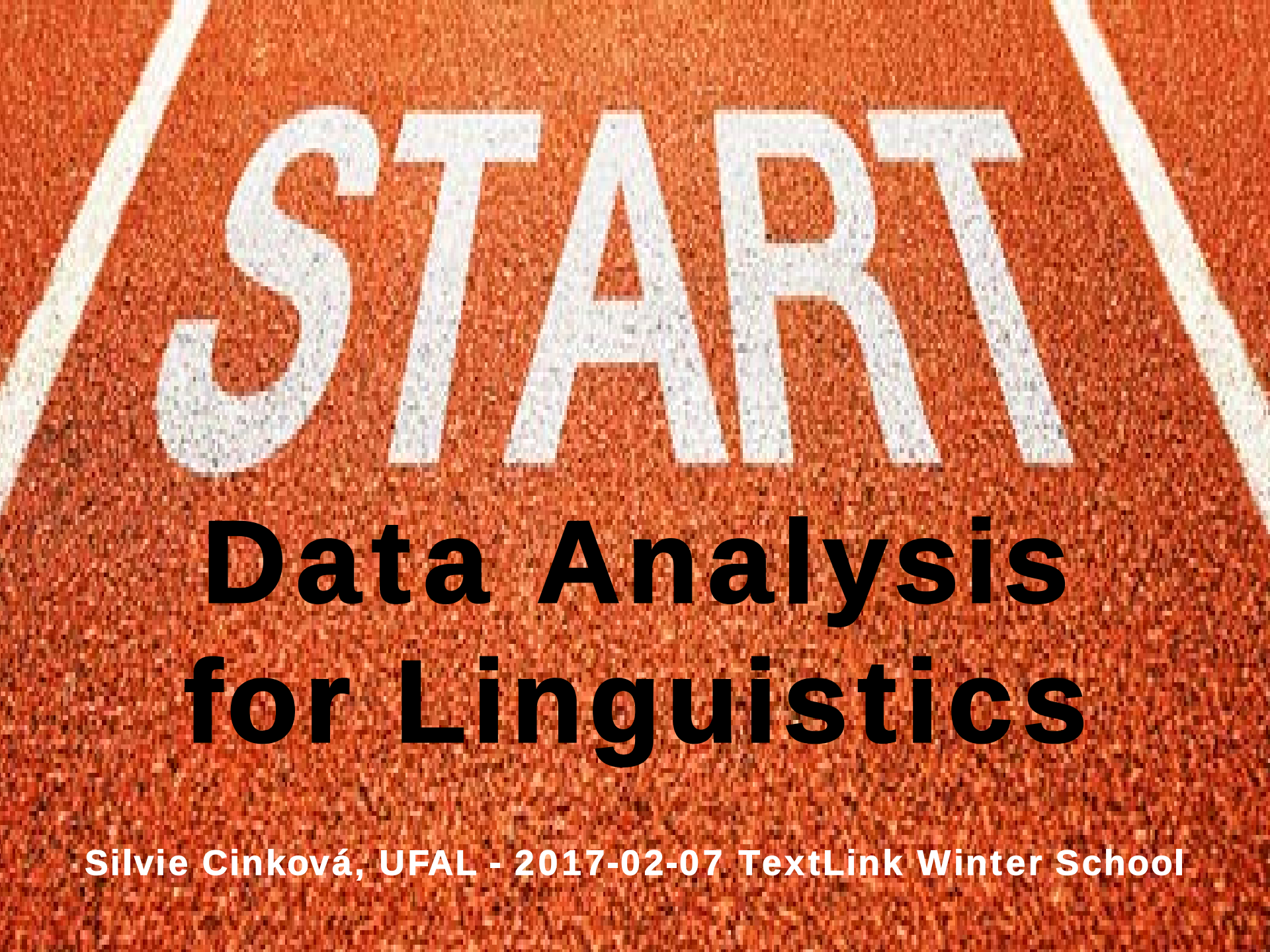




START

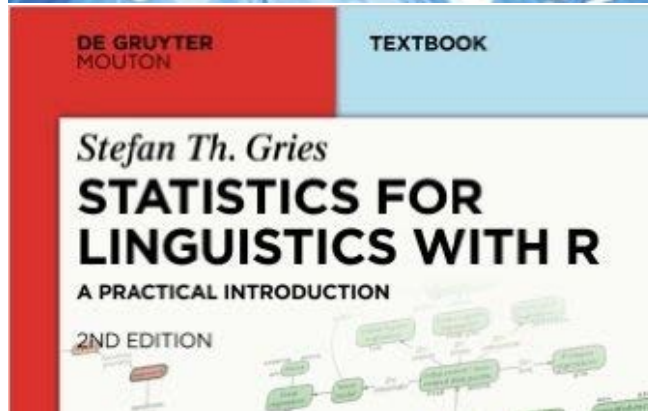
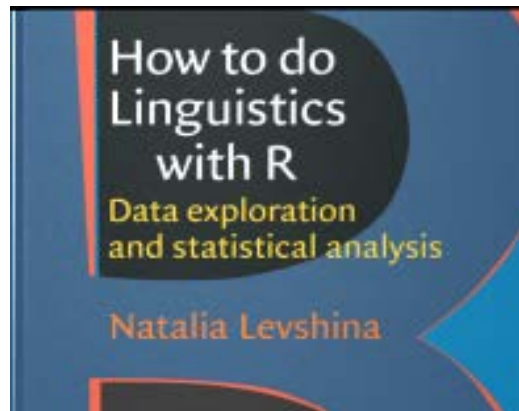
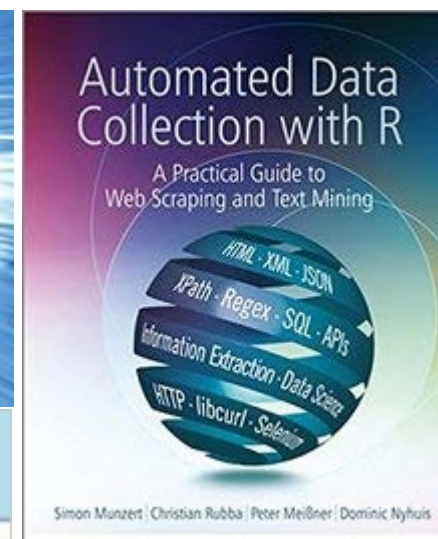
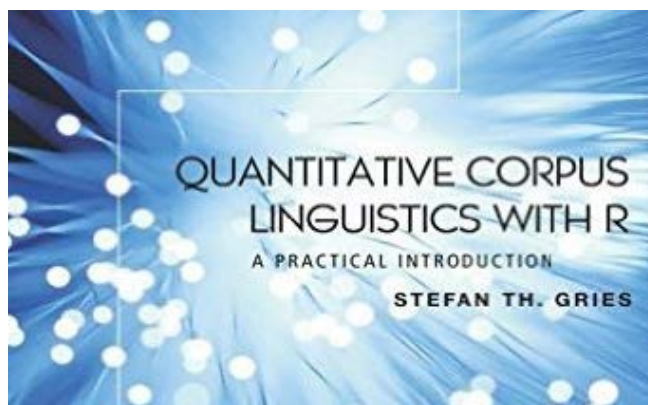
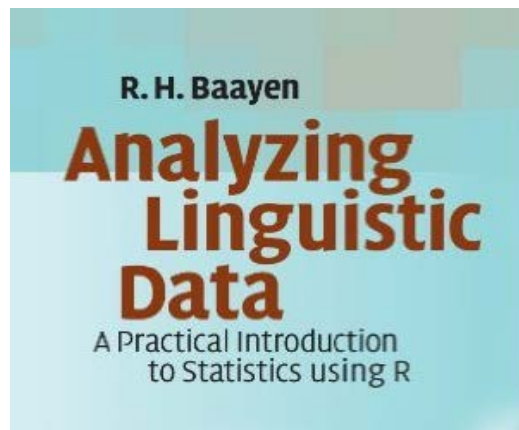
Data Analysis for Linguistics

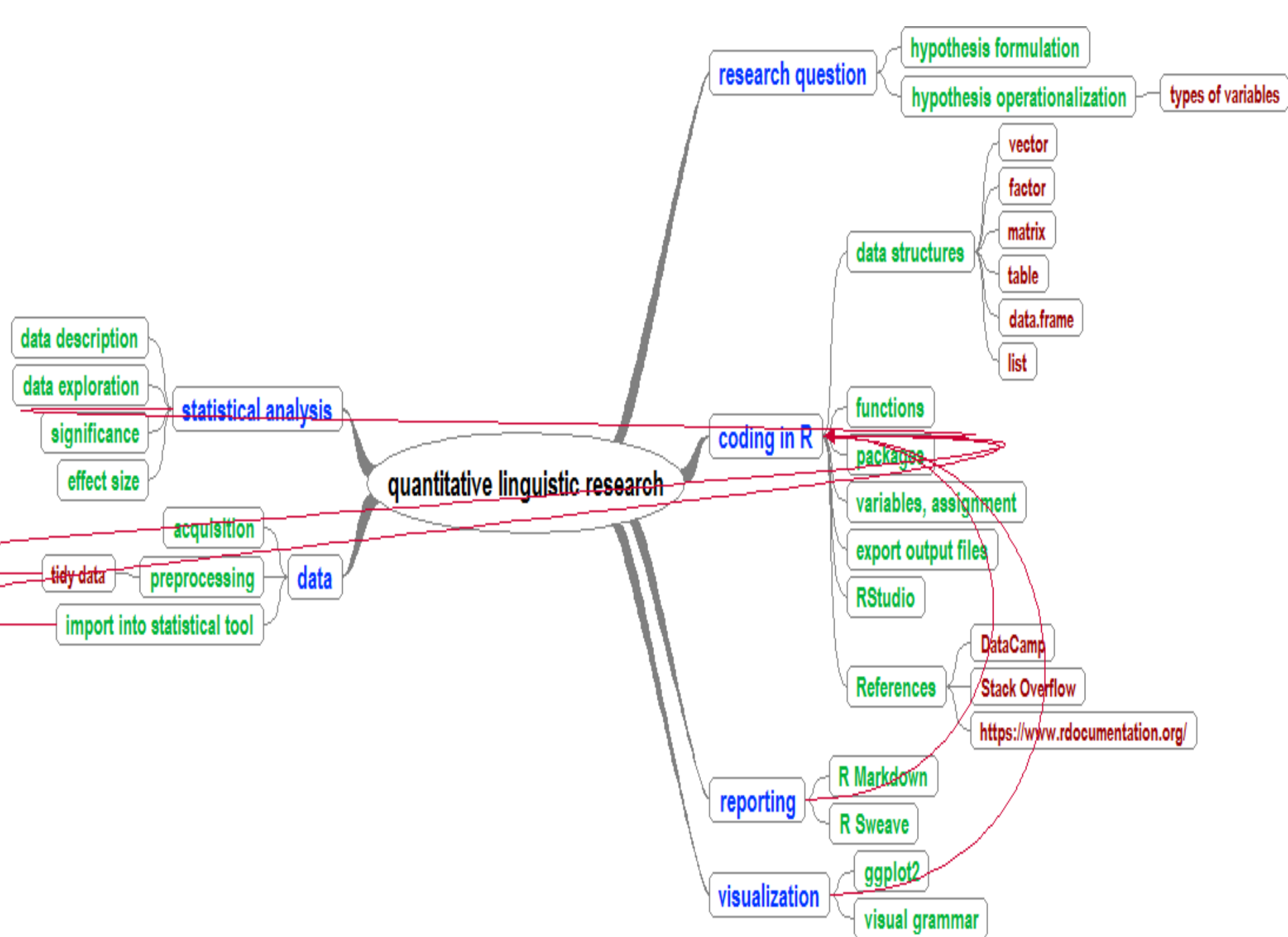
Silvie Cinková, UFAL - 2017-02-07 TextLink Winter School

*The math a linguist needs ain't rocket science.
And no one is asking you to be brilliant, just
competent.*

The Lousy Linguist, 2010

<http://thelousylinguist.blogspot.cz/2010/01/why-linguists-should-study-math.html>





Is discourse structure genre-specific?

PDT
30



Are [they] *or* aren't they right?

Mají, *či* nemají pravdu?

contingency,
contrast,
expansion,
temporal

advice, caption, collection, comment,
description, essay, invitation, letter,
news, other, overview, plot, program,
review, sport, survey, topic_interview,
weather

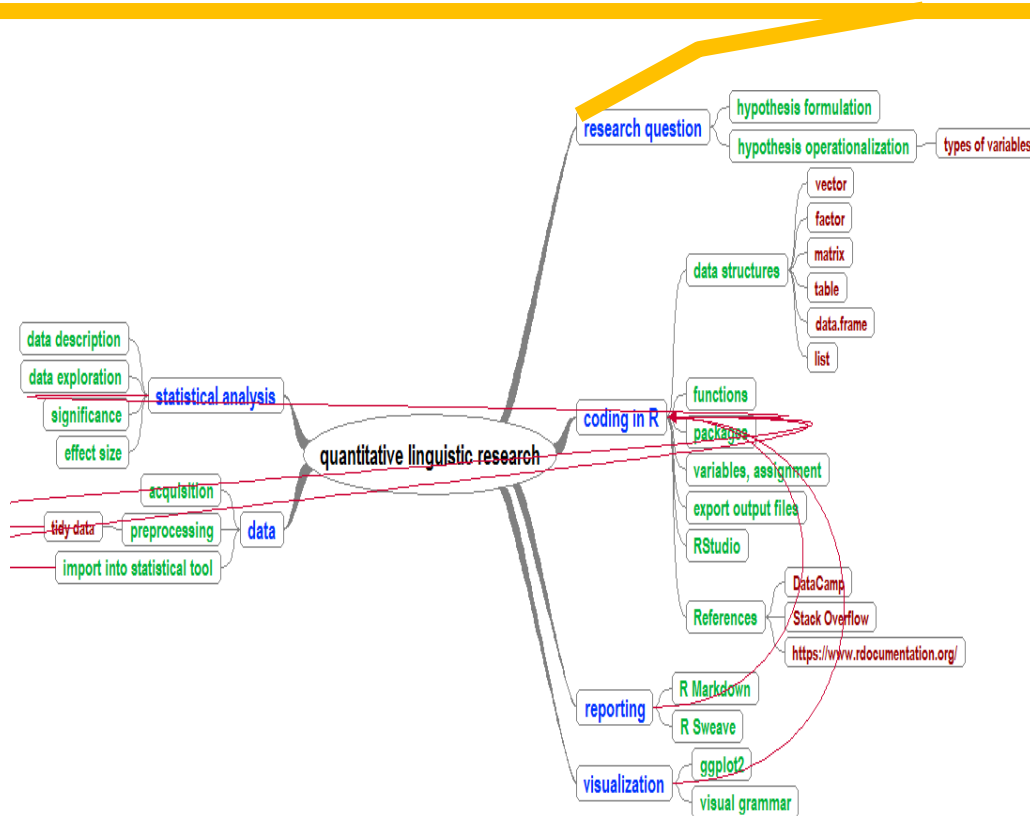
	document_id^	genre^	number_of_sentences^	sentence_id^	discourse_type^	discourse_class^	
1	cmpr9410_001	comment	33	t-cmpr9410-001-p10s2	conc	CONTRAST	^
2	cmpr9410_001	comment	33	t-cmpr9410-001-p11s2	opp	CONTRAST	
3	cmpr9410_001	comment	33	t-cmpr9410-001-p15s1	spec	EXPANSION	
4	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	restr	CONTRAST	
5	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	conj	EXPANSION	

research question

hypothesis formulation

hypothesis operationalization

types of variables

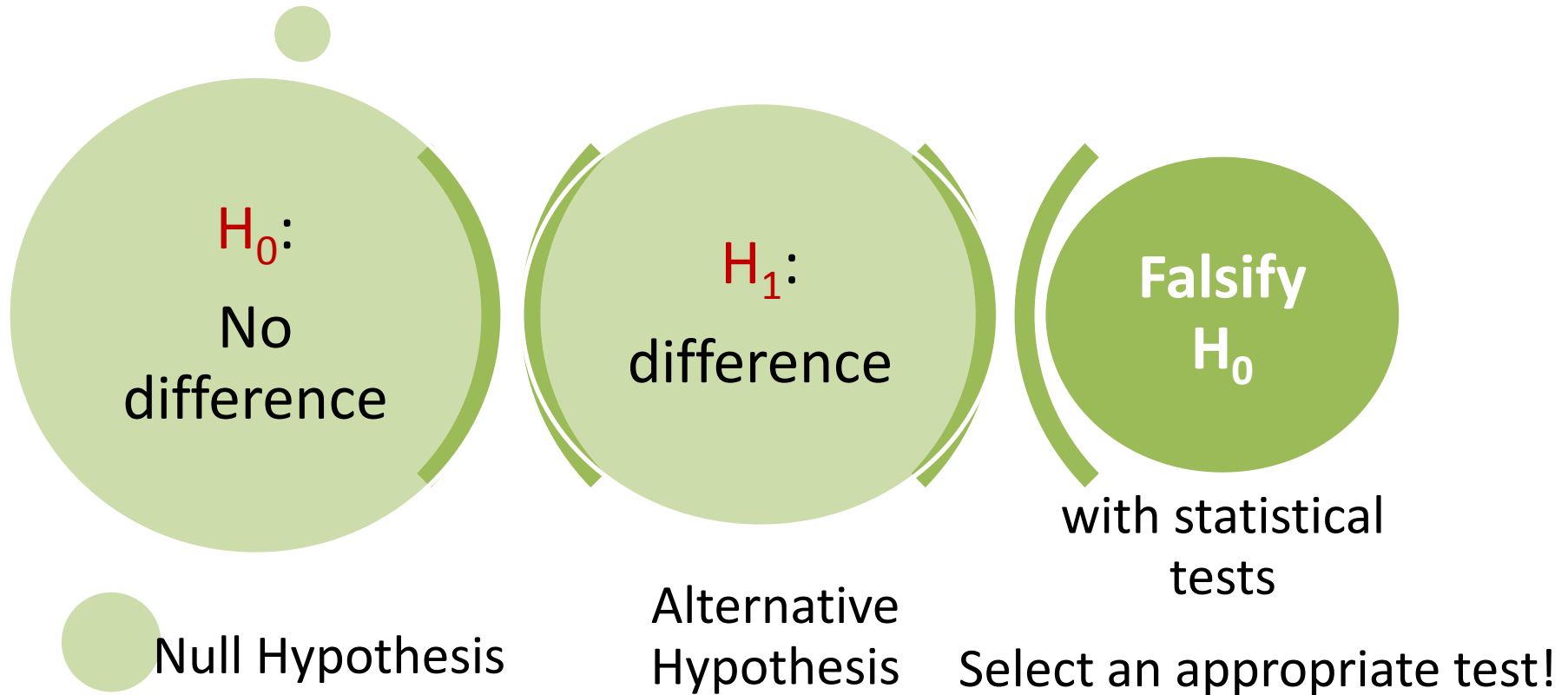


Hypothesis

- general statement concerned with more than just a singular event
- *if... then, the... the...* and their paraphrases
- Potentially falsifiable
 - no hedges: *may, possibly, slightly, often...*
- variables operationalized
 - length of sentence: *tokens? characters? time?*

Hypothesis Testing

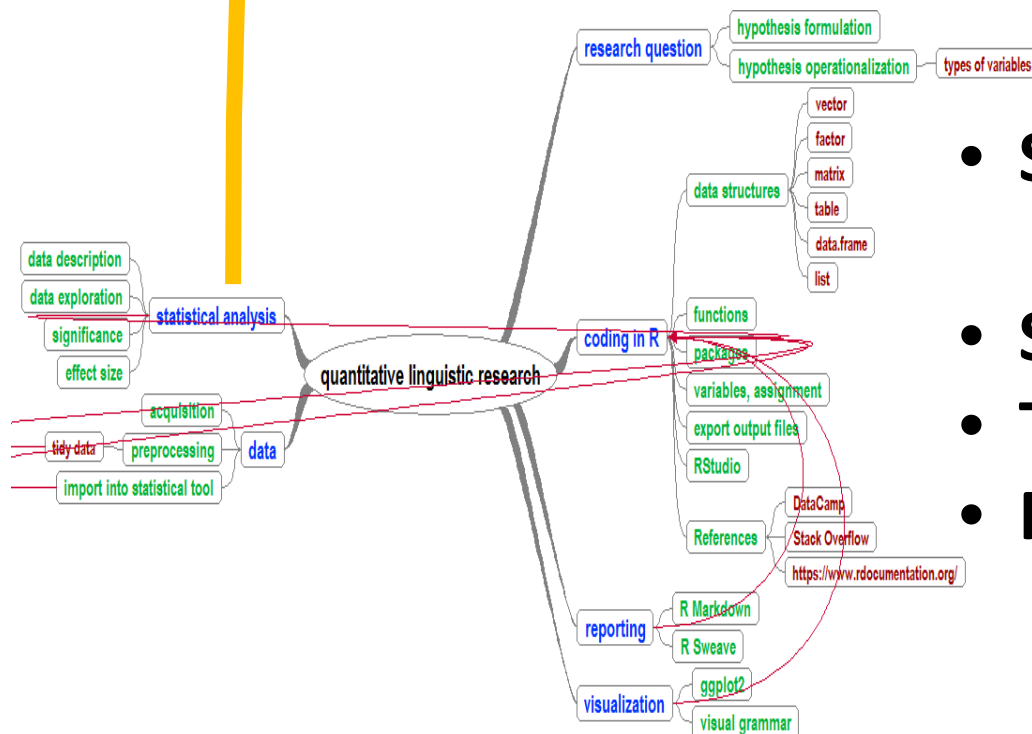
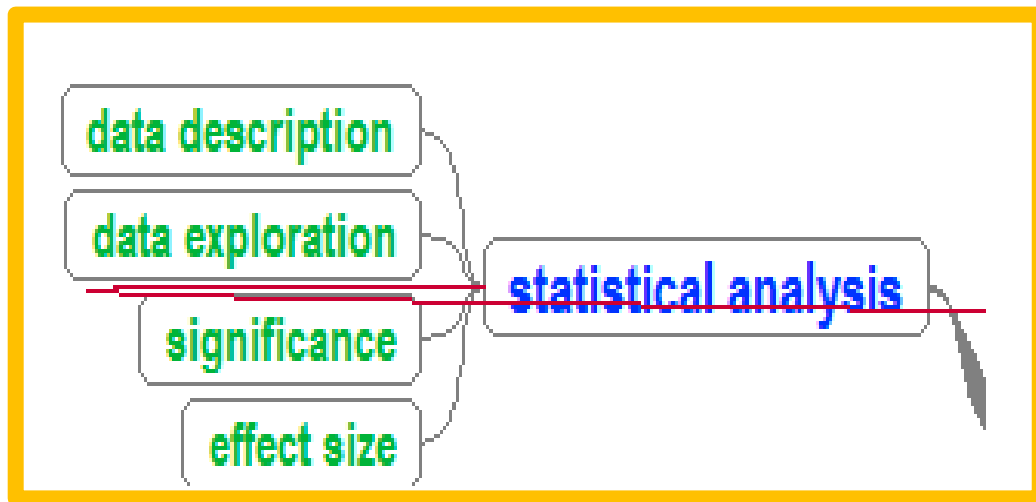
Hypotheses cannot be verified, just falsified. Falsify the opposite!



Operationalization

- Discourse structure = distribution of different discourse classes
- Hypothesis: Genres differ wrt the distribution of different discourse classes. Genres and distribution of discourse classes are dependent of each other. There is an association between genre and discourse class.

	document_id^	genre ^	number_of_sentences^	sentence_id ^	discourse_type^	discourse_class^	
1	cmpr9410_001	comment	33	t-cmpr9410-001-p10s2	conc	CONTRAST	^
2	cmpr9410_001	comment	33	t-cmpr9410-001-p11s2	opp	CONTRAST	
3	cmpr9410_001	comment	33	t-cmpr9410-001-p15s1	spec	EXPANSION	
4	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	restr	CONTRAST	
5	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	conj	EXPANSION	



- **Set alpha level**
(aka significance level)
- **Select appropriate test**
- **Test H_0**
- **Reject or keep H_0**

Test Selection

- How many indep. variables?
- Which types of variables?
- How many observations?
- Distribution of variable values?
- Computing power?



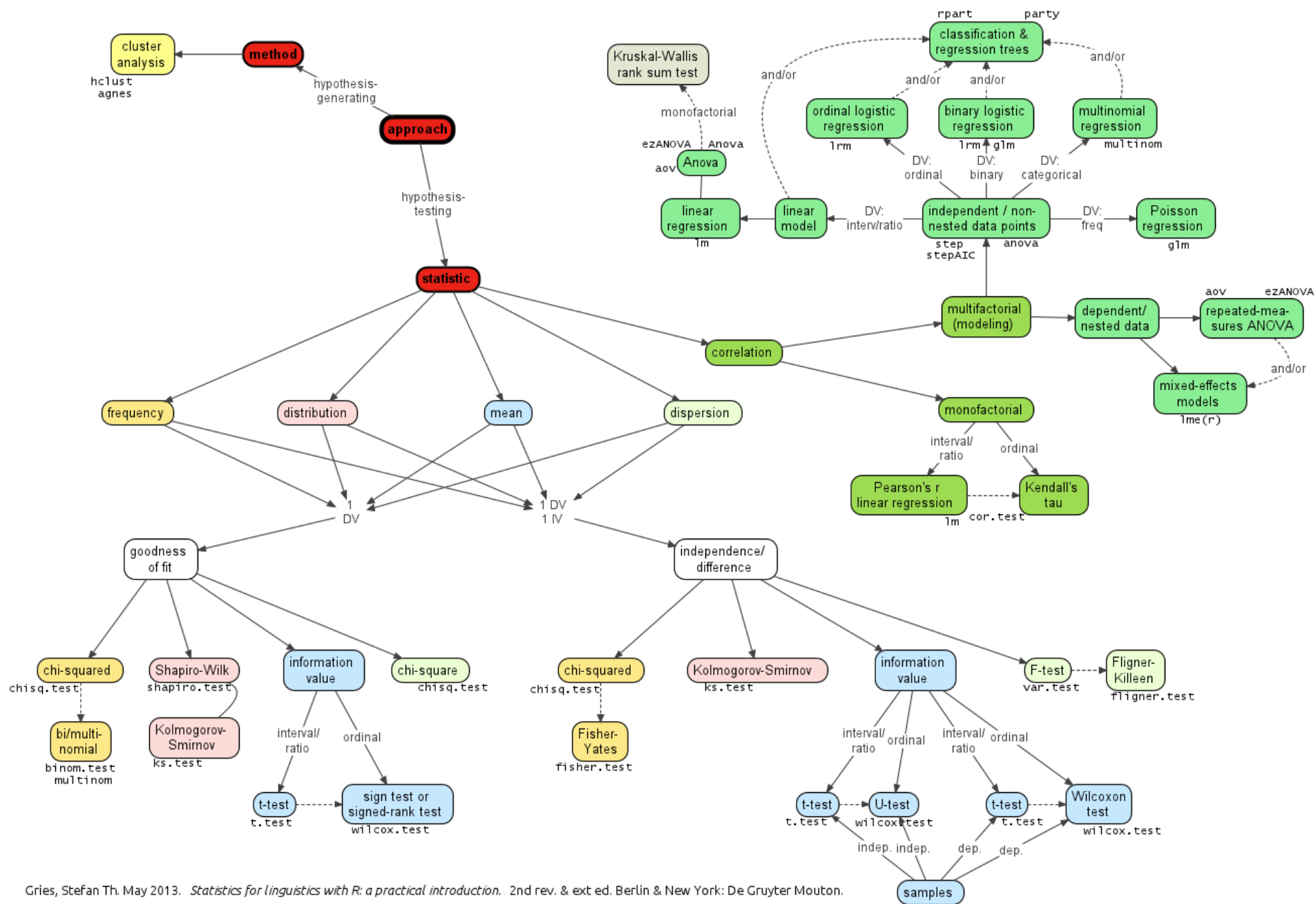
Some tests assume the sample data fit a certain distribution
Test *goodness-of-fit* to choose right

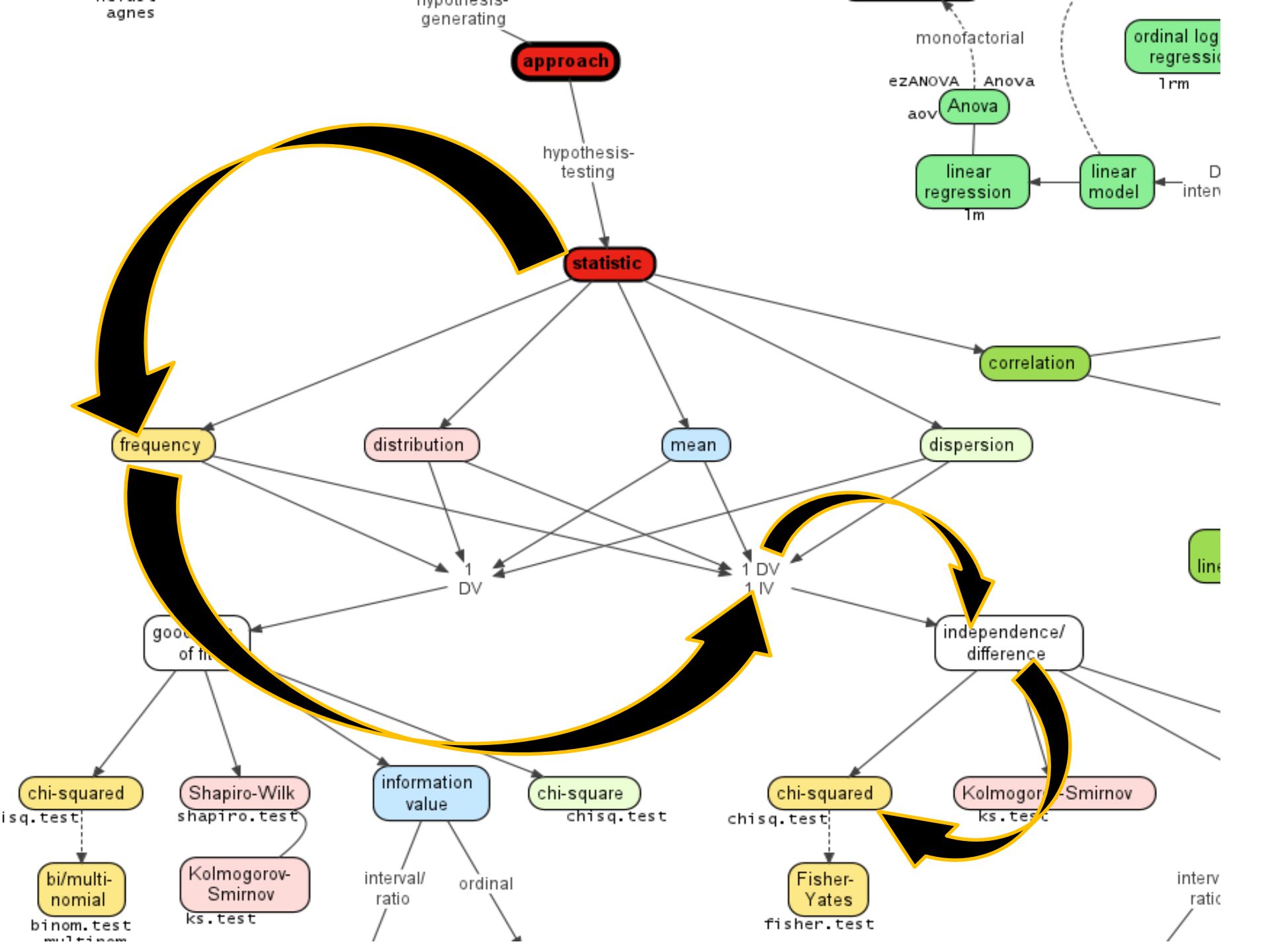
Variable Types

- dependent
 - the observed one
- independent
 - the controlled one

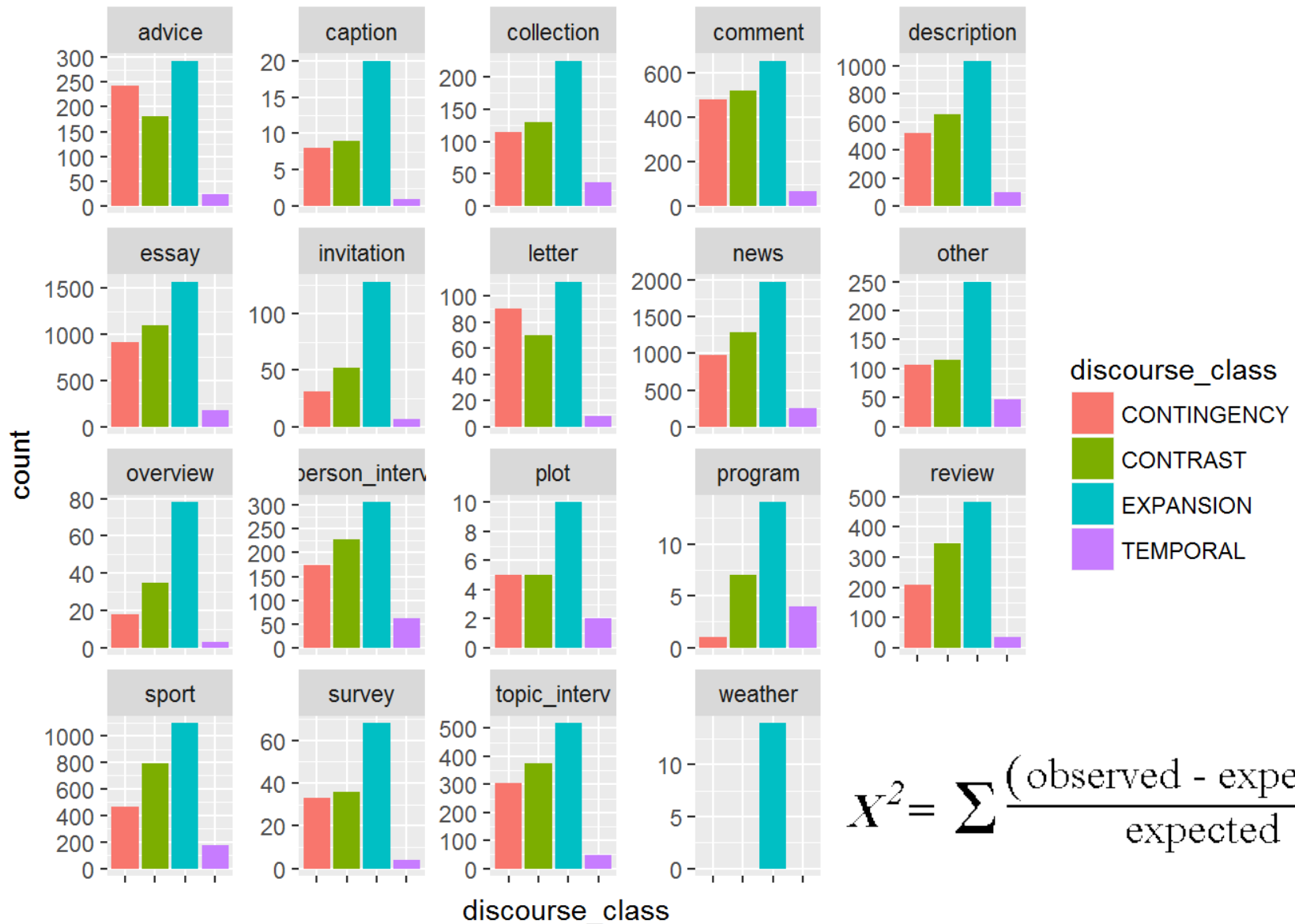
- quantitative + continuous
 - weight, height...
- quantitative + discrete
 - frequencies

- categorical
 - male, female
 - brown, blue, green
- categorical + ordinal
 - beginner < intermediate < advanced
- quantitative + interval
 - no zero point (Fahrenheit scale)
- quantitative + ratio
 - zero point





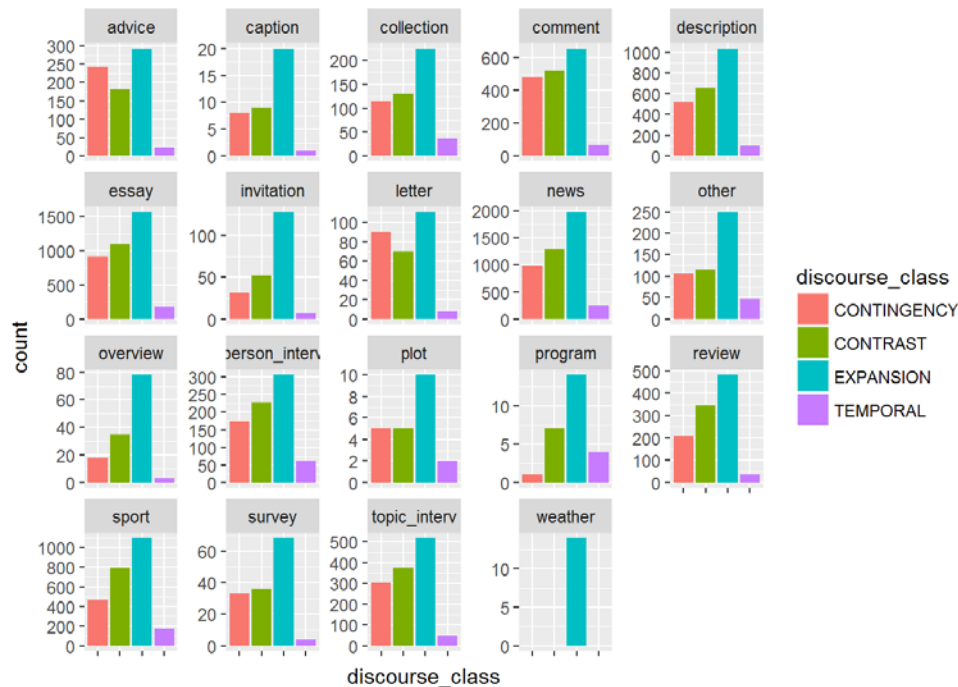
Pearsons' Chi-Squared Test of Independence

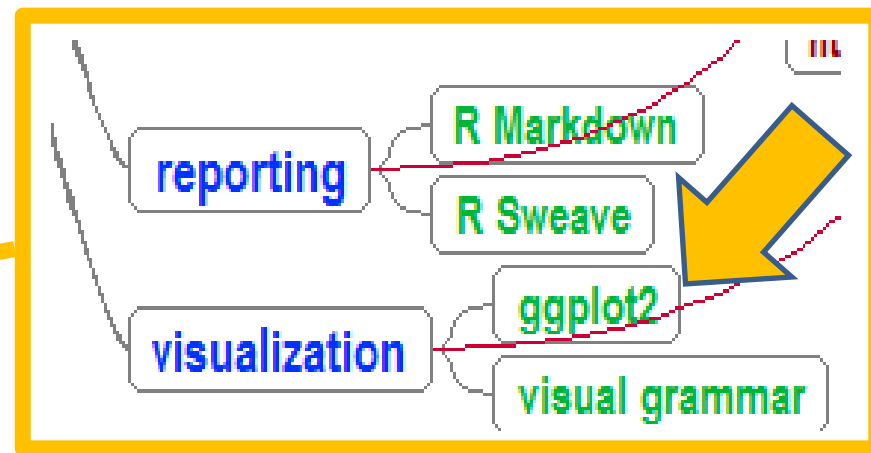
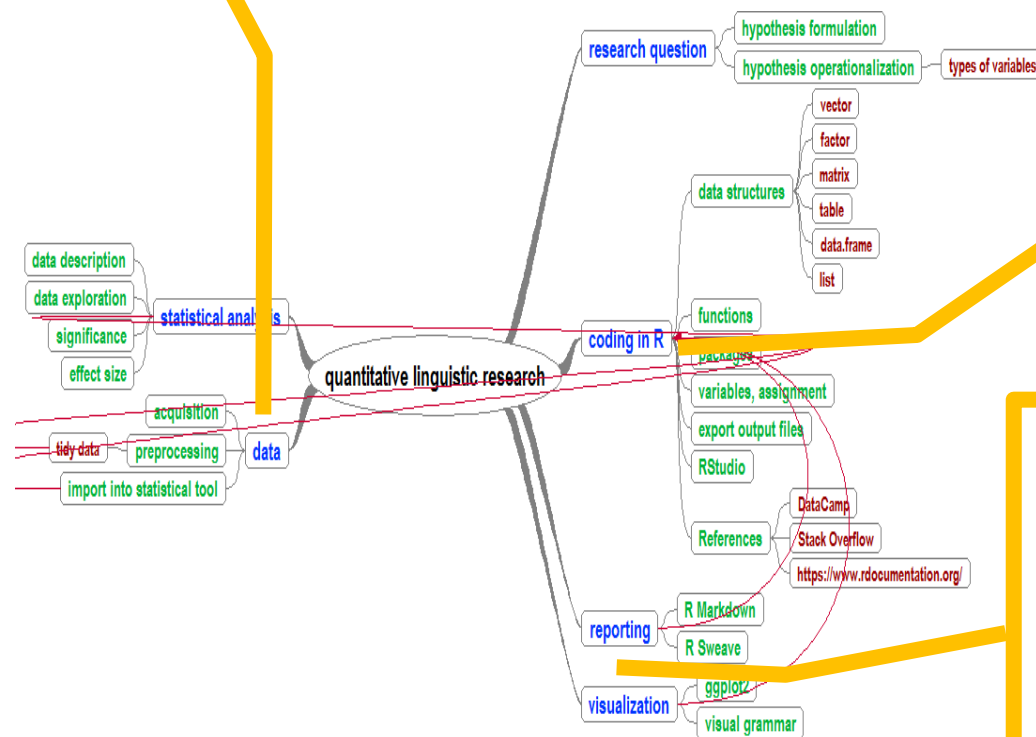
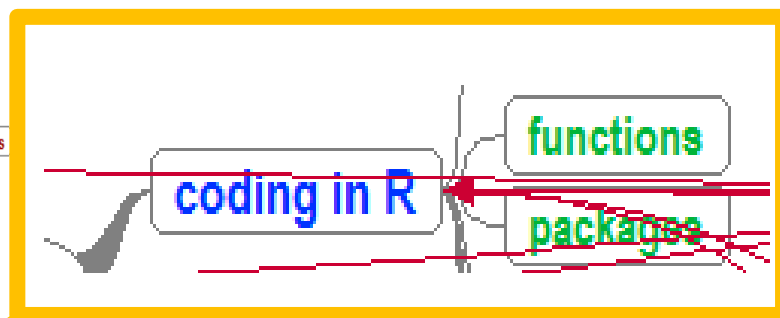
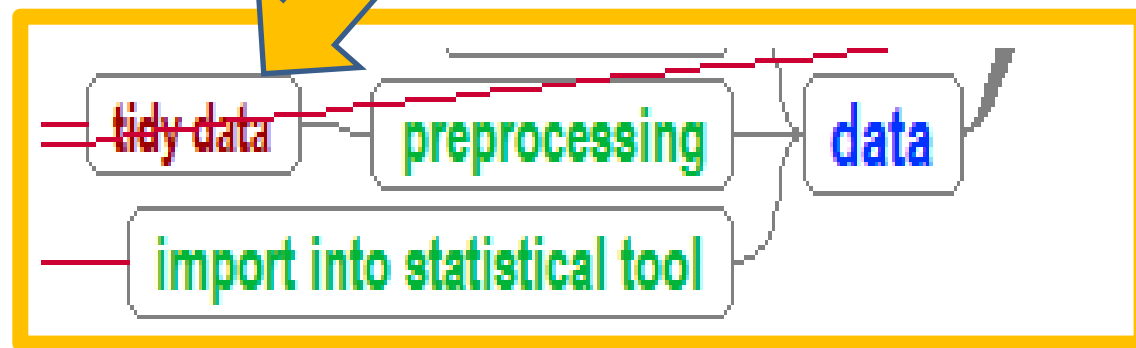


$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Assumptions of Chisq

1. Samples must be independent
(i.e. observations not paired)
2. All **expected** values must exceed 5





Initial Questions about Data Set

Data Cleaning

- What is one observation?
- Distribution of text lengths
 - (how many how long texts?)
- How much bulk text for each genre?
 - no matter in how many documents
- Distribution of text lengths genrewise
 - (how many how long texts for each genre?)
- How many connectives in the texts genrewise?
- Examine strange cases, remove errors

What is one observation here, conceptually?

	document_id^	genre^	number_of_sentences^	sentence_id^	discourse_type^	discourse_class^	
1	cmpr9410_001	comment	33	t-cmpr9410-001-p10s2	conc	CONTRAST	^
2	cmpr9410_001	comment	33	t-cmpr9410-001-p11s2	opp	CONTRAST	
3	cmpr9410_001	comment	33	t-cmpr9410-001-p15s1	spec	EXPANSION	
4	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	restr	CONTRAST	
5	cmpr9410_001	comment	33	t-cmpr9410-001-p17s1	conj	EXPANSION	

Summarizing the Corpus I

```
Classes 'tbl_df', 'tbl' and 'data.frame':    20556 obs. of  6 variables:
 $ document_id      : Factor w/ 2580 levels "cmpr9410_001",...: 1 1 1 1 1
 $ genre            : Factor w/ 19 levels "advice","caption",...: 4 4 4 4
 $ number_of_sentences: int   33 33 33 33 33 33 33 33 33 33 33 ...
 $ sentence_id      : Factor w/ 16112 levels "t-cmpr9410-001-p10s2",...: 1
 $ discourse_type    : Factor w/ 23 levels "conc","cond",...: 1 16 22 21 4
 $ discourse_class   : Factor w/ 4 levels "CONTINGENCY",...: 2 2 3 2 3 2 2
```

Summarizing the Corpus II

```
summary(pdt30)
```

```
##          document_id          genre  number_of_sentences
## ln94207_73 :   123    news          :4510    Min.      :   1.0
## ln94207_76 :   118    essay          :3757    1st Qu.:  16.0
## ln94210_95 :   102    sport          :2525    Median :  26.0
## ln94211_92 :    87    description :2305    Mean     :  37.3
## ln94200_127:    80    comment         :1731    3rd Qu.:  42.0
## ln94207_79 :    77    topic_interv:1247    Max.      :231.0
## (Other)      :19969    (Other)         :4481
##
##          sentence_id    discourse_type    discourse_class
## t-ln94203-125-p8s11 :     6    conj      :7498    CONTINGENCY:4701
## t-ln94207-73-p13s7  :     6    opp       :3196    CONTRAST     :5938
## t-ln94207-90-p4s9   :     6    reason    :2632    EXPANSION     :8851
## t-ln95048-061-p7s3  :     6    cond      :1369    TEMPORAL      :1066
## t-cmpr9415-033-p11s4:     5    conc      : 880
## t-ln94203-99-p8s1   :     5    preced    : 840
## (Other)              :20522    (Other) :4141
```

Contingency Table (2-Way Table)

Genre vs. Discourse Class

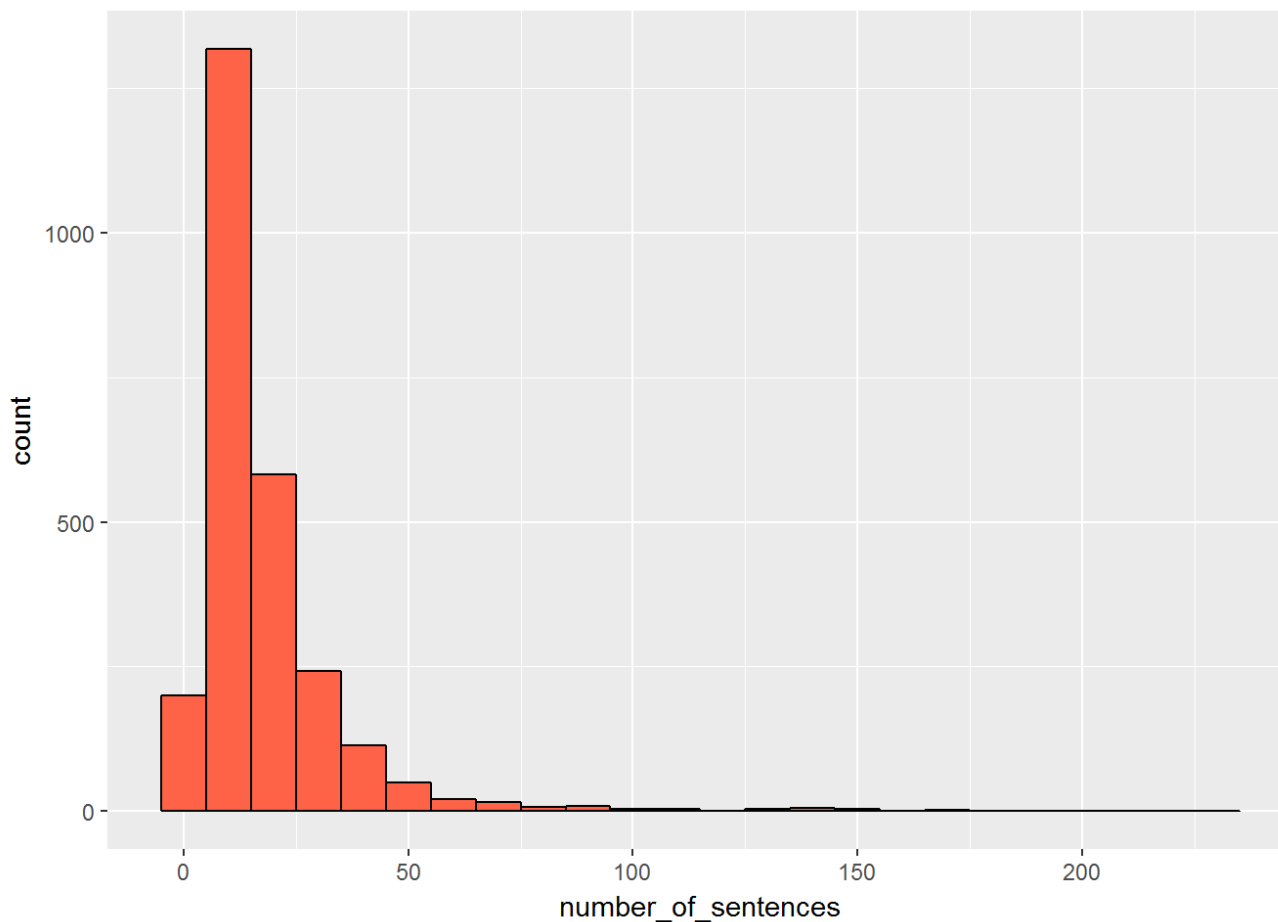
##		discourse_class			
## genre		CONTINGENCY	CONTRAST	EXPANSION	TEMPORAL
##	advice	242	182	293	24
##	caption	8	9	20	1
##	collection	115	130	225	37
##	comment	483	523	656	69
##	description	520	651	1034	100
##	essay	913	1092	1569	183
##	invitation	31	52	128	7
##	letter	90	70	111	8
##	news	983	1289	1979	259
##	other	107	116	250	48
##	overview	18	35	78	3
##	person_interv	173	227	306	61
##	plot	5	5	10	2
##	program	1	7	14	4
##	review	209	346	482	35
##	sport	464	794	1095	172
##	survey	33	36	68	4
##	topic_interv	306	374	518	49
##	weather	0	0	14	0

Summary of Text Lengths

```
# A tibble: 10 × 3
  document_id      genre number_of_sentences
  <fctr>         <fctr>          <int>
1 mf920925_116    news              7
2 mf920925_120 person_interv         52
3 ln94203_75      description        25
4 cmpr9415_018    comment           26
5 ln95045_059     news              6
6 ln95046_078     news              5
7 ln95047_120     news             13
8 cmpr9410_008    advice           64
9 cmpr9413_052    essay            58
10 ln95045_110    news             11
```

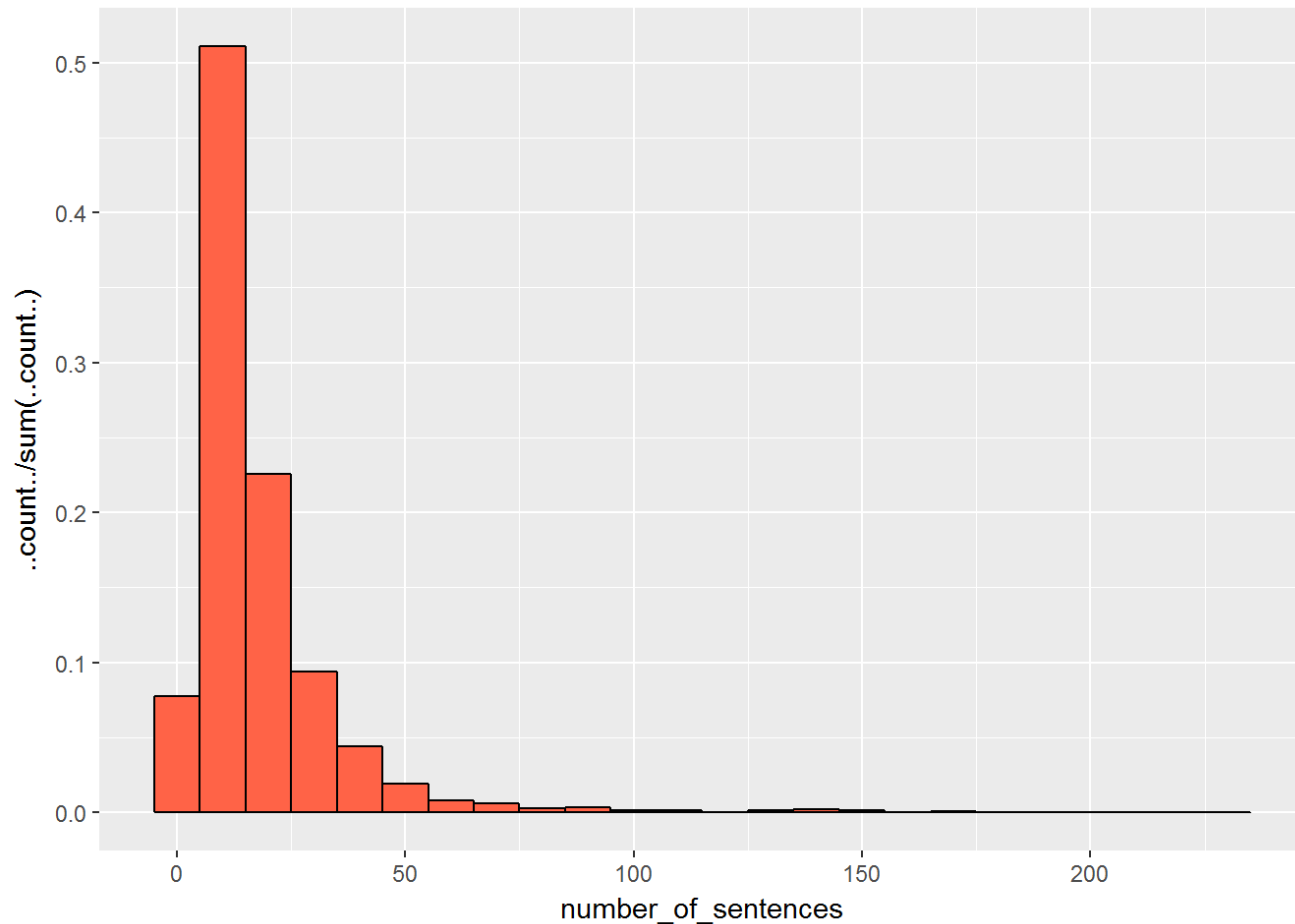
Distribution of Text Lengths

Histogram with Bins, *Absolute Frequencies*

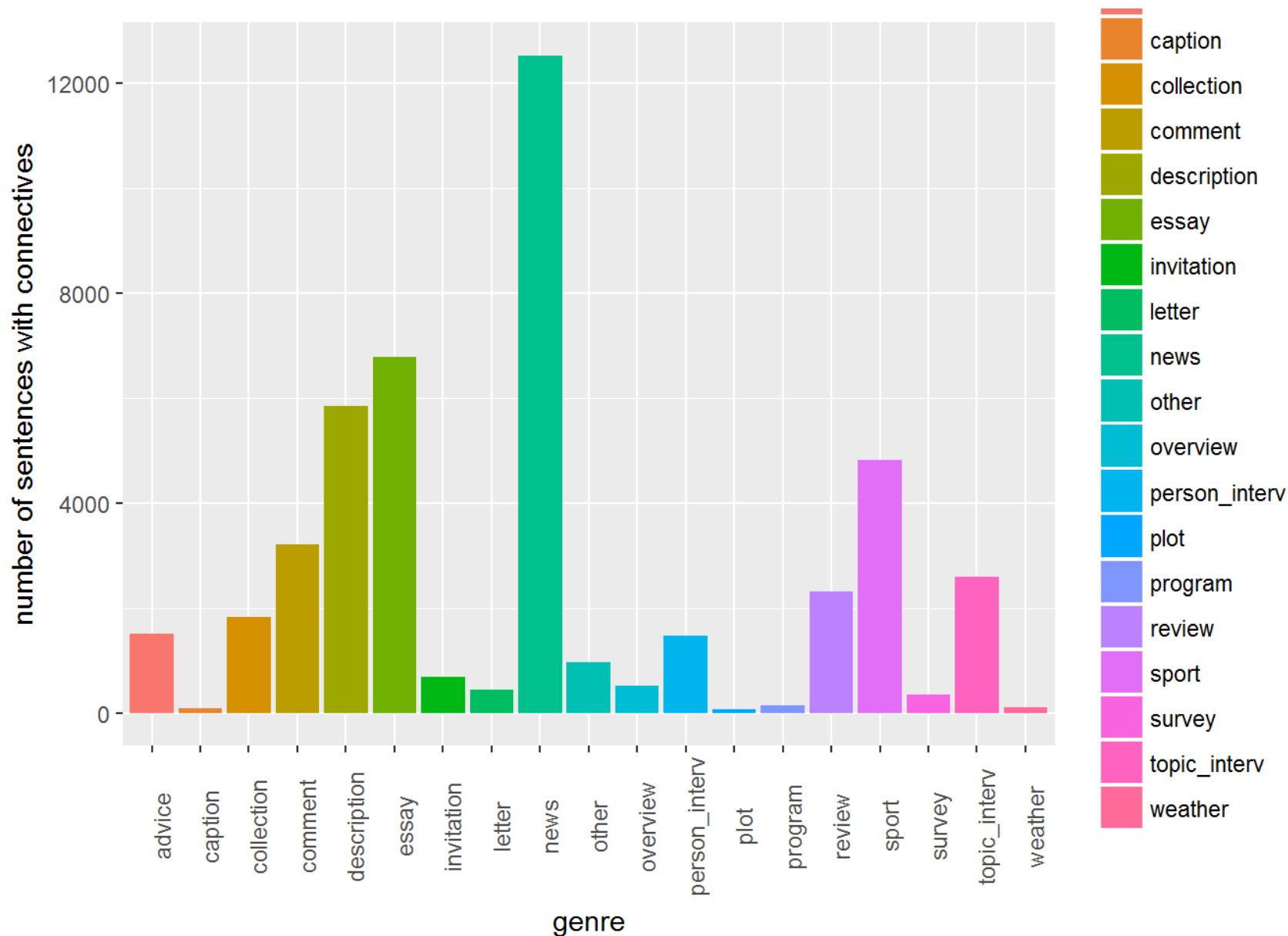


Distribution of Text Lengths

Histogram with Bins, Relative Frequencies

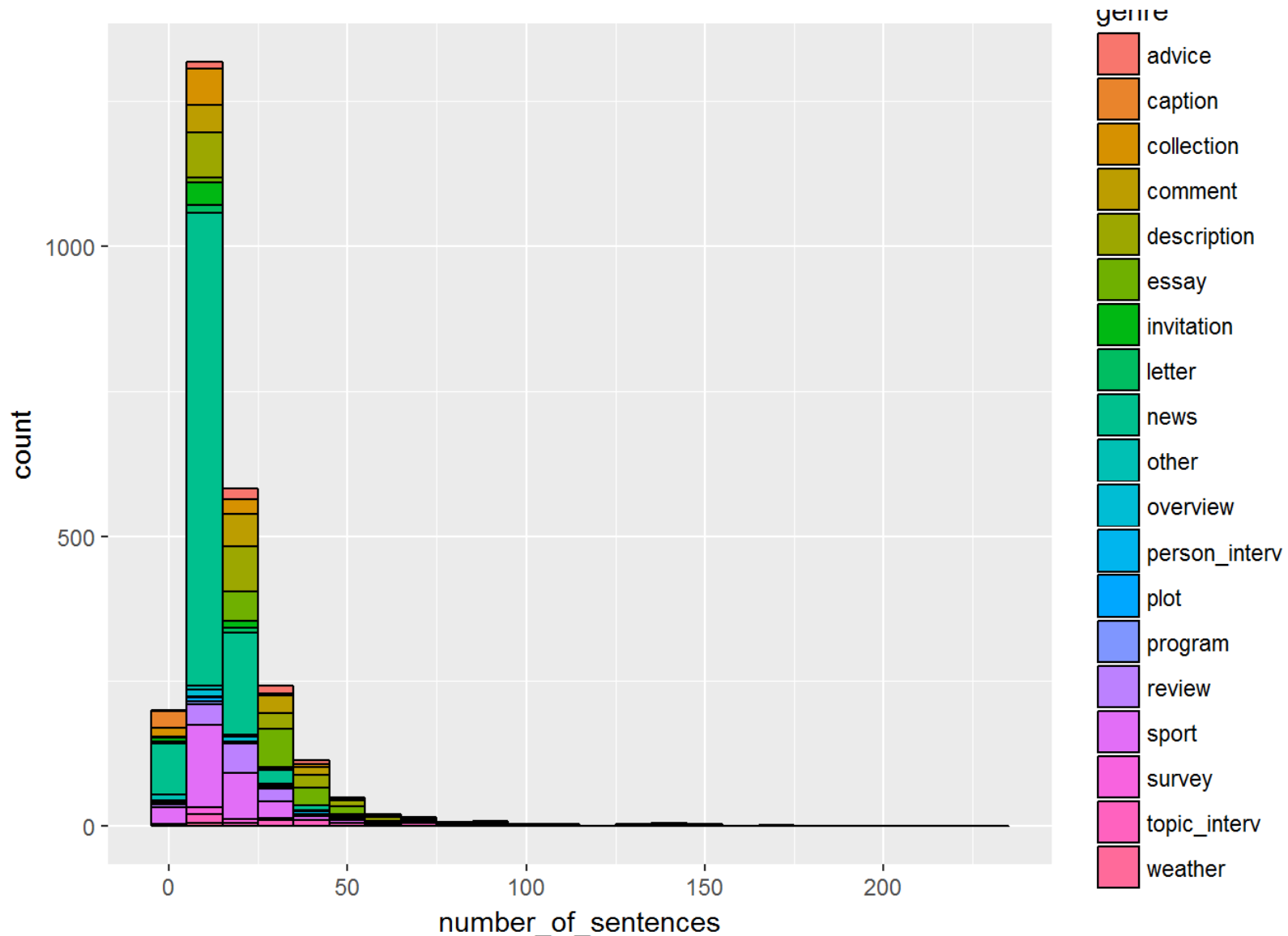


Genre vs. Bulk Text Length - Bar Plot



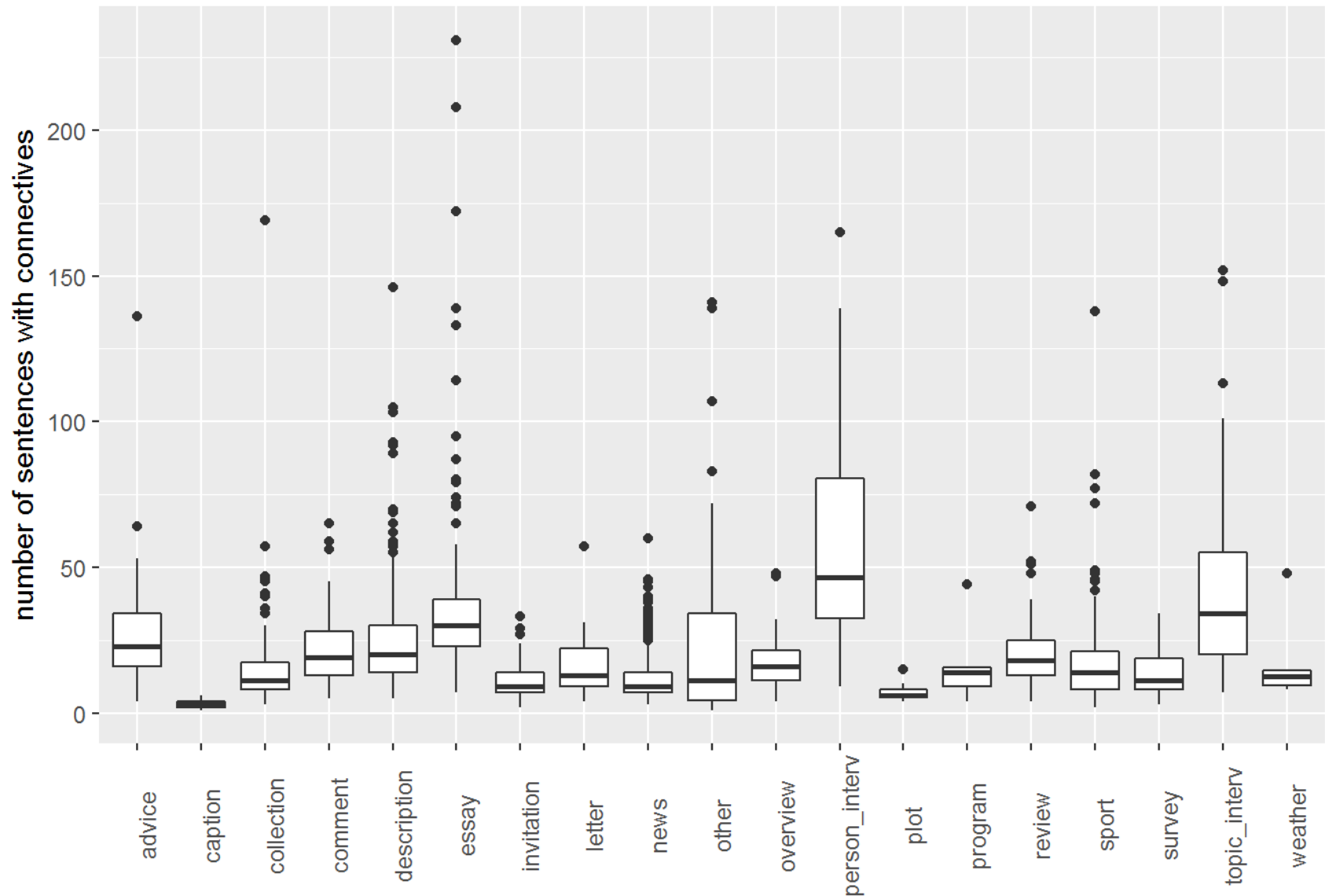
Distribution of Text Lengths Genrewise

Histogram with Additional Variable

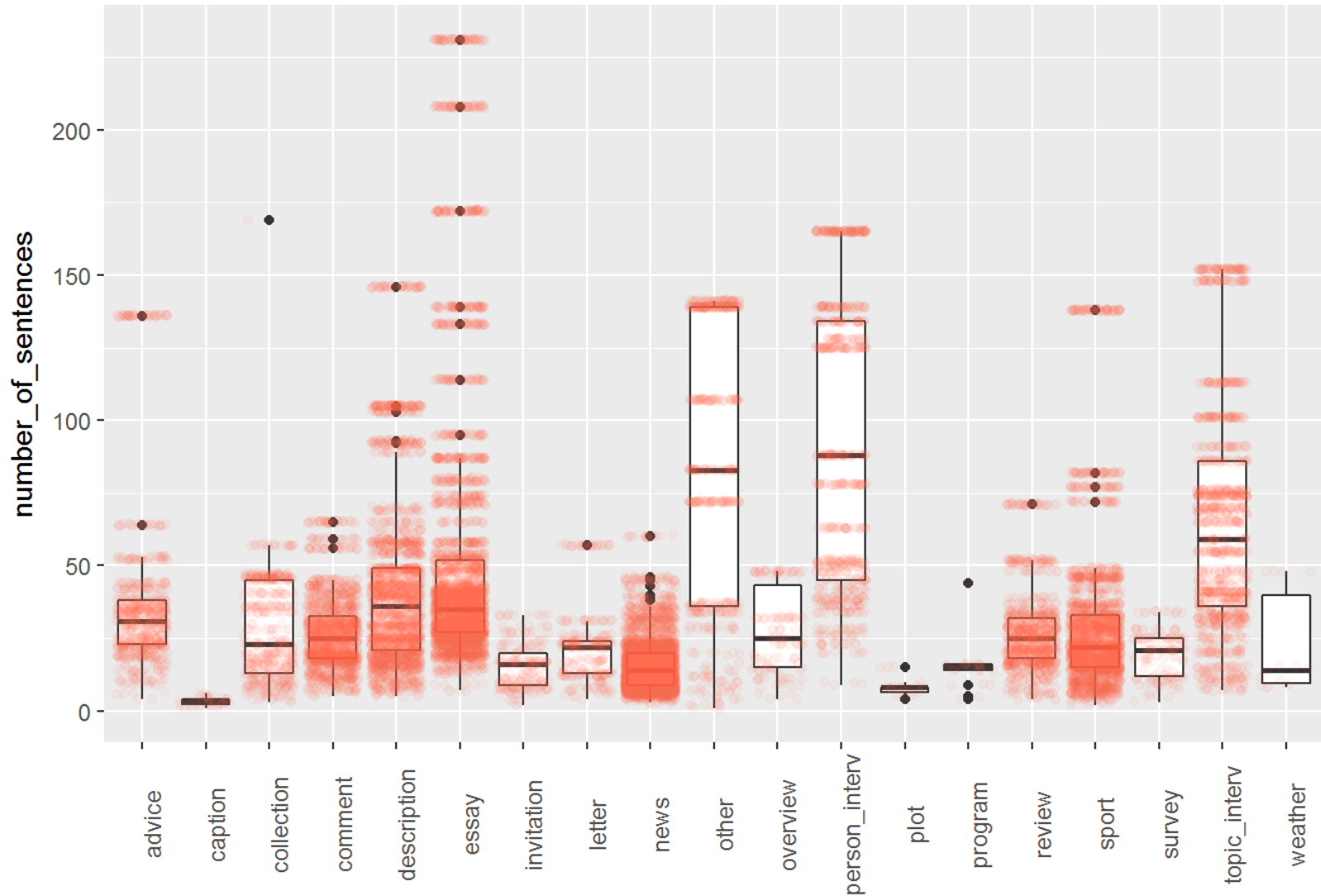


Distribution of Text Lengths Genrewise

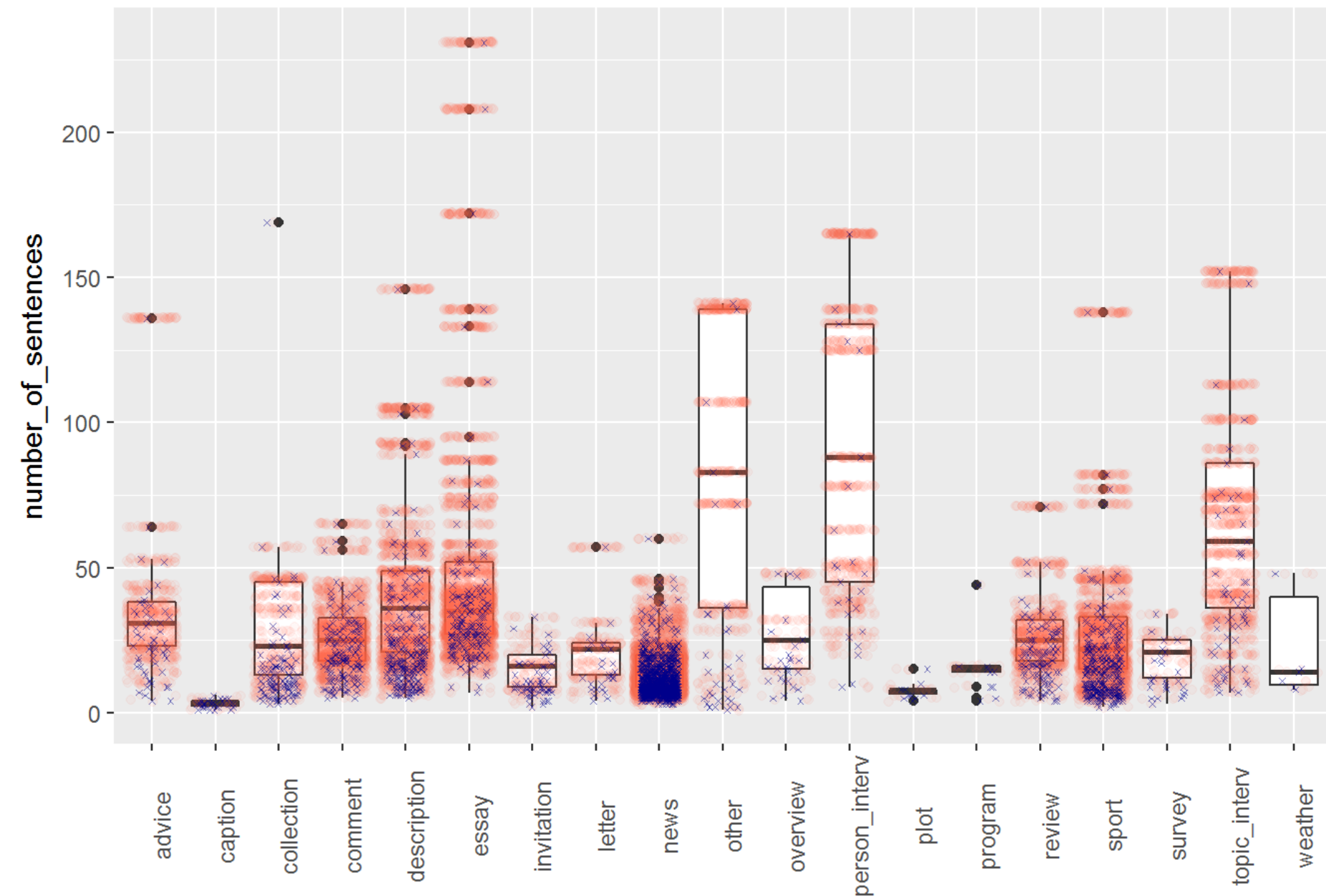
Box Plots with Whiskers



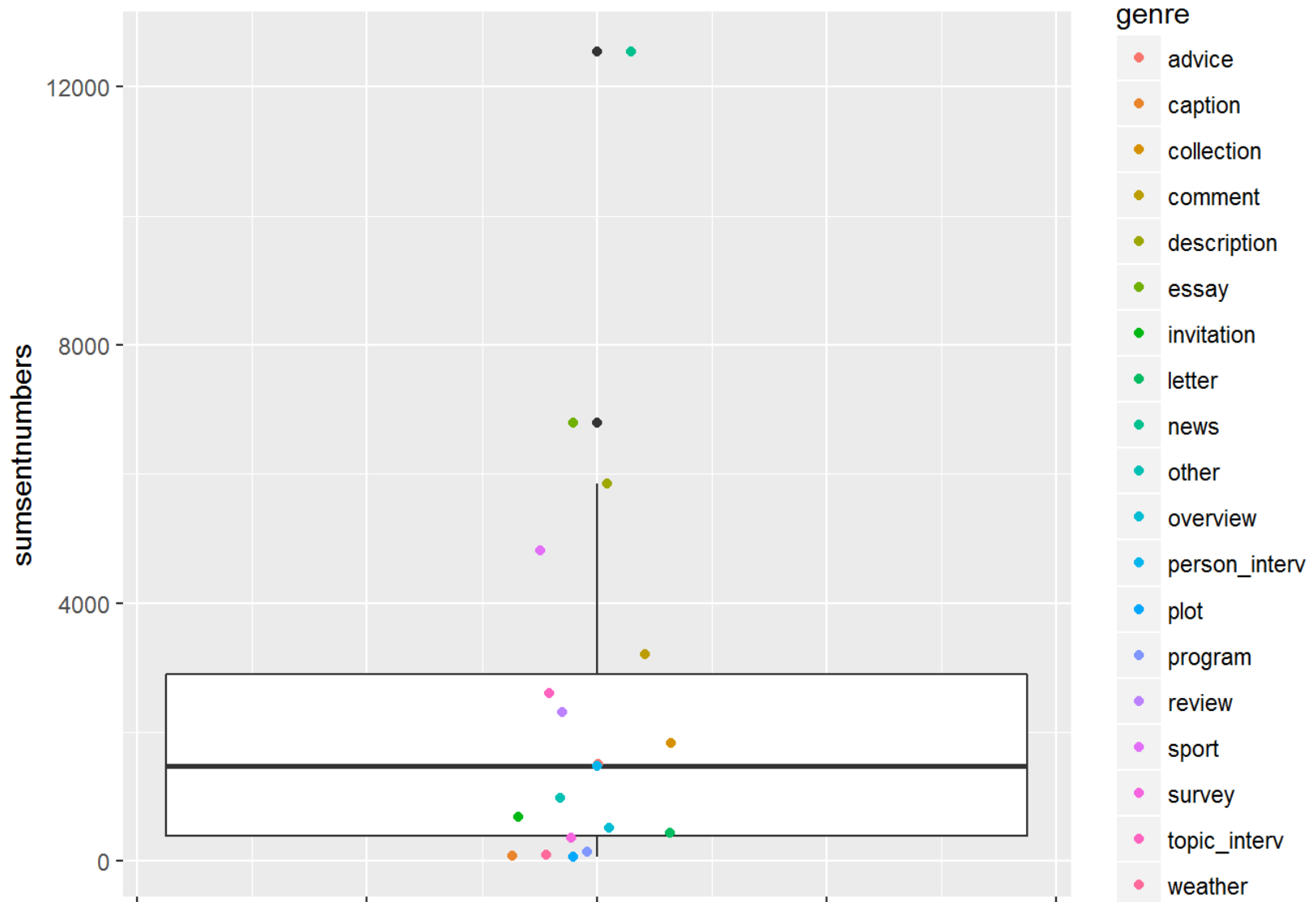
Numbers of Connectives in Different Texts Genrewise



Adding Unique Docs



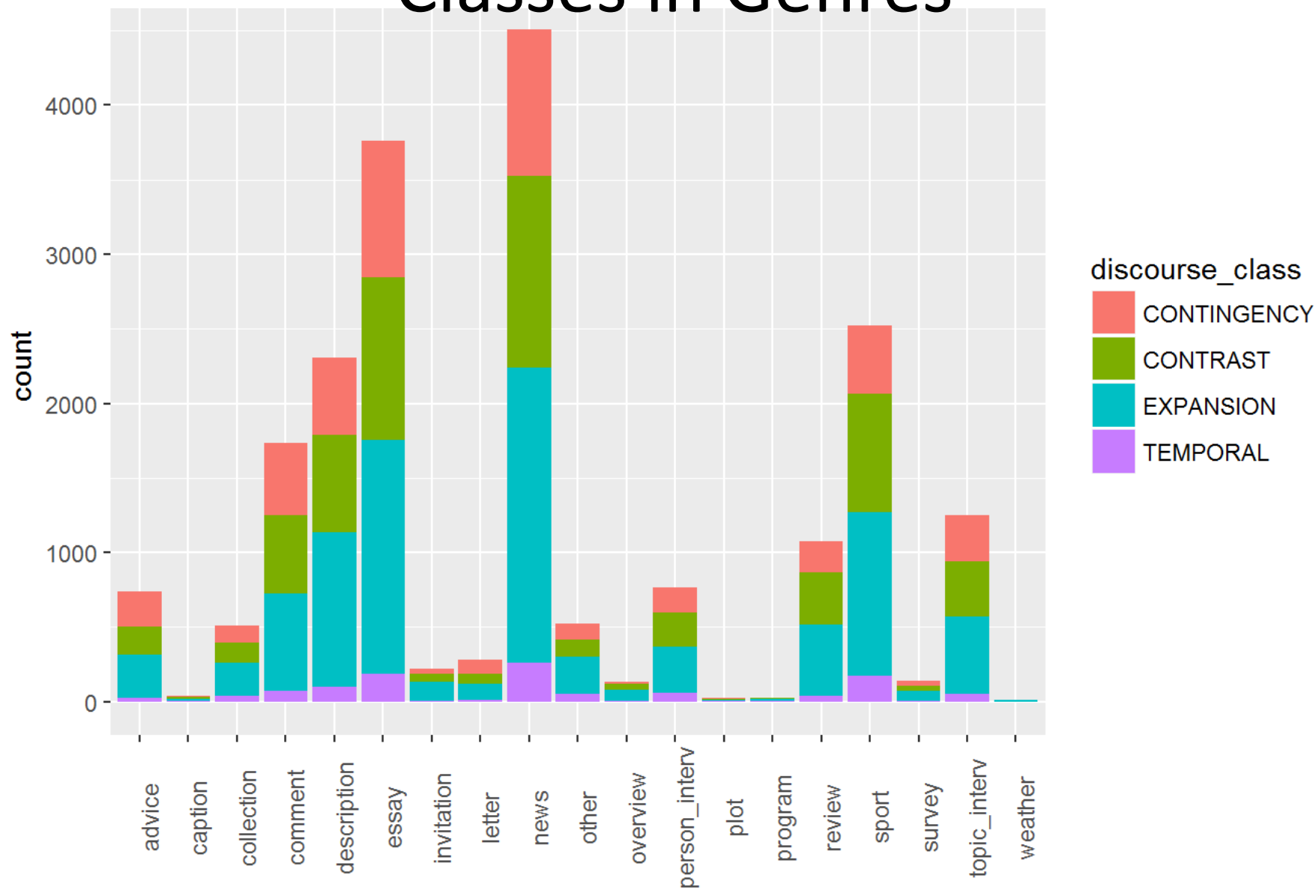
Bulk Text Outlier Genres



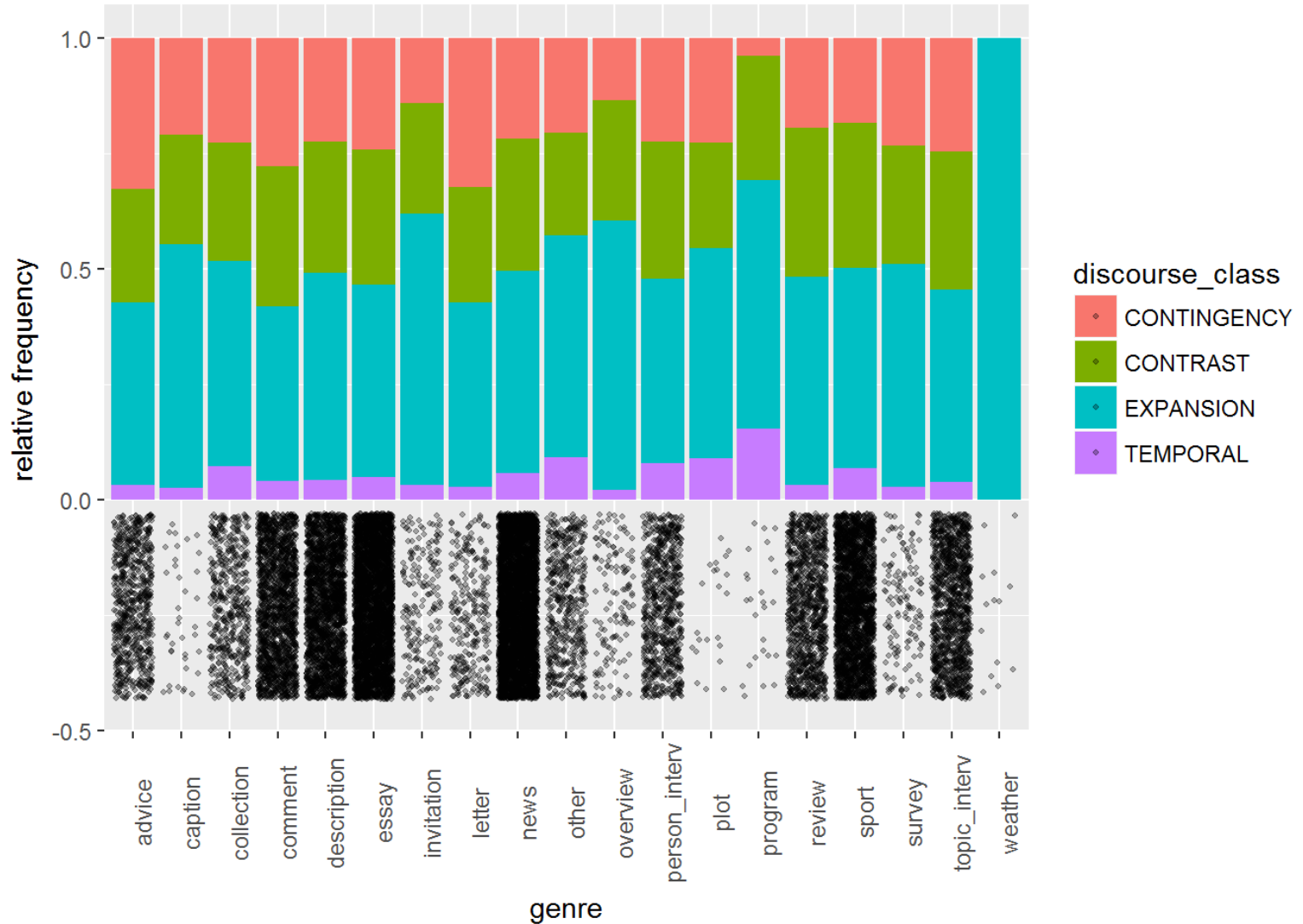
How Many Connective Each Genre?



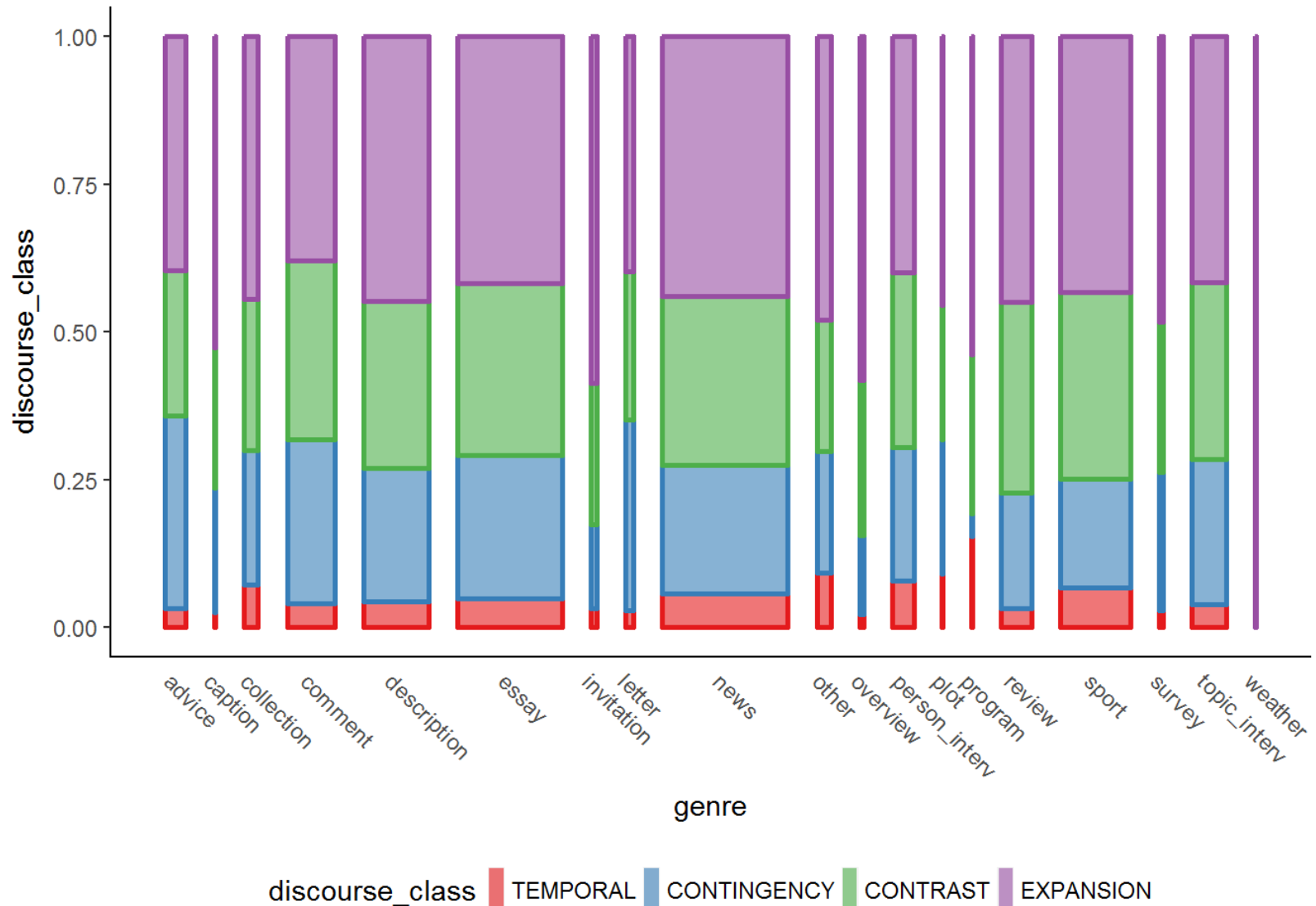
Absolute Frequencies of Discourse Classes in Genres



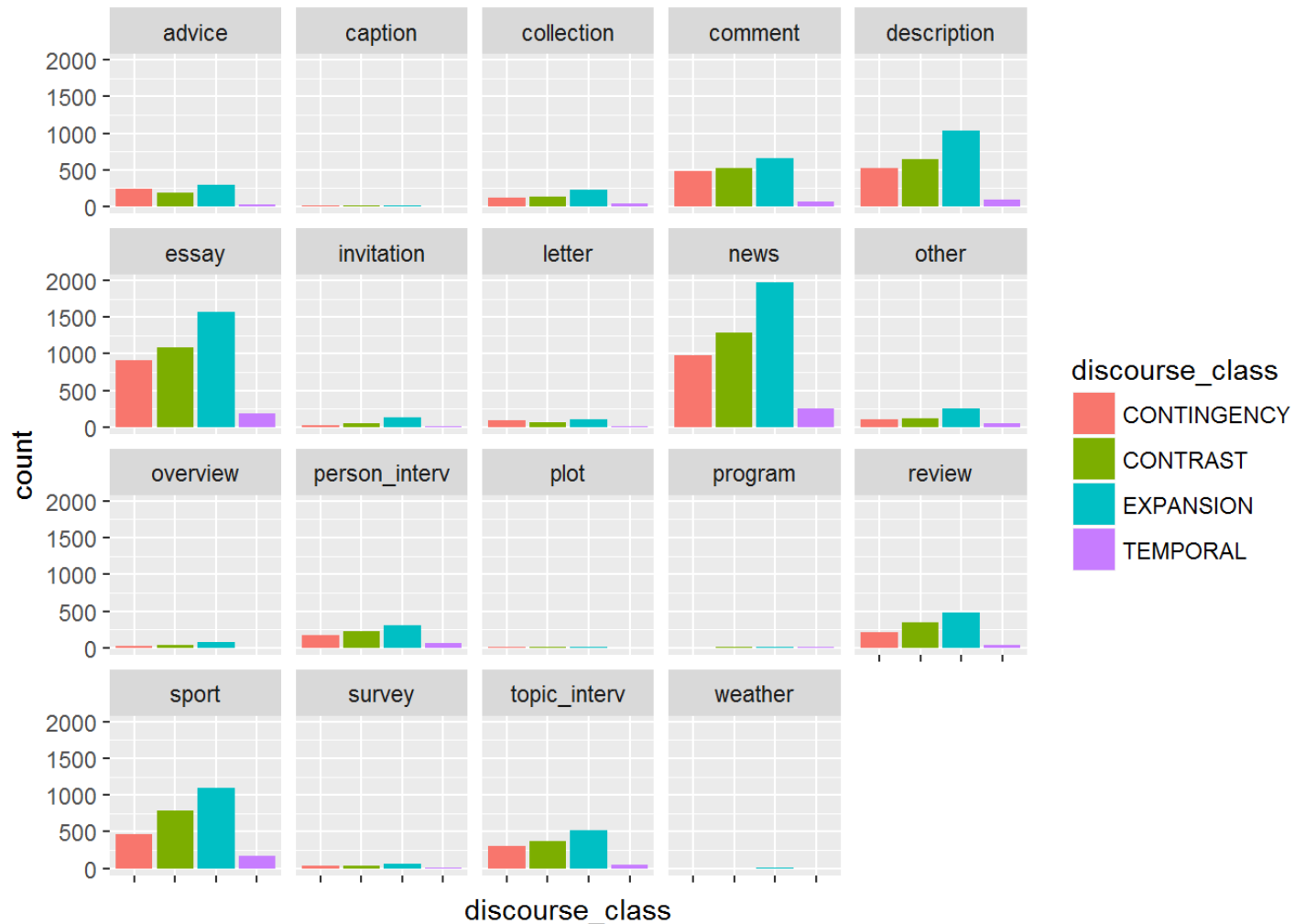
Proportions with Rugs



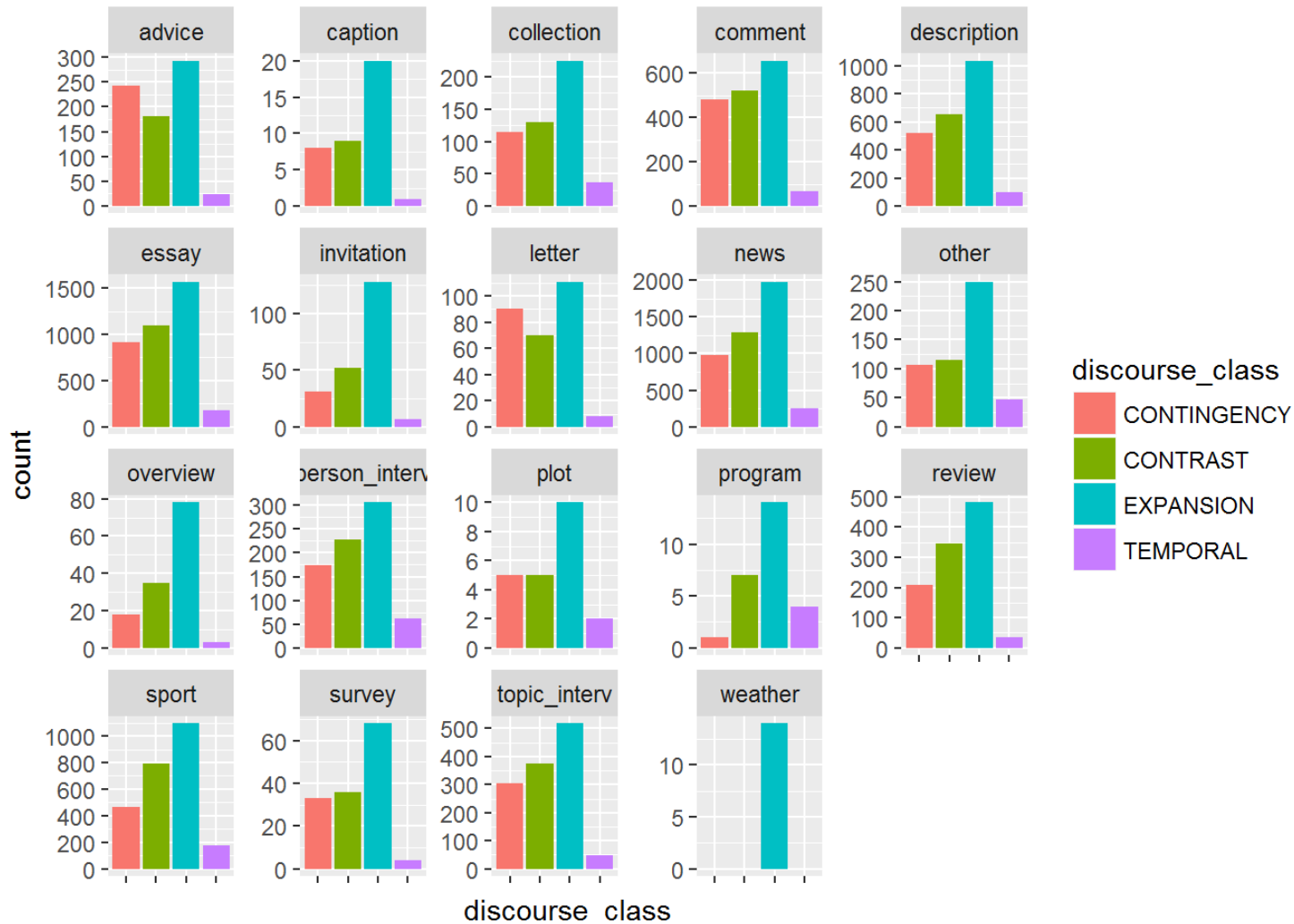
Mosaic Plot - Proportions of Discourse Classes Genrewise



Faceted Absolute Frequencies of Discourse Classes Genrewise, Unified Scale



Faceted Absolute Frequencies of Discourse Classes Genrewise, Individual Scales



Expected Value in Each Cell

discourse_class					
	CONTINGENCY	CONTRAST	EXPANSION	TEMPORAL	
advice	242	182	293	24	741
caption	8	9	20	1	38
collection	115	130	225	37	510
comment	483	523	656	69	1731
description	520	651	1034	1	2206
essay	913	1092	1569	1	3575
invitation	31	52	1	1	85
letter	90	70	1	1	162
news	983	1289	1	1	2274
other	107	1	1	1	210
overview	18	1	1	1	21
person_interv	1	1	306	1	309
plot	1	1	1	1	4
program	1	7	1	1	9
review	209	3	1	35	1072
sport	464	1	195	172	2525
survey	33	1	68	4	141
topic_interv	306	37	518	49	1247
weather	0	0	14	0	14
	4 701	5 398	8 850	1 066	20 555

$$\frac{\text{column_total}}{\text{total}} \times \frac{\text{row_total}}{\text{total}} \times \text{total}$$

$$\frac{4701}{20555} \times \frac{741}{20555} \times 20555 = 169.4693$$

Expected value must be larger than 5!
(sometimes disputed...)

##	CONTINGENCY	CONTRAST	EXPANSION	TEMPORAL
## advice	169.5	214.1	319.0	38.4
## caption	8.7	11.0	16.4	2.0
## collection	116.0	146.5	218.3	26.3
## comment	395.9	500.1	745.3	89.8
## description	527.2	665.9	992.4	119.5
## essay	859.2	1085.3	1617.6	194.8
## invitation	49.9	63.0	93.9	11.3
## letter	63.8	80.6	120.1	14.5
## news	1031.5	1302.9	1941.8	233.9
## other	119.2	150.5	224.3	27.0
## overview	30.6	38.7	57.7	6.9
## person_interv	175.4	221.6	330.2	39.8
## plot	5.0	6.4	9.5	1.1
## program	5.9	7.5	11.2	1.3
## review	245.2	309.7	461.6	55.6
## sport	577.5	729.4	1087.1	130.9
## survey	32.2	40.7	60.7	7.3
## topic_interv	285.2	360.2	536.9	64.7
## weather	3.2	4.0	6.0	0.7

Chisq Test

```
(result <- chisq.test(pdt30$genre, pdt30$discourse_class))
```

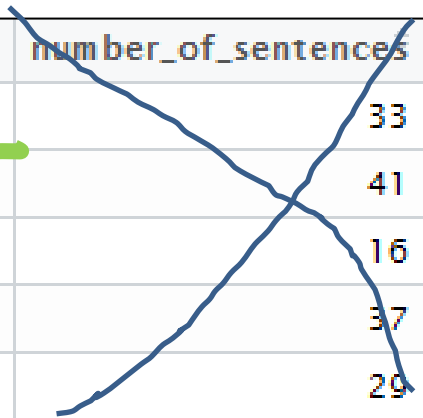
```
##  
##  Pearson's Chi-squared test  
##  
## data:  pdt30$genre and pdt30$discourse_class  
## X-squared = 273.34, df = 42, p-value < 2.2e-16
```

Standardized Residuals

pdt30\$genre	pdt30\$discourse_class			
	CONTINGENCY	CONTRAST	EXPANSION	TEMPORAL
advice	6.4615765	-2.6468625	-1.9676630	-2.4346740
other	-2.0398666	-3.8116713	3.3434255	4.1887851
collection	-0.1020027	-1.6335066	0.6094034	2.1712639
comment	5.2095058	1.2713551	-4.5289432	-2.3526586
description	-0.3769206	-0.7255357	1.8561264	-1.9478011
essay	2.3100952	0.2653760	-1.7707918	-0.9637207
invitation	-3.0571787	-1.6490094	4.6946794	-1.3221224
letter	3.7590890	-1.4095414	-1.1107431	-1.7585667
news	-1.9443364	-0.5155531	1.2665335	1.9083080
overview	-2.6096549	-0.7095047	3.5542873	-1.5436120
person_interv	-0.2116568	0.4405930	-1.8010955	3.5221289
review	-2.7016956	2.5136831	1.2955250	-2.9136409
sport	-5.7410089	3.0270799	0.3371150	3.9337022
survey	0.1514688	-0.8823696	1.2445159	-1.2623203
topic_interv	1.4474858	0.8871888	-1.1151822	-2.0648269

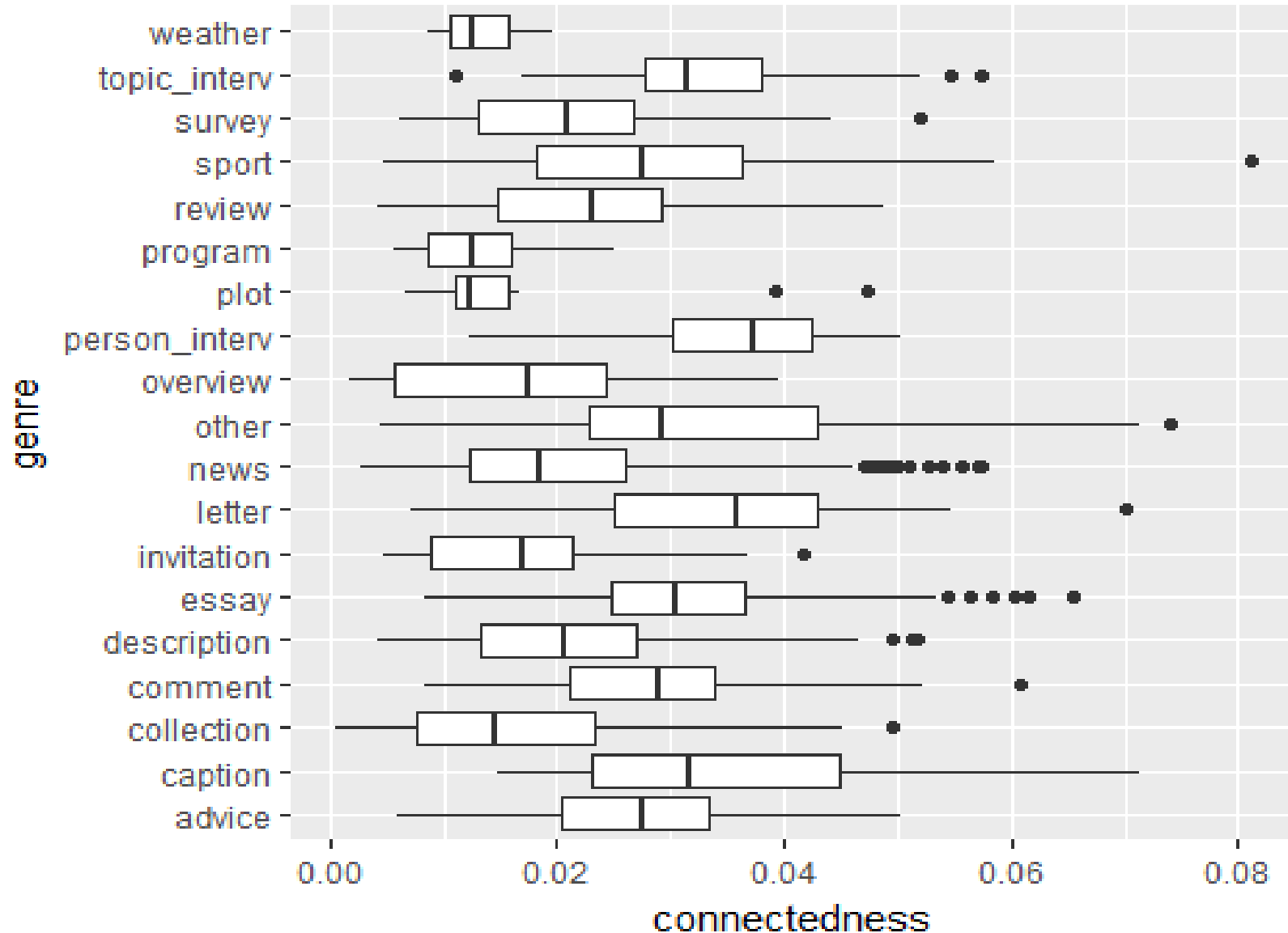
t-test or Wilcoxon-Mann-Whitney-U test to Compare Connectedness ?

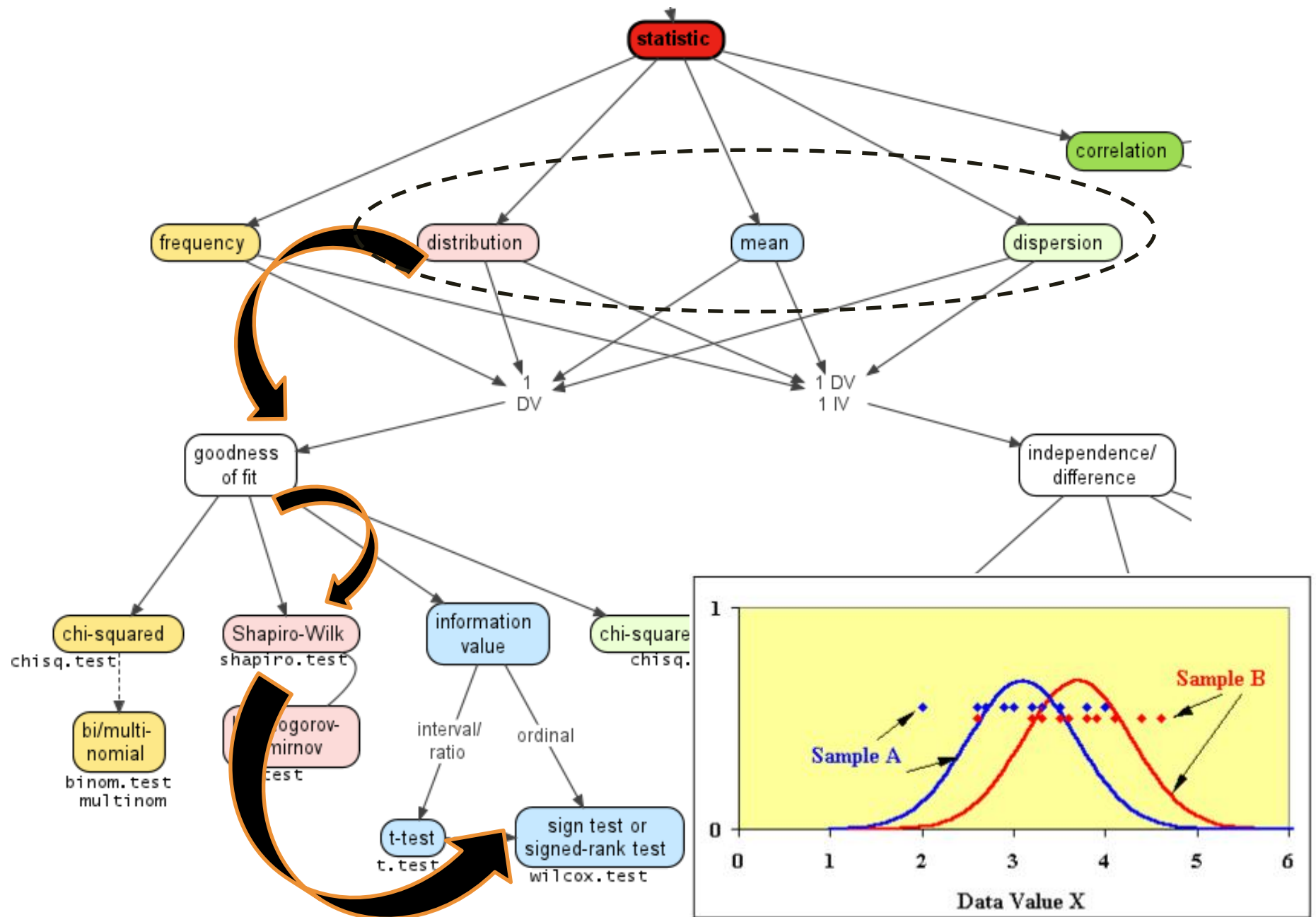
- *document connectedness* (a toy variable):
 - number of connectives/number of words
- Hypothesis H1: Genres differ in document connectedness
- Distribution of a single quantitative variable compared across genres



	document_id	genre	number_of_sentences	word_count	total_connectives	connectedness
1	cmpr9410_001	comment	33	644	12	0.018633540
2	cmpr9410_002	description	41	629	10	0.015898251
3	cmpr9410_003	advice	16	202	5	0.024752475
4	cmpr9410_004	description	37	597	13	0.021775544
5	cmpr9410_005	advice	29	382	16	0.041884817

Distribution of document connectedness in each genre

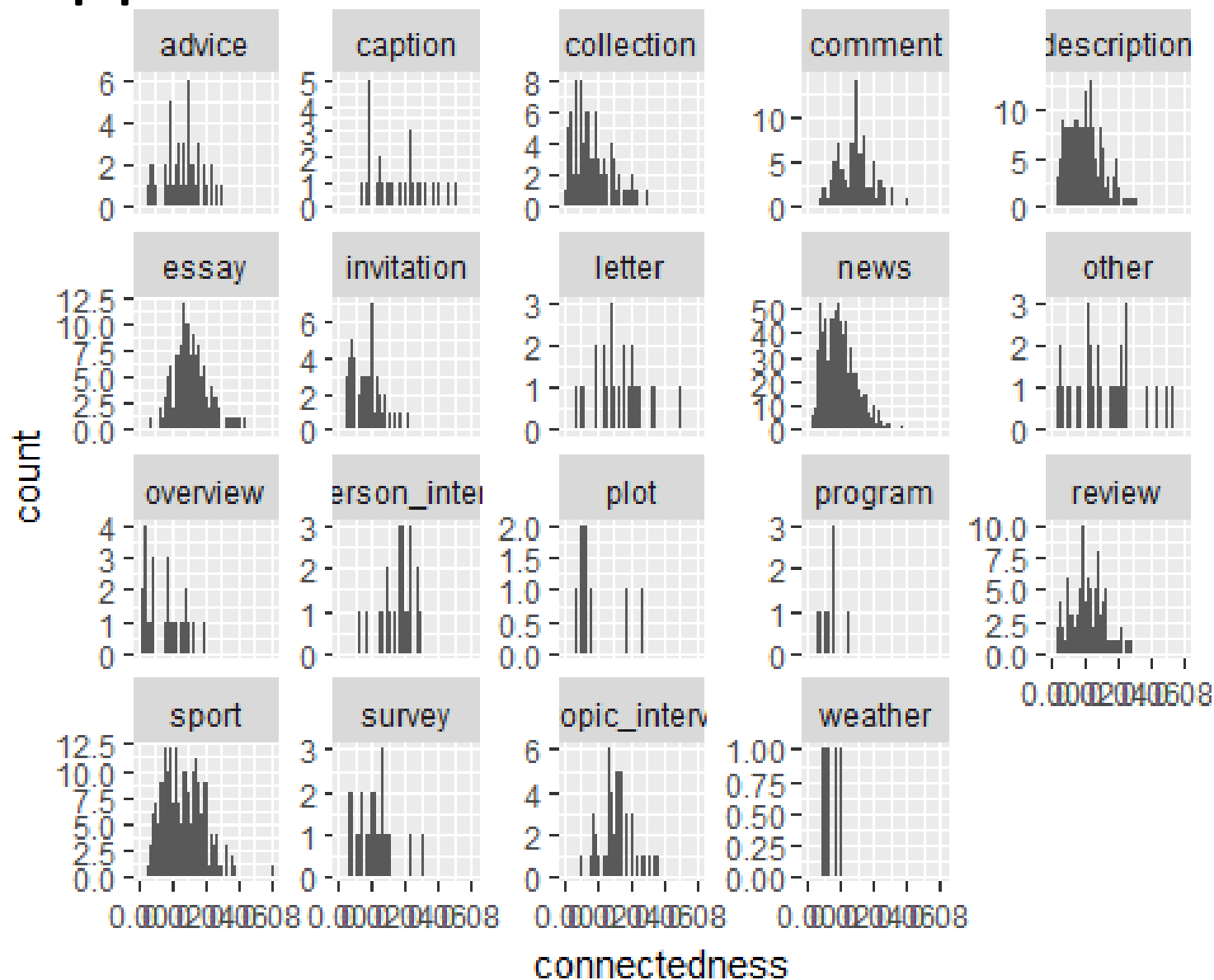




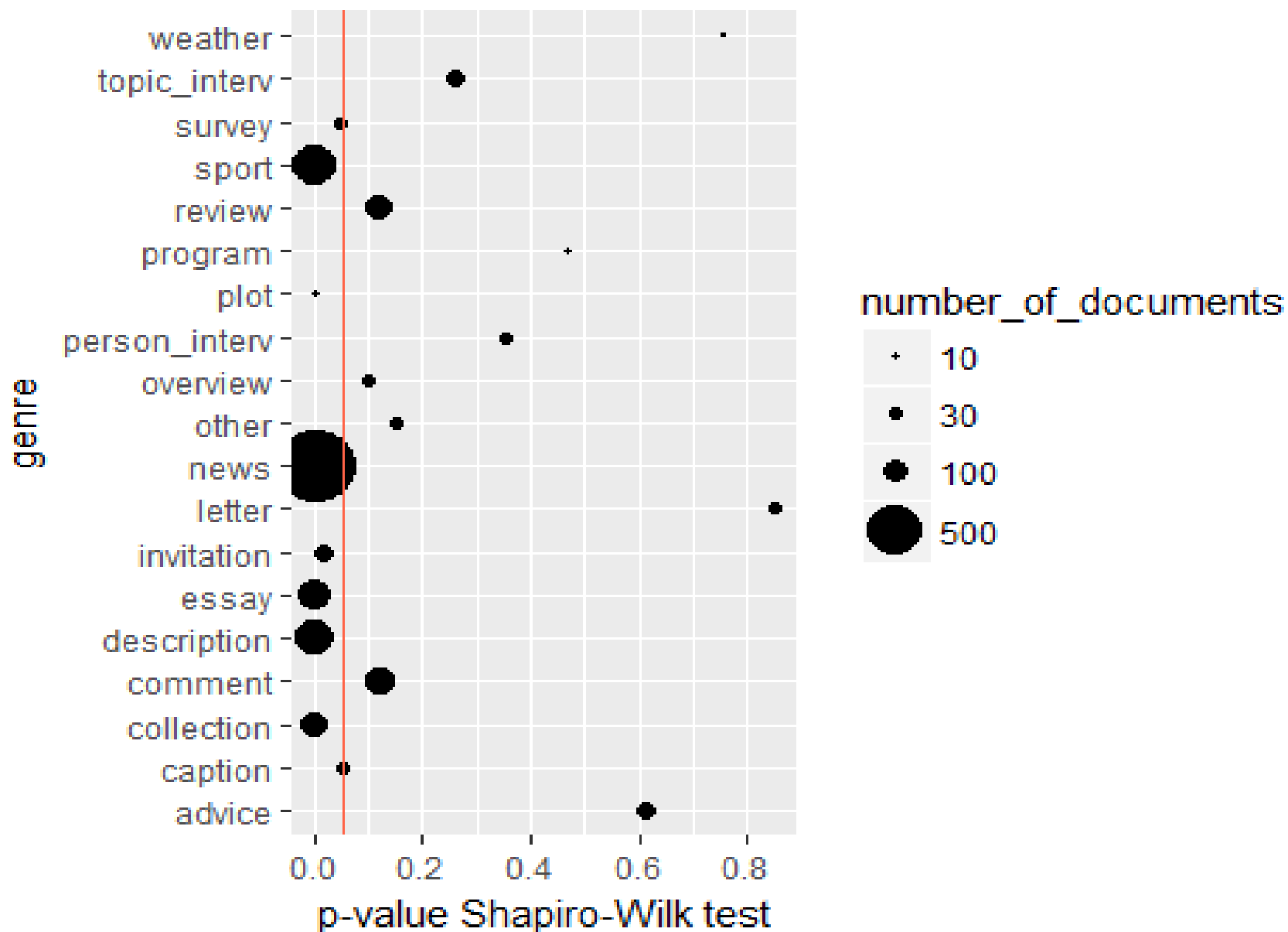
t-test Assumptions

- samples have normal distribution
- each sample contains at least 30 observations

Histograms of Connectedness + Approx. Number of Texts in Genres



Shapiro-Wilk Test + Sample Sizes



Wilcoxon Test Outcomes

```
## advice vs. caption
## List of 9
## $ statistic : Named num 624
## ..- attr(*, "names")= chr "W"
## $ parameter : NULL
## $ p.value : num 0.0514
## $ null.value : Named num 0
## ..- attr(*, "names")= chr "location shift"
## $ alternative: chr "two.sided"
## $ method : chr "Wilcoxon rank sum test with continuity correction"
## $ data.name : chr "x_connectedness$connectedness and y_connectedness$connectedness"
## $ conf.int : atomic [1:2] -0.013205 0.000021
## ..- attr(*, "conf.level")= num 0.95
## $ estimate : Named num -0.00672
## ..- attr(*, "names")= chr "difference in location"
## - attr(*, "class")= chr "htest"
##
## -----
```

Normalized Deviation of Proportions

(Gries 2008, Lijffijt & Gries 2012)

```
norm_DP_word <- (sum(abs(obs_word_prop - exp_word_prop))/2)/(1-  
min(exp_word_prop)).
```

A value between 0 and 1 for a given word. The lower, the more evenly is the word spread across the observed groups. Our application:

Which discourse classes evenly spread across genres? Which biased towards some? (The test will not tell to which.)

Our result: all four discourse classes have NDP approx. 0,2. No dramatic bias, no dramatic differences.

Visualization Layers in ggplot2



Useful References

- www.datacamp.com/
 - a high-quality on-line course hub, very affordable for academics
- www.stackoverflow.com
 - questions and answers community portal for many prog. lang.
- <http://cran.r-project.org/web/views/NaturalLanguageProcessing.html>
 - language-oriented R libraries
- <http://r4ds.had.co.nz>
 - preprint of a book about data science with R, the most up-to-date methods
- http://kbroman.org/knitr_knutshell/
 - reporting - how to write RMarkdown documents