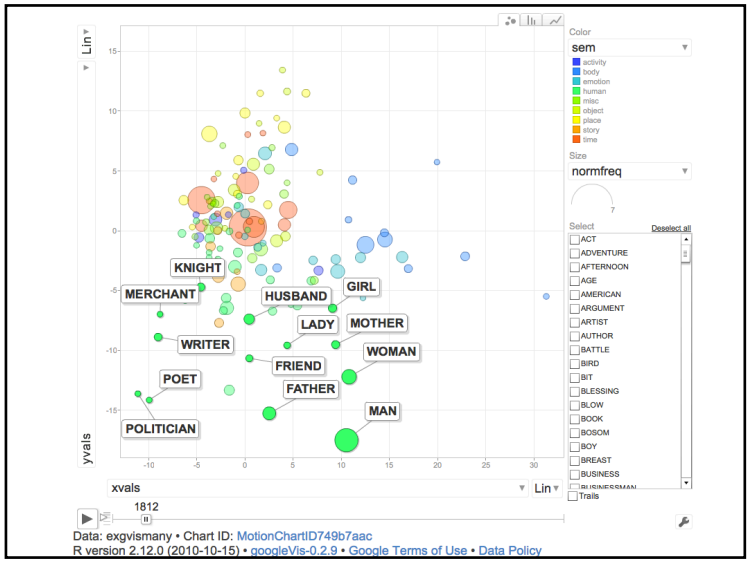
33



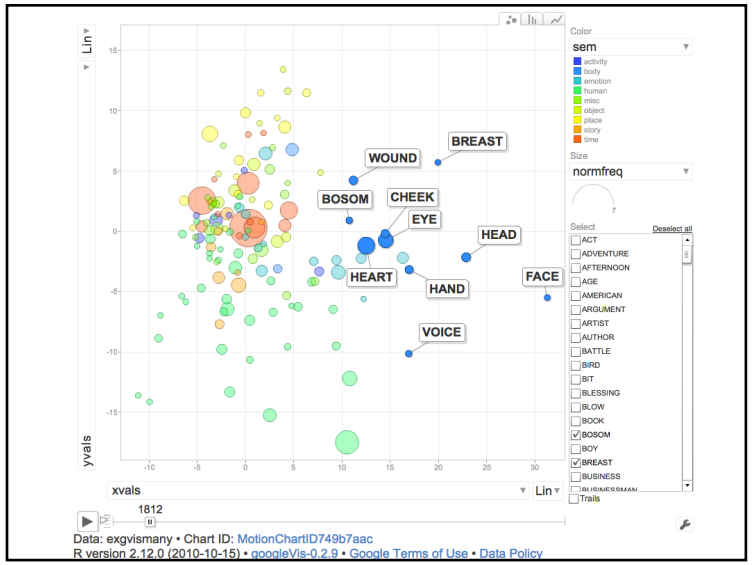
34



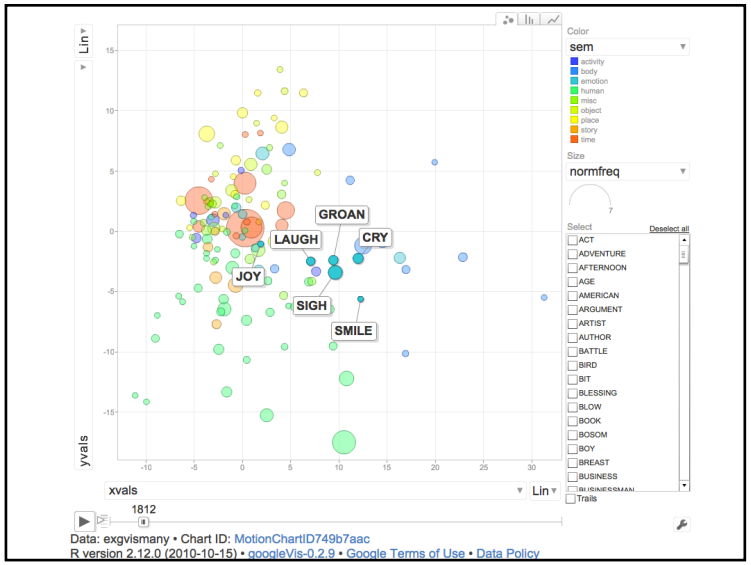
35



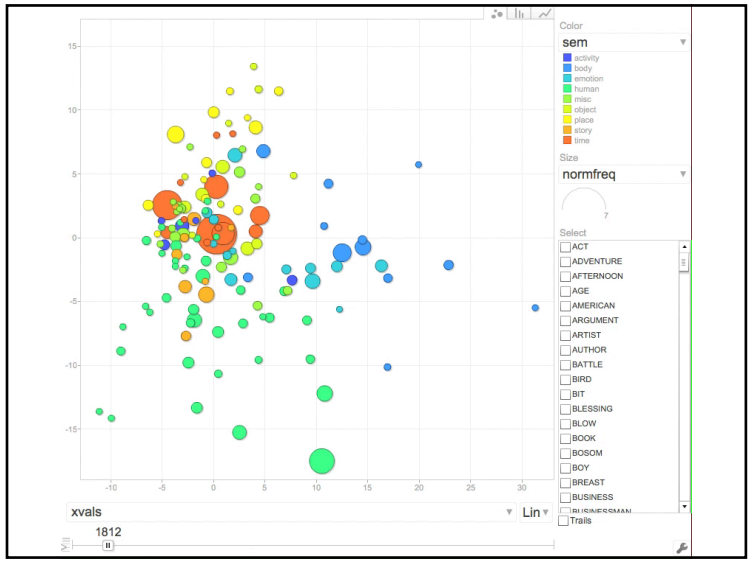
36


WOUND
BREAST
BOSOM
CHEEK
EYE
HEAD
HEART
HAND
FACE
VOICE
Color
sem
Size
normfreq
Select
Deselect all
xvals
yvals
Lin
1812
Data: exgvismany • Chart ID: MotionChartID749b7aac
R version 2.12.0 (2010-10-15) • googleVis-0.2.9 • Google Terms of Use • Data Policy

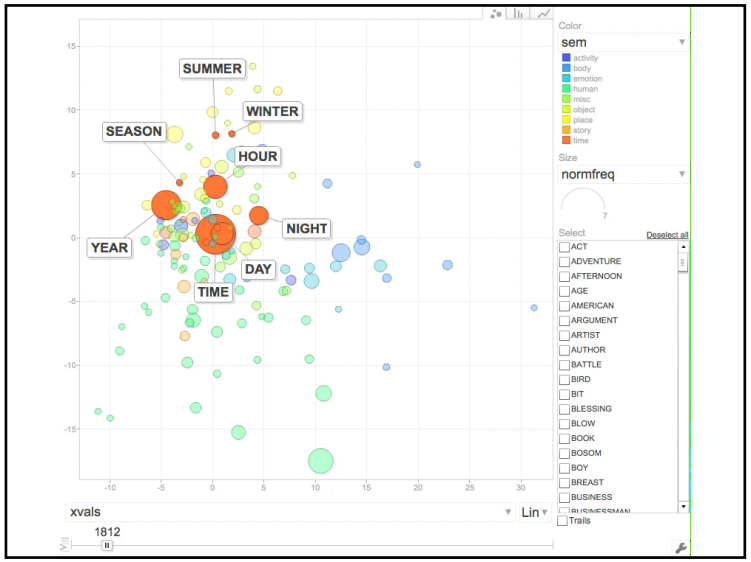
37


GROAN
LAUGH
CRY
JOY
SIGH
SMILE
Color
sem
Size
normfreq
Select
Deselect all
xvals
yvals
Lin
1812
Data: exgvismany • Chart ID: MotionChartID749b7aac
R version 2.12.0 (2010-10-15) • googleVis-0.2.9 • Google Terms of Use • Data Policy

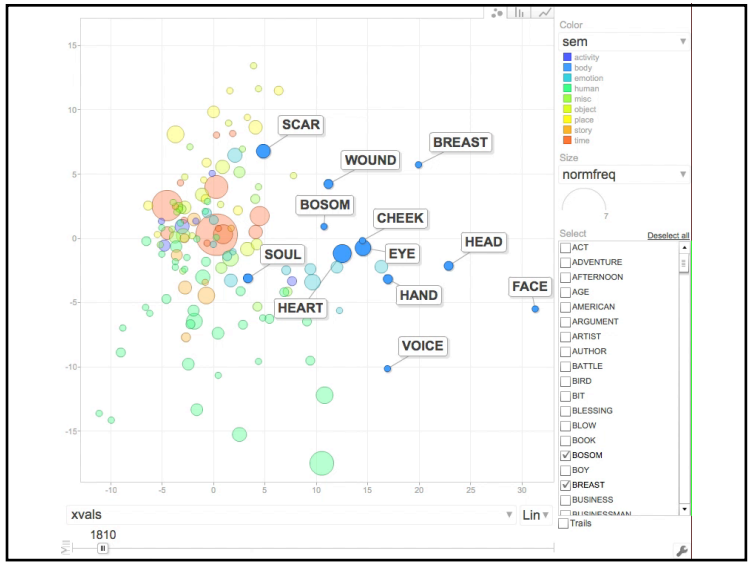
38


Color
sem
Size
normfreq
Select
xvals
Lin
1812

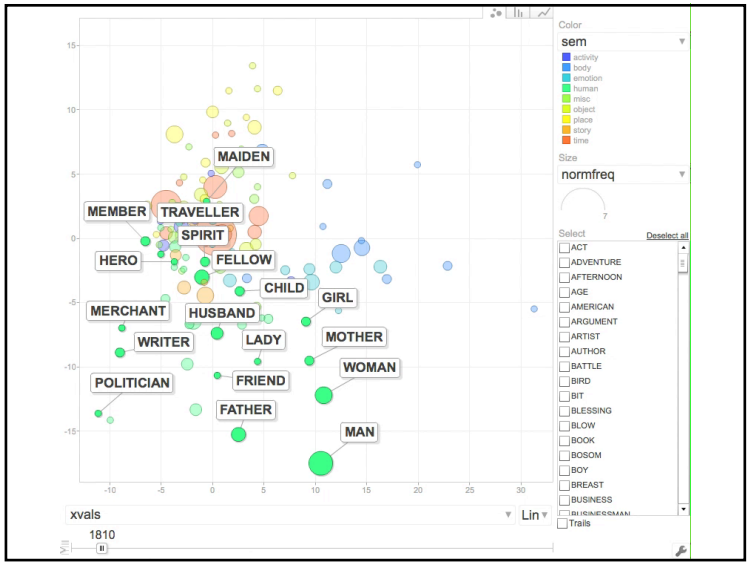
39


SUMMER
WINTER
SEASON
HOUR
NIGHT
YEAR
DAY
TIME
Color
sem
Size
normfreq
Select
xvals
Lin
1812

40



41



42

interpretation

- Whence *many a sausage roll*?
- The *many a noun* construction does not recede into a single semantic niche.
- Time nouns remain strong, human being nouns remain strong.
- Words like *time* or *man* are highly diffuse in their collocational behavior, besides them there is a sizable residue of semantically diverse types.
- Speakers thus experience the *many a NOUN* construction as **semantically unrestricted**.

43

Making your own motion charts

44

what you need

- a corpus with comparable data from different historical periods
  - Mark Davies' corpora

- the open-source software R
  - the package "googleVis"
- a spreadsheet software
- internet access

45

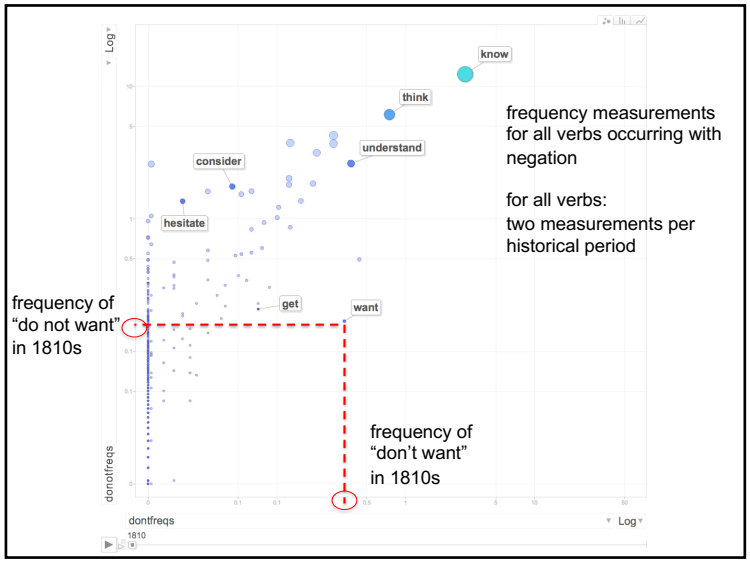
data:
what kind of data, how to organize it

46

What kind of corpus data?

- easiest scenario:
  - a comparison of two alternative linguistic forms
  - negation:
    - contracted form plus verb: *don't VERB*
    - non-contracted form plus verb: *do not VERB*
  - what is compared is the frequency of elements in the VERB slot
    - Are certain verbs more likely to occur with the contracted form? How does this change over time?
    - contracted frequencies on the x-axis
    - non-contracted frequencies on the y-axis

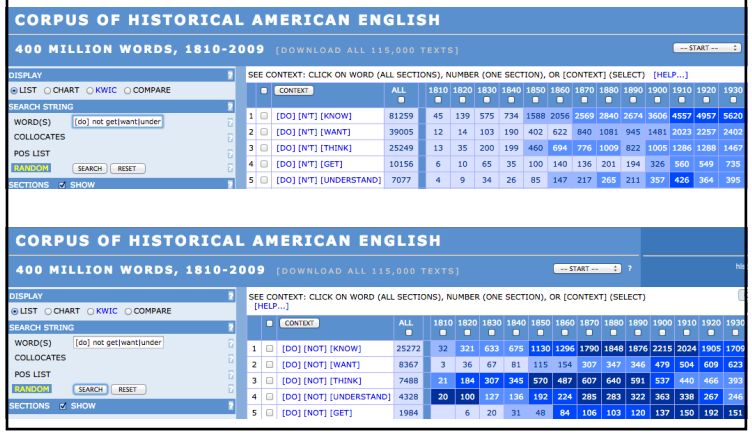
47



48

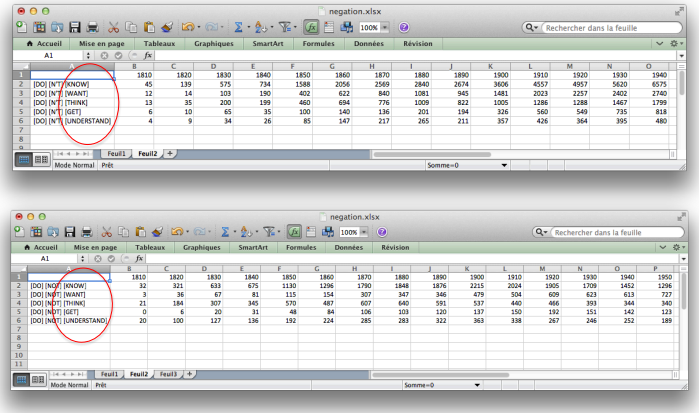
retrieving data

do not / don't know, want, think, get, understand



49

paste data into a spreadsheet, bring verbs into identical sequence



50

getting the data into the right format

(requires some manual copy & paste)

verb	time period	contracted	non-contracted
know	1810	45	32
want	1810	12	3
think	1810	13	21
get	1810	...	...
understand	1810	...	...
know	1820	139	321
want	1820	...	...
think	...	...	...
...	...	...	...

51

problem: If the corpus periods have different sizes, the frequencies need to be normalized!

verb	time period	contracted	non-contracted
know	1810	45	32
want	1810	12	3
think	1810	13	21
get	1810	...	...
understand	1810	...	...
know	1820	139	321
want	1820	...	...
think	...	...	...
...	...	...	...

52

period sizes of the COHA

COMPOSITION OF THE CORPUS

Download all 115,000 texts, for use on your own computer.

DECADE	FICTION	POPULAR MAGAZINES	NEWSPAPERS	NON-FICTION BOOKS	TOTAL	% FICTION
1810s	641,164	88,316	0	451,542	1,181,022	0.54
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005	0.54
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987	0.55
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854	0.55
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826	0.55
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102	0.55
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160	0.55
1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855	0.54
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383	0.53
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232	0.53
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252	0.53

53

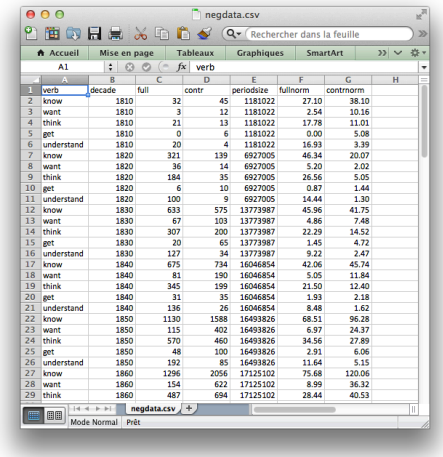
45 examples in a 1.18 million word corpus:
38.1 instances per million words

verb	time period	contracted	non-contracted
know	1810	45	32
want	1810	12	3
think	1810	13	21
get	1810	…	…
understand	1810	…	…
know	1820	139	321
want	1820	…	…
think	…	…	…
…	…	…	…

139 examples in a 6.92 million word corpus:
20.6 instances per million words

54

suggested format, saved as .csv

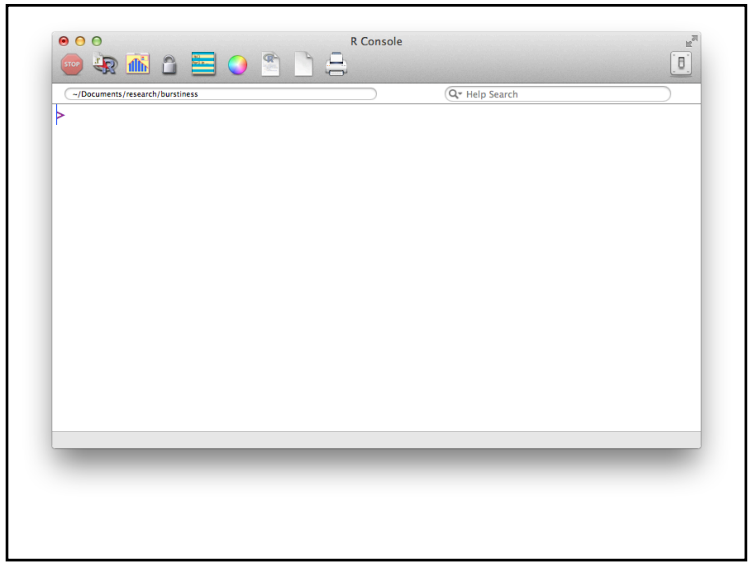


	A	B	C	D	E	F	G
1	verb	decade	full	contr	periodsize	fullnorm	contnorm
2	know	1810	32	45	1181022	27.10	38.10
3	want	1810	3	12	1181022	2.54	10.16
4	think	1810	21	13	1181022	17.78	11.01
5	get	1810	0	6	1181022	0.00	5.08
6	understand	1810	20	4	1181022	16.93	3.39
7	know	1820	321	139	6927005	46.34	20.07
8	want	1820	36	14	6927005	5.20	2.02
9	think	1820	184	35	6927005	26.56	5.05
10	get	1820	6	10	6927005	0.87	1.44
11	understand	1820	100	9	6927005	14.44	1.30
12	know	1830	633	575	13773987	45.96	41.75
13	want	1830	67	103	13773987	4.86	7.48
14	think	1830	307	200	13773987	22.29	14.52
15	get	1830	20	65	13773987	1.45	4.72
16	understand	1830	127	34	13773987	9.22	2.47
17	know	1840	675	734	16046854	42.06	45.74
18	want	1840	81	190	16046854	5.05	11.84
19	think	1840	345	199	16046854	21.50	12.40
20	get	1840	31	35	16046854	1.93	2.18
21	understand	1840	136	26	16046854	8.48	1.62
22	know	1850	1130	1588	16493826	68.51	96.28
23	want	1850	115	402	16493826	6.97	24.37
24	think	1850	570	460	16493826	34.56	27.89
25	get	1850	48	100	16493826	2.91	6.06
26	understand	1850	192	85	16493826	11.64	5.15
27	know	1860	1296	2056	17125102	75.68	120.06
28	want	1860	154	622	17125102	8.99	36.32
29	think	1860	487	694	17125102	28.44	40.53

55

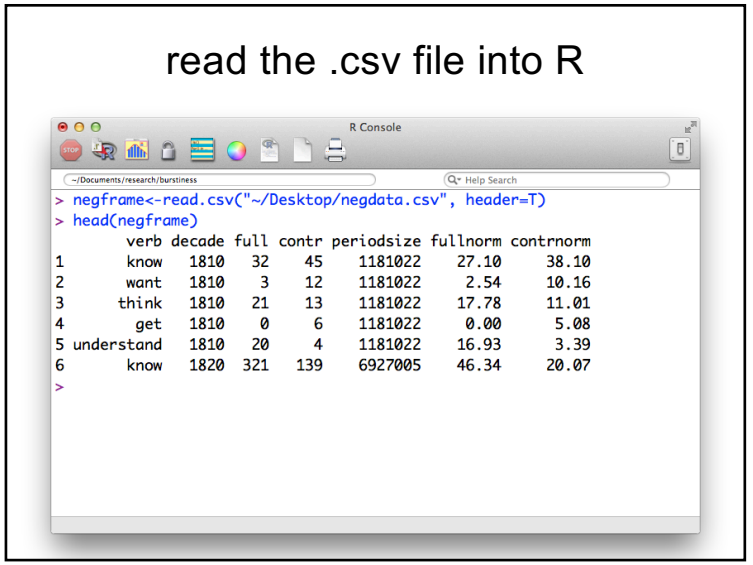
creating the motion chart:
install and open R (www.r-project.org)

56

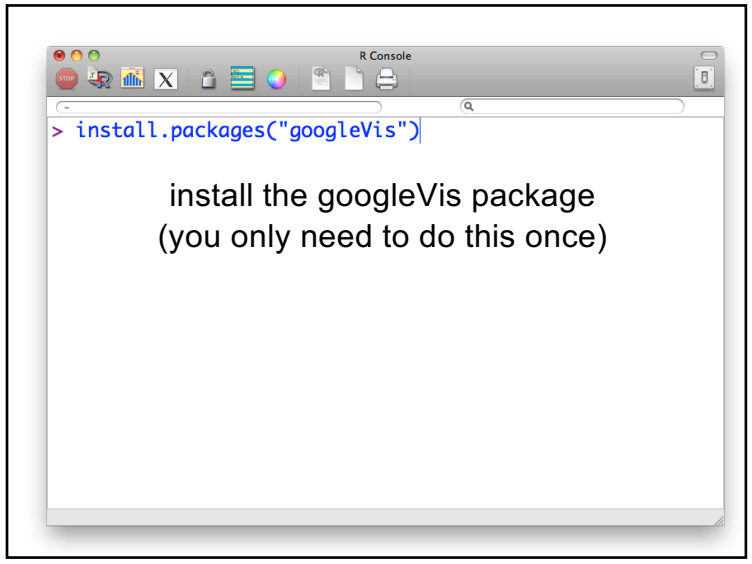


57

read the .csv file into R

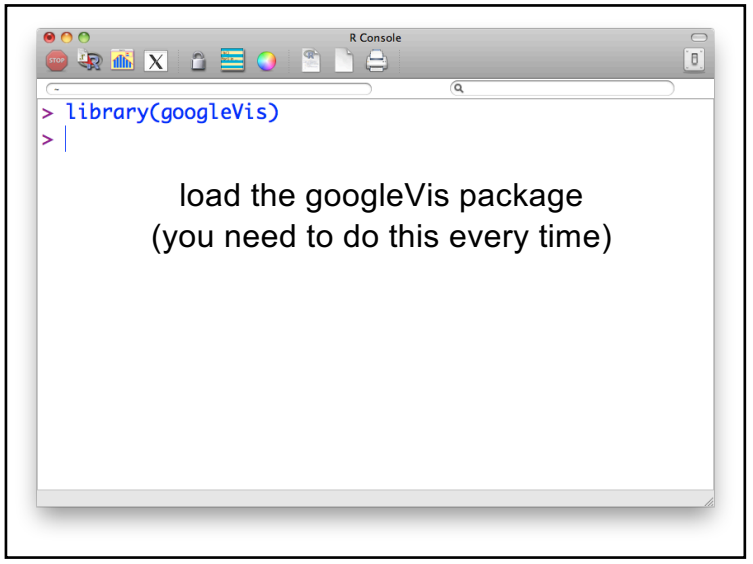


58



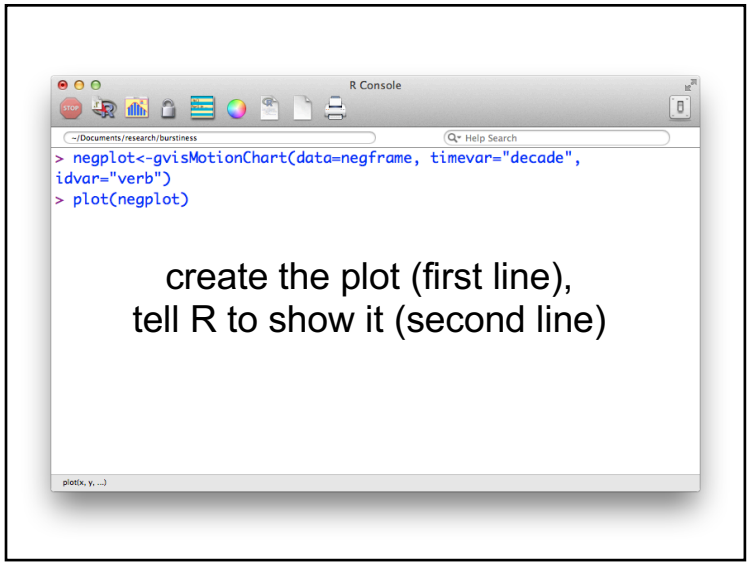
install the googleVis package
(you only need to do this once)

59

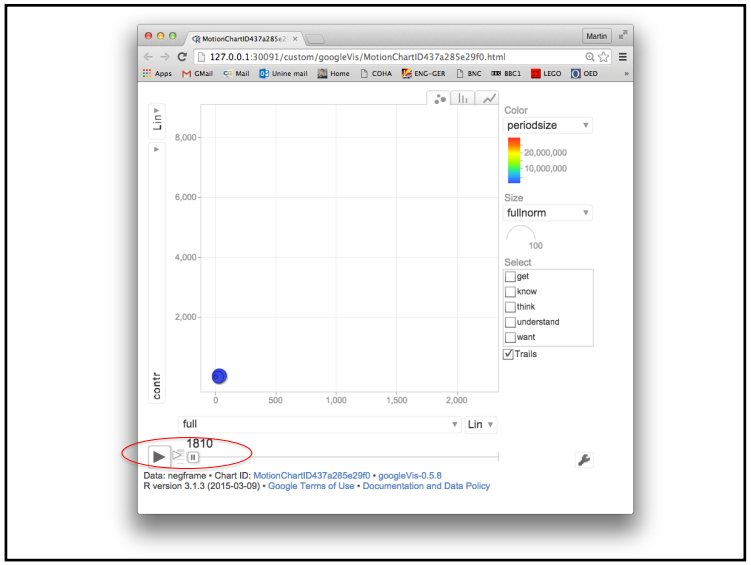


load the googleVis package
(you need to do this every time)

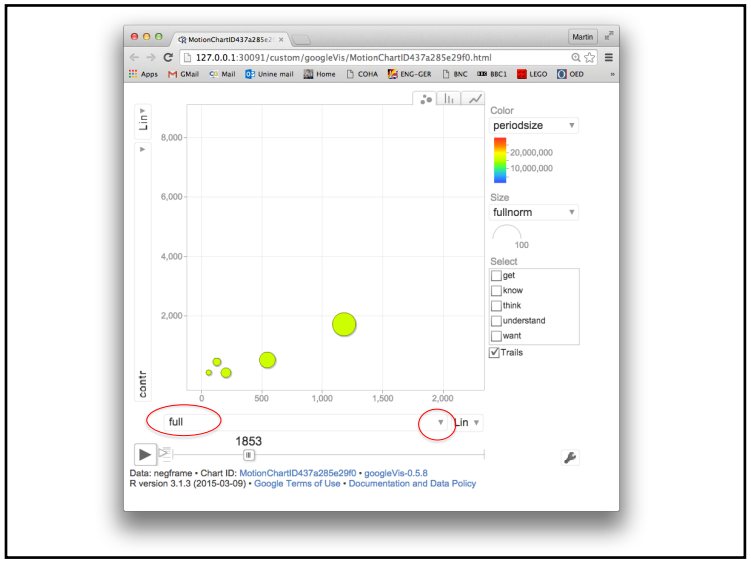
60


R Console
~/Documents/research/burstiness
Help Search
> negplot<-gvisMotionChart(data=negframe, timevar="decade",
idvar="verb")
> plot(negplot)
create the plot (first line),
tell R to show it (second line)

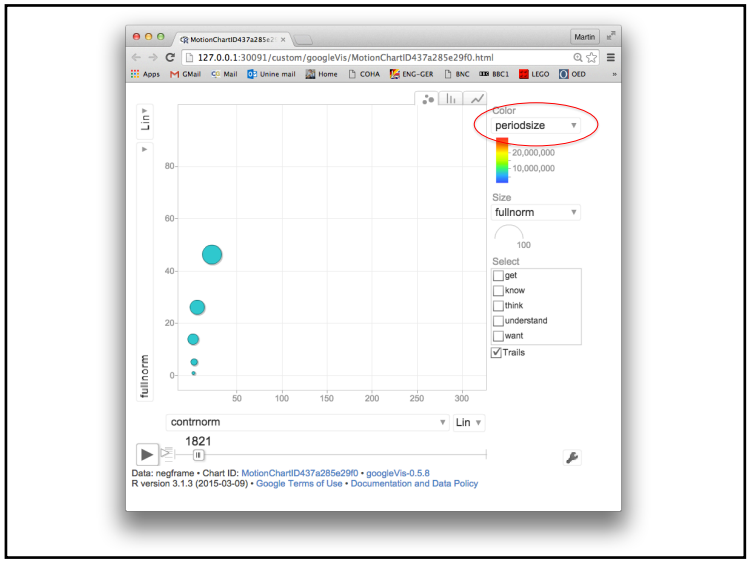
61


127.0.0.1:30091/custom/googleVis/MotionChartID437a285e29f0.html
Color
periodsize
Size
fullnorm
Select
get
know
think
understand
want
Trails
contr
full
Lin
1810

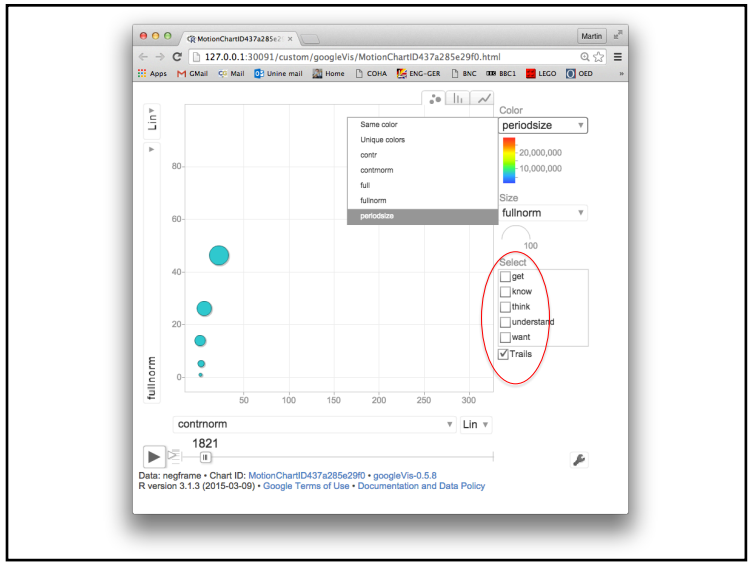
62


127.0.0.1:30091/custom/googleVis/MotionChartID437a285e29f0.html
Color
periodsize
Size
fullnorm
Select
get
know
think
understand
want
Trails
contr
full
Lin
1853

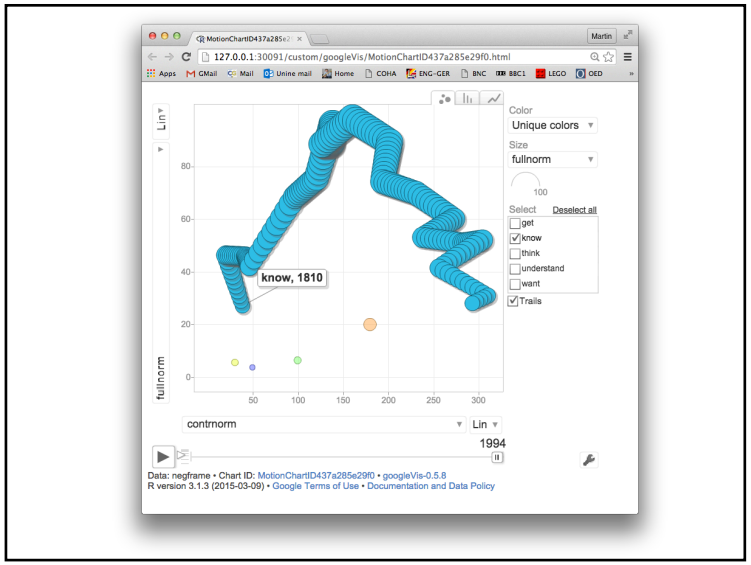
63


127.0.0.1:30091/custom/googleVis/MotionChartID437a285e29f0.html
Color
periodsize
Size
fullnorm
Select
get
know
think
understand
want
Trails
fullnorm
contrnorm
Lin
1821

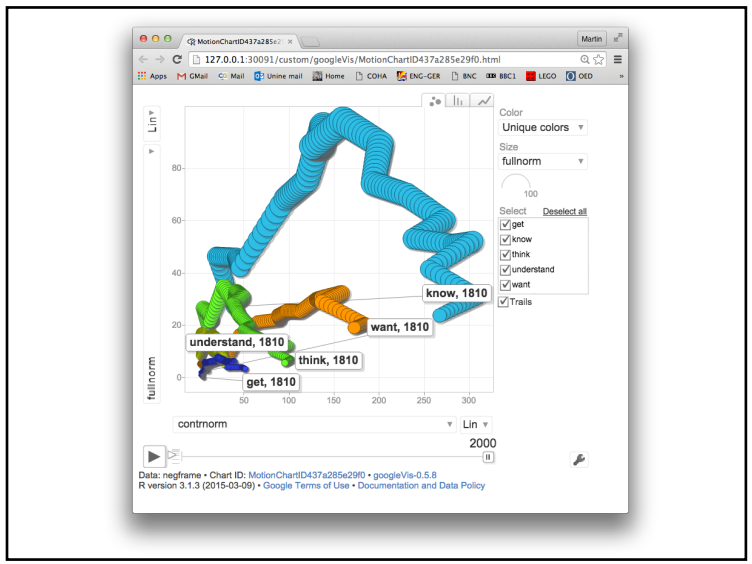
64


Color
periodsize
Same color
Unique colors
contr
contrnorm
full
fullnorm
periodsize
20,000,000
10,000,000
Size
fullnorm
Select
get
know
think
understand
want
Trails
fullnorm
contrnorm
Lin
1821

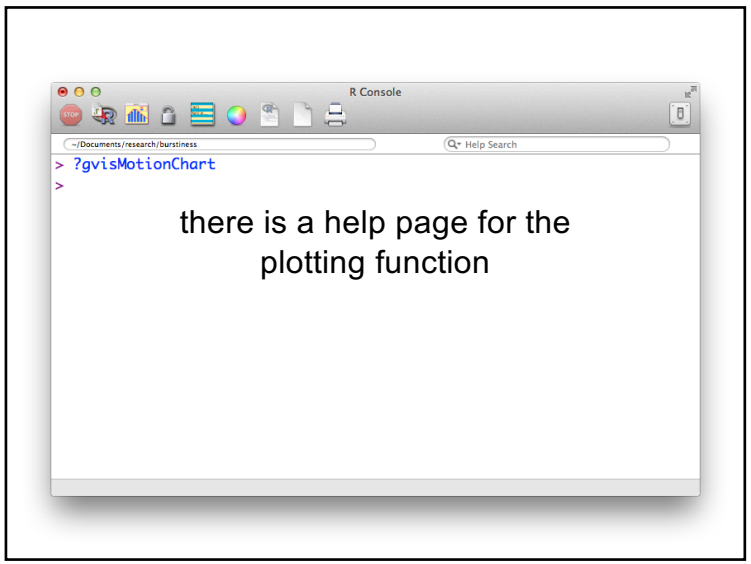
65


Color
Unique colors
Size
fullnorm
Select
Deselect all
get
know
think
understand
want
Trails
know, 1810
fullnorm
contrnorm
Lin
1994

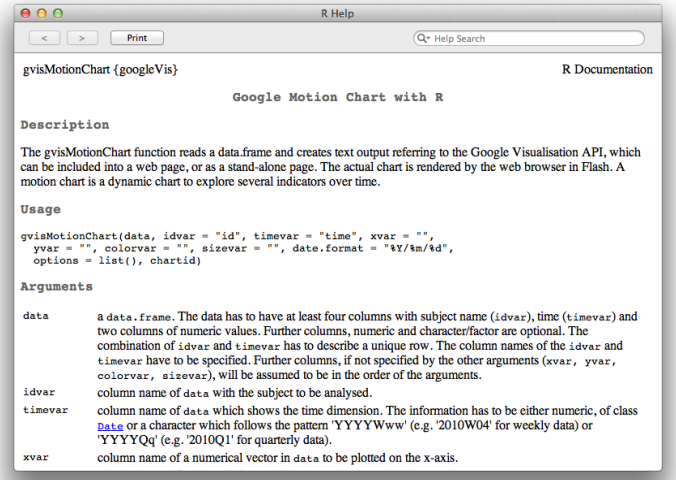
66


Color
Unique colors
Size
fullnorm
Select
Deselect all
get
know
think
understand
want
Trails
know, 1810
want, 1810
understand, 1810
think, 1810
get, 1810
fullnorm
contrnorm
Lin
2000

67


R Console
~/Documents/research/burstiness
Help Search
> ?gvisMotionChart
>
there is a help page for the
plotting function

68


R Help

Print

Help Search

gvisMotionChart {googleVis} R Documentation

Google Motion Chart with R

Description

The gvisMotionChart function reads a data.frame and creates text output referring to the Google Visualisation API, which can be included into a web page, or as a stand-alone page. The actual chart is rendered by the web browser in Flash. A motion chart is a dynamic chart to explore several indicators over time.

Usage

```
gvisMotionChart(data, idvar = "id", timevar = "time", xvar = "",
  yvar = "", colorvar = "", sizevar = "", date.format = "%Y/%m/%d",
  options = list(), chartid)
```

Arguments

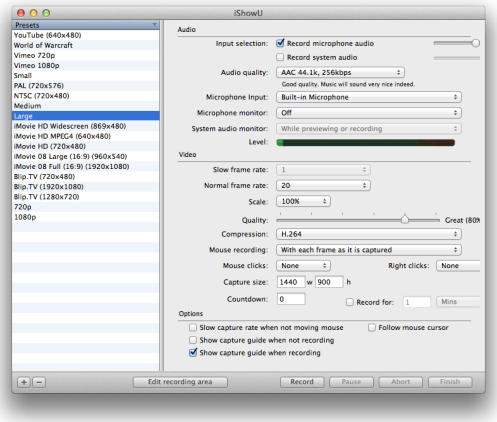
data	a data.frame. The data has to have at least four columns with subject name (idvar), time (timevar) and two columns of numeric values. Further columns, numeric and character/factor are optional. The combination of idvar and timevar has to describe a unique row. The column names of the idvar and timevar have to be specified. Further columns, if not specified by the other arguments (xvar, yvar, colorvar, sizevar), will be assumed to be in the order of the arguments.
idvar	column name of data with the subject to be analysed.
timevar	column name of data which shows the time dimension. The information has to be either numeric, of class Date or a character which follows the pattern 'YYYYWww' (e.g. '2010W04' for weekly data) or 'YYYYQq' (e.g. '2010Q1' for quarterly data).
xvar	column name of a numerical vector in data to be plotted on the x-axis.

69

How can I present a motion chart when I give a talk?

70

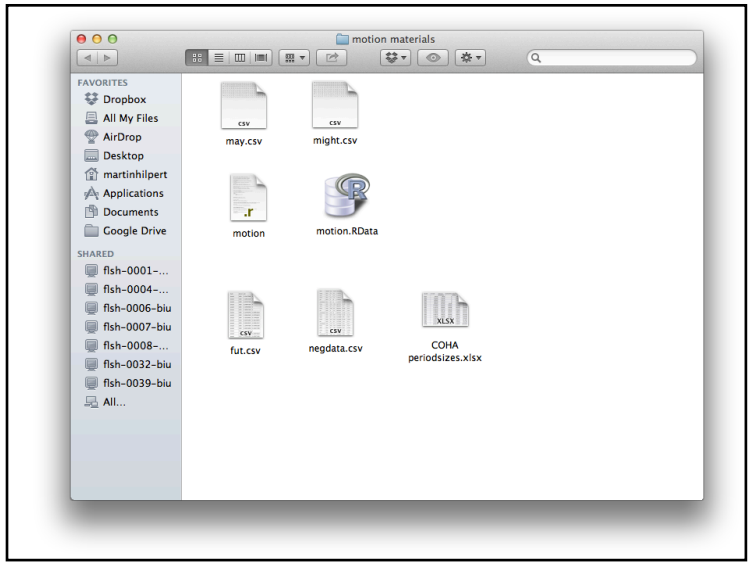
screen capture software, creates .mov files



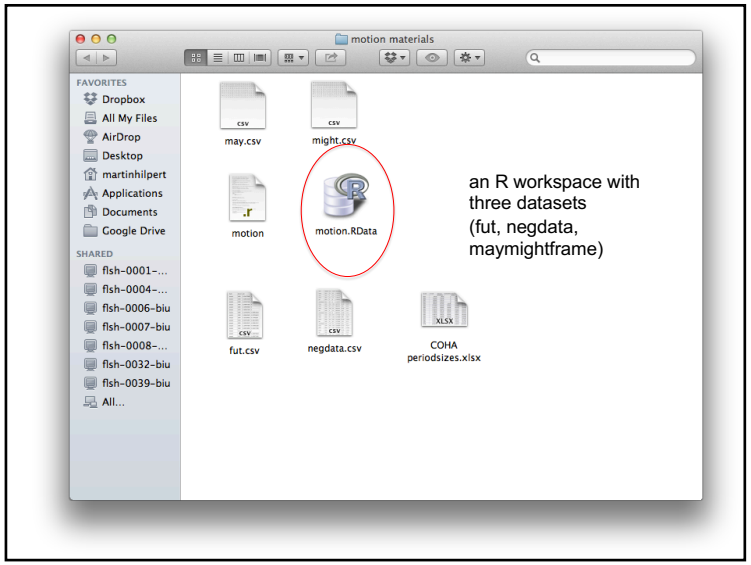
71

Some exercise materials:
http://members.unine.ch/martin.hilpert/motion.zip

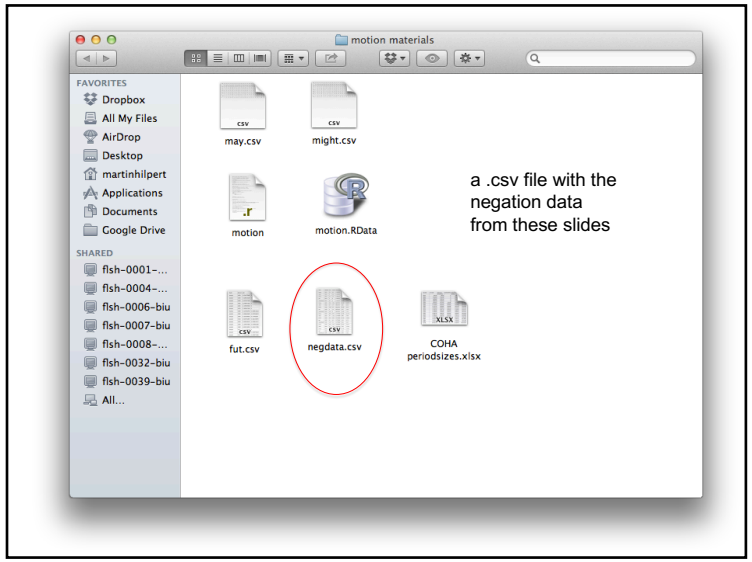
72


motion materials
FAVORITES
Dropbox
All My Files
AirDrop
Desktop
martinhilpert
Applications
Documents
Google Drive
SHARED
fish-0001-...
fish-0004-...
fish-0006-biu
fish-0007-biu
fish-0008-...
fish-0032-biu
fish-0039-biu
All...
may.csv
might.csv
motion
motion.RData
fut.csv
negdata.csv
COHA periodsizes.xlsx

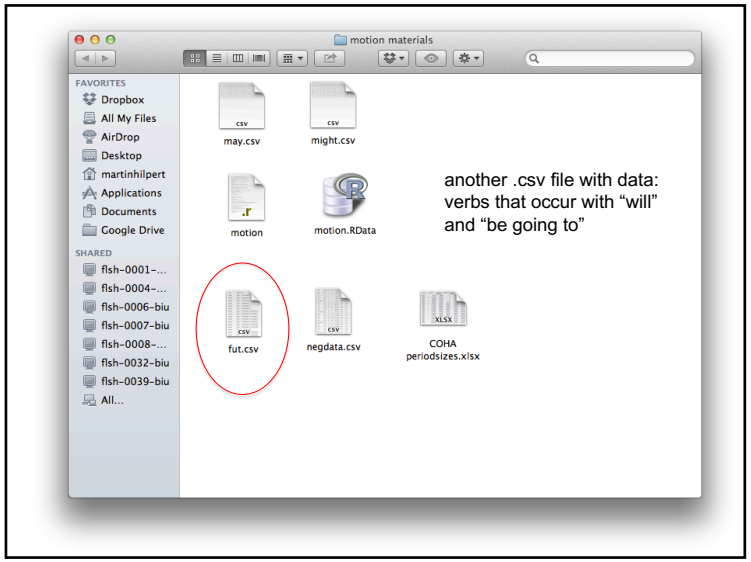
73


motion materials
FAVORITES
Dropbox
All My Files
AirDrop
Desktop
martinhilpert
Applications
Documents
Google Drive
SHARED
fish-0001-...
fish-0004-...
fish-0006-biu
fish-0007-biu
fish-0008-...
fish-0032-biu
fish-0039-biu
All...
may.csv
might.csv
motion
motion.RData
fut.csv
negdata.csv
COHA periodsizes.xlsx
an R workspace with three datasets (fut, negdata, maymightframe)

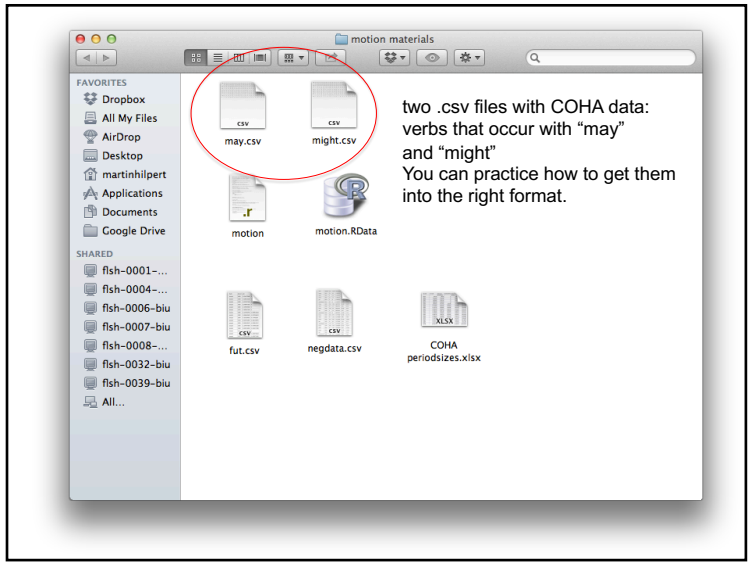
74


motion materials
FAVORITES
Dropbox
All My Files
AirDrop
Desktop
martinhilpert
Applications
Documents
Google Drive
SHARED
fish-0001-...
fish-0004-...
fish-0006-biu
fish-0007-biu
fish-0008-...
fish-0032-biu
fish-0039-biu
All...
may.csv
might.csv
motion
motion.RData
fut.csv
negdata.csv
COHA periodsizes.xlsx
a .csv file with the negation data from these slides

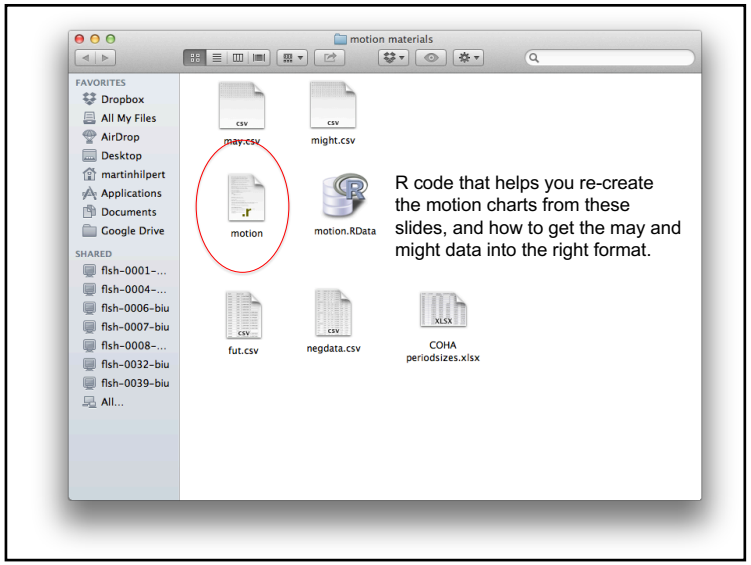
75


motion materials
FAVORITES
Dropbox
All My Files
AirDrop
Desktop
martinhilpert
Applications
Documents
Google Drive
SHARED
fish-0001-...
fish-0004-...
fish-0006-biu
fish-0007-biu
fish-0008-...
fish-0032-biu
fish-0039-biu
All...
may.csv
might.csv
motion
motion.RData
fut.csv
negdata.csv
COHA periodsizes.xlsx
another .csv file with data: verbs that occur with “will” and “be going to”

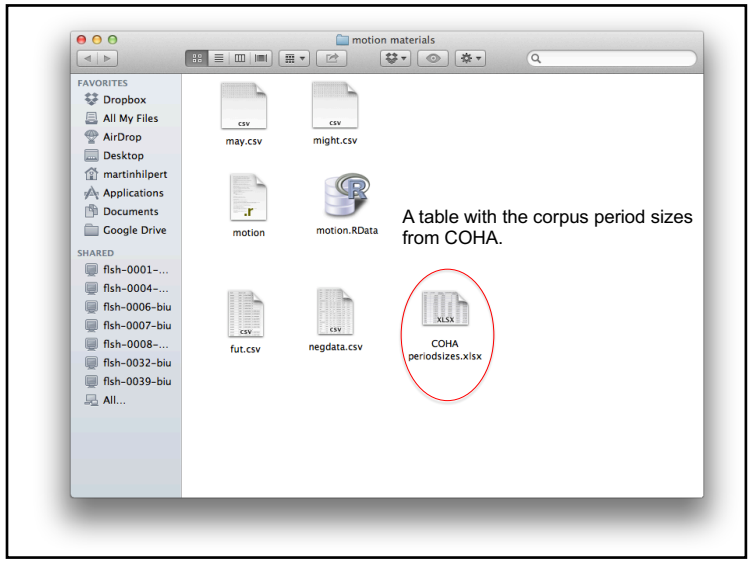
76



77



78



79

Many thanks!

martin.hilpert@unine.ch
http://members.unine.ch/martin.hilpert

80