(In)formal modifiers

Chi-square and Odds Ratio

Question

- Is there a relation between the `formality' of a text and the formality of the (gradable) modifiers of a certain adjective?

Corpora

- Formal: Written Dutch
  - Twente Nieuws Corpus
- Informal: Spoken Dutch
  - Corpus Gesproken Nederlands

Data

- Select an adjective
  - Sufficiently present in both corpora
  - “Leuk(e)” and “interressant(e)”
- Select only those cases of the adjectives that are proceeded by a gradable modifier
- Judge wether the modifier is formal, informal or neutral

Data

106	heel	leuke	N	0	0	tw
125	geen	leuke	N	0	0	tw
74	zo'n	leuke	N	0	0	tw
73	hele	leuke	N	0	0	tw
71	minder	leuke	N	0	0	tw
63	erg	leuke	N	0	0	tw
45	wel	leuke	N	0	0	tw
42	hartstikke	leuke	I	0	1	tw
27	zulke	leuke	N	0	0	tw
26	ontzettend	leuke	N	0	0	tw
18	echt	leuke	N	0	0	tw
10	zo	leuke	N	0	0	tw
9	minst	leuke	N	0	0	tw
9	geweldig	leuke	N	0	0	tw
7	niet	leuke	N	0	0	tw
6	vreselijk	leuke	I	0	1	tw
6	best	leuke	I	0	1	tw
5	verrassend	leuke	F	1	0	tw
7	te	leuke	N	0	0	tw

- Chi-square
- Odds Ratio

Pearson's Chi-square test

- If a pair of categorical variables are related.

- H0: The formality of the modifiers is distributed similarly over the different corpora.
- H1: The formality of the modifiers is not distributed similarly over the different corpora.

Chi-square test

- Observed values compared with expected values

Modifiers	Formal	Informal	Neutral	Total
GN corpus	A	B	C	A+B+C
TW corpus	D	E	F	D+E+F
Total	A+D	B+E	C+F	N

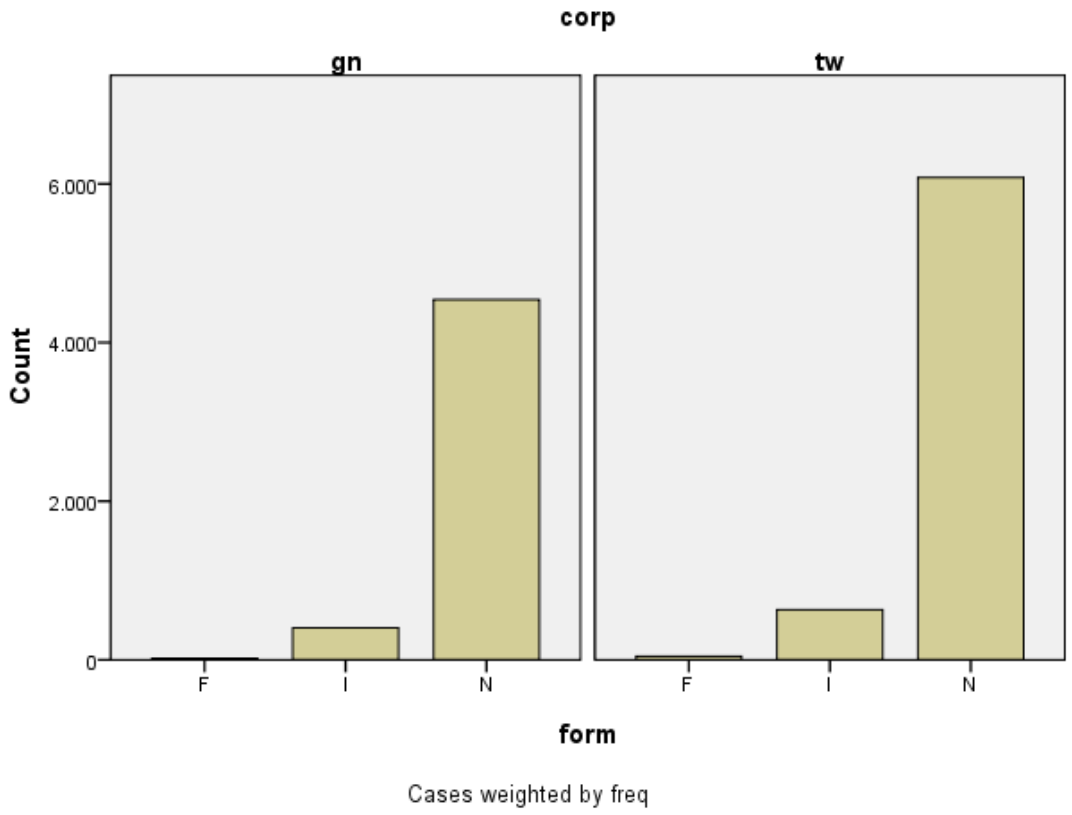
- Expected cell frequency= row total * column total / N

Chi-square test

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- Df = (nRows -1)(nColums-1)
- Observed value > Exptected value → Effects
- Observed value < Expected value → No effects

Leuk(e)


corp
gn
tw
Count
6.000
4.000
2.000
0
F
I
N
F
I
N
form
Cases weighted by freq

Chi-square test: Leuk(e)

corp * form Crosstabulation

Count

		form			
		F	I	N	Total
corp	gn	20	403	4540	4963
	tw	46	630	6079	6755
	Total	66	1033	10619	11718

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9,344[a]	2	,009
Likelihood Ratio	9,527	2	,009
N of Valid Cases	11718		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 27,95.

$p<0.05$ → Significant → H0 rejected

Odds ratio

- Statistics to assess the risk of a particular outcome if a certain factor is present
- Medical reports
- Way of presenting probabilities
- Not the same as relative risk
- 2x2

Odds ratio

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1},$$

Modifiers	Formal	Informal/ neutral	Total
GN corpus	A	B	A+B
TW corpus	C	D	C+D
Total	A+c	B+D	N

p(A)/p(B) /

p(C)/p(D)

Odds ratio

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1},$$

Modifiers	Informal	Formal/neutral	Total
GN corpus	A	B	A+B
TW corpus	C	D	C+D
Total	A+c	B+D	N

p(A)/p(B) /

p(C)/p(D)

Odds ratio: leuk(e)

formal * corp Crosstabulation

Count

		corp		
		gn	tw	Total
formal	non-formal	4943	6709	11652
	formal	20	46	66
	Total	4963	6755	11718

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for formal (non-formal / formal)	1,695	1,001	2,868
For cohort corp = gn	1,400	,970	2,020
For cohort corp = tw	,826	,704	,969
N of Valid Cases	11718		

Odds ratio: leuk(e)

informal * corp Crosstabulation

Count

		corp		
		gn	tw	Total
informal	non-informal	4560	6125	10685
	informal	403	630	1033
	Total	4963	6755	11718

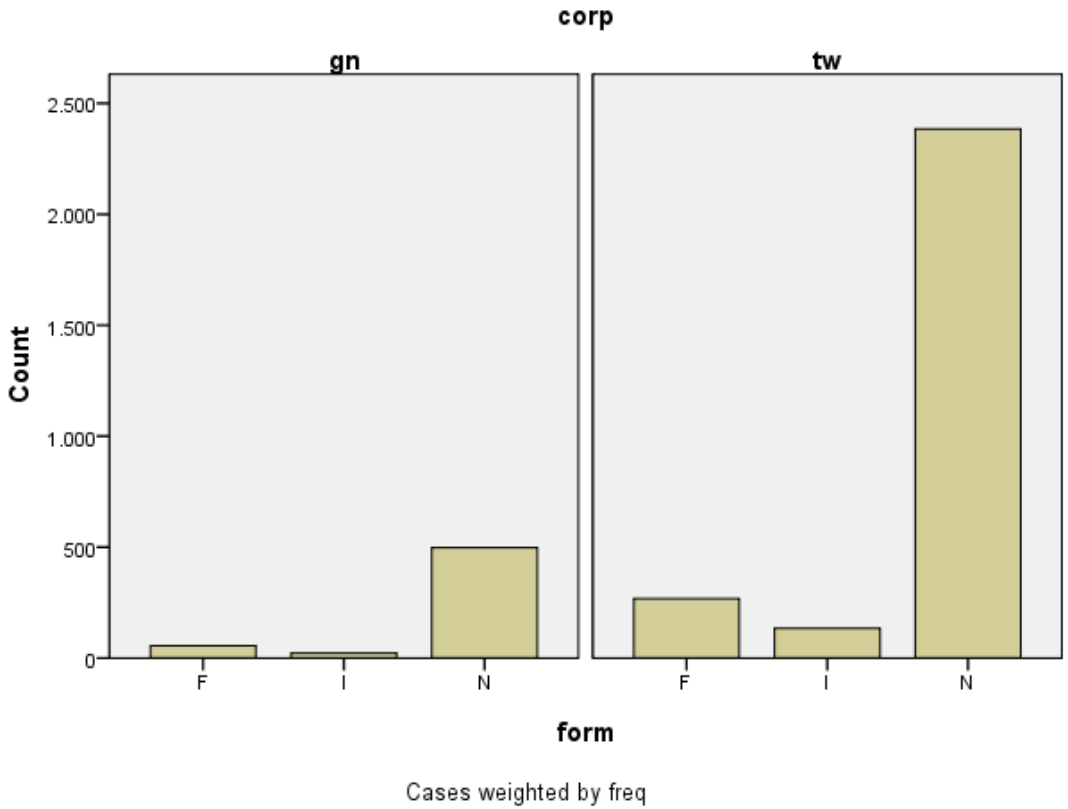
Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for informal (non-informal / informal)	1,164	1,021	1,326
For cohort corp = gn	1,094	1,010	1,184
For cohort corp = tw	,940	,893	,990
N of Valid Cases	11718		

Result: leuk(e)

- Modifiers of the written corpus are more likely to be formal or informal (instead of neutral) than the ones from the spoken corpus.

Interessant(e)



Interessant(e)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,033[a]	2	,597
Likelihood Ratio	1,082	2	,582
N of Valid Cases	3361		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 26,64.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for formal (non-informal / formal)	1,004	,740	1,362
For cohort corp = gn	1,003	,779	1,292
For cohort corp = tw	,999	,949	1,053
N of Valid Cases	3361		

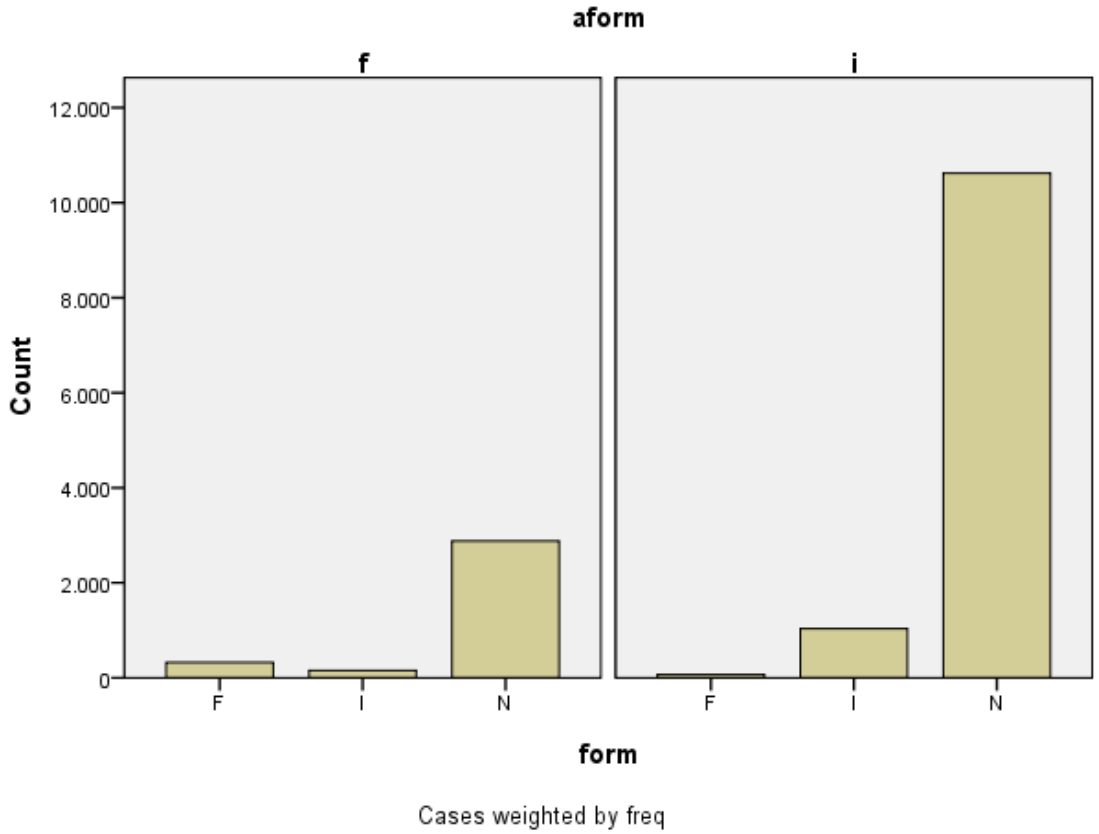
Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for informal (non-informal / informal)	1,267	,800	2,008
For cohort corp = gn	1,221	,823	1,812
For cohort corp = tw	,964	,903	1,029
N of Valid Cases	3361		

Results

- Only modifiers that refer to ‘leuk(e)’ have a relation with the sort of text they are found in.
- Is there a relation between the nature (formality) of the adjective and the formality of its modifiers?

Leuk(e)/interessant(e)



Leuk(e)/interessant(e)

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for formal (non-formal / formal)	,053	,041	,070
For cohort aform = f	,249	,236	,263
For cohort aform = i	4,675	3,752	5,825
N of Valid Cases	15079		

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8,933E2	2	,000
Likelihood Ratio	720,787	2	,000
N of Valid Cases	15079		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 86,71.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for informal (non-informal / informal)	1,986	1,671	2,361
For cohort aform = f	1,759	1,515	2,042
For cohort aform = i	,885	,865	,907
N of Valid Cases	15079		

Conclusion

- The formality of modifiers that refer to ‘leuk(e)’ can be related to the sort of text (written/spoken, formal/informal) they are found in.
- But the formality of both modifiers is related to the type of adjective they refer to (formal/informal).
  - The formal adjective (interessant(e)) is more likely to have a formal modifier.
  - The informal adjective (leuk(e)) is more likely to have an informal modifier.

Further

- I could have made errors or inconsistency in collecting and constructing the data. Certain words can sometimes function as modifier and other times not.
- For `leuk(e)' and `interessant(e)' results are like this. Other adjectives might have another outcome.
- Perhaps I need to consider also the odds ratio of formal vs informal modifiers instead of formal vs informal+neutral and informal vs formal+neutral