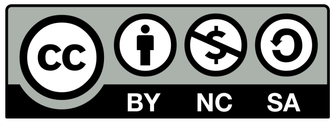


Statistik untuk Linguistik Korpus

Gede Primahadi Wijaya RAJEG
Universitas Udayana, Bali
https://orcid.org/0000-0002-2047-8621
@PrimahadiWijaya

5 November 2020

Alur presentasi

- Pengantar
  - LKorp sebagai bidang ilmu distribusional
- Pemahaman kuantitatif dasar untuk linguis korpus
  - Uji signifikansi dengan chi-square
  - Praktik

Pengantar

- Linguistik korpus pada hakikatnya adalah bidang ilmu distribusional
- Korpus tidak mengandung apapun (mis. makna/fungsi) selain data frekuensi distribusional (Gries 2010: 269):
  - Frekuensi kemunculan (*freq. of occurrence*) unsur linguistik
  - Frekuensi kemunculan bersama (*freq. of co-occurrence*) unsur linguistik

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Pengantar

Frekuensi kemunculan

- Seberapa sering (i.e. berapa kali) suatu morfem atau kata atau konstruksi grammatikal muncul di dalam suatu korpus?
  - Pertanyaan ini bisa dijawab dengan data daftar frekuensi
- Seberapa merata (*how evenly*) suatu morfem atau kata atau konstruksi gramatikal tersebar di dalam korpus?
  - Diberikan oleh statistik dispersi

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Pengantar

Frekuensi kemunculan bersama (*freq. of co-occurrence*)

- Seberapa sering (i.e. berapa kali) suatu unsur linguistik, mis. morfem atau kata atau konstruksi grammatikal, **muncul bersama dengan** unsur linguistik lain?

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Alur presentasi

- Pengantar ✅
  - LKorp sebagai bidang ilmu distribusional ✅
- Pemahaman kuantitatif dasar untuk linguis korpus
  - Uji signifikansi dengan chi-square
  - Praktik

Pemahaman kuantitatif dasar untuk linguistik korpus

	CxN A	CxN B	Total
Collexeme L	Freq of L in A	Freq of L in B	Total Freq of L
Collexeme ¬L	Freq of ¬L in A	Freq of ¬L in B	Total Freq of ¬L
Total	Freq of CxN A	Freq of CxN B	Total Freq of CxN A & B

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

- Distribusi kata kerja modal *will* dan *shall* di dua ragam korpus bahasa Inggris

	LOB	KOLHAPUR
***Will*+infinitive**	2316	1974
***Shall*+infinitive**	363	363

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

- Distribusi kata kerja modal *will* dan *shall* di dua ragam korpus bahasa Inggris

Logika tabel dua dimensi dan signifikansi statistik akan sangat relevan dengan konsep kolokasi dan koligasi

	LOB	KOLHAPUR
***Will*+infinitive**	2316	1974
***Shall*+infinitive**	363	363

Bagaimana kita menentukan apakah distribusi kata kerja modal secara potensial penting? Dengan kata lain, apakah distribusi ini **signifikan secara statistik**?

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

- ⚠️ Perihal signifikansi statistik **berbeda** dengan perihal adanya sesuatu yang menarik dari suatu distribusi.

	LOB	KOLHAPUR
***Will*+infinitive**	2316	1974
***Shall*+infinitive**	363	363

Signifikansi statistik: secara ekslusif merujuk pada perihal apakah distribusi tersebut kemungkinan merupakan suatu kebetulan

Signifikansi statistik: TIDAK sama sekali menyatakan apakah distribusi tersebut signifikan secara linguistik

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

- ⚠️ Perihal signifikansi statistik **berbeda** dengan perihal adanya sesuatu yang menarik dari suatu distribusi.

	LOB	KOLHAPUR
***Will*+infinitive**	2316	1974
***Shall*+infinitive**	363	363

NAMUN, jika suatu distribusi TIDAK signifikan secara statistik, kemungkinan adanya hal menarik secara linguistik tidak akan muncul

JADI, signifikansi statistik adalah pra-kondisi kemungkinan adanya signifikansi linguistik, namun BUKAN suatu jaminan!

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Frekuensi harapan (*expected frequency*)

Pemahaman kuantitatif dasar untuk linguistik korpus

Frekuensi harapan

- Untuk mengetahui apakah suatu distribusi yang kita amati (**frekuensi riil/pengamatan** [*observed frequency*]) adalah suatu kebetulan, kita perlu menghitung bagaimana distribusi yang kita harapkan atas dasar kebetulan (**frekuensi harapan** [*expected frequency*])

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Frekuensi harapan

- Kita lalu bandingkan frekuensi pengamatan dengan frekuensi harapan
  - Apakah ada perbedaan/selisih yang cukup besar antara frekuensi pengamatan dan frekuensi harapan sehingga kita bisa mengatakan bahwa frekuensi pengamatan itu bukan suatu kebetulan

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR
Will+infinitive	2316	1974
Shall+infinitive	363	363

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316	1974	4290
Shall+infinitive	363	363	726
TOTAL	2679	2337	5016

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 ?	1974 ?	4290
Shall+infinitive	363 ?	363 ?	726
TOTAL	2679	2337	5016

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 (2291.25)	1974 ?	4290
Shall+infinitive	363 ?	363 ?	726
TOTAL	2679	2337	5016

(2679 * 4290)/5016

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 (2291.25)	1974 ?	4290
Shall+infinitive	363 (387.75)	363 ?	726
TOTAL	2679	2337	5016

(2679 * 726)/5016

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 (2291.25)	1974 (1998.75)	4290
Shall+infinitive	363 (387.75)	363 ?	726
TOTAL	2679	2337	5016

(2337 * 4290)/5016

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung frekuensi harapan

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 (2291.25)	1974 (1998.75)	4290
Shall+infinitive	363 (387.75)	363 (338.25)	726
TOTAL	2679	2337	5016

(2337 * 726)/5016

Pemahaman kuantitatif dasar untuk linguistik korpus

Membandingkan frekuensi pengamatan vs. frekuensi harapan

	LOB	KOLHAPUR	TOTAL
***Will*+infinitive**	2316 > **(2291.25)**	1974 < **(1998.75)**	4290
***Shall*+infinitive**	363 < **(387.75)**	363 > **(338.25)**	726
TOTAL	2679	2337	5016

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung selisih frekuensi pengamatan dengan frekuensi harapan

	LOB	KOLHAPUR
Will+infinitive	$(Obs-Exp)^2/Exp$ $(2316-2291.25)^2/2291.25$ **0.27**	$(Obs-Exp)^2/Exp$ $(1974-1998.75)^2/1998.75$ **0.31**
Shall+infinitive	$(Obs-Exp)^2/Exp$ $(363-387.75)^2/387.75$ **1.58**	$(Obs-Exp)^2/Exp$ $(363-338.25)^2/338.25$ **1.81**

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis statistik *Chi-Square* (lihat MS Excel sheet)

	LOB	KOLHAPUR
Will+infinitive	(Obs-Exp)²/Exp (2316-2291.25)²/2291.25 **0.27**	(Obs-Exp)²/Exp (1974-1998.75)²/1998.75 **0.31**
Shall+infinitive	(Obs-Exp)²/Exp (363-387.75)²/387.75 **1.58**	(Obs-Exp)²/Exp (363-338.25)²/338.25 **1.81**

Degree of Freedom (df)

X^2 / Nilai Chi-square

p-value

Degree of freedom (df) = 1

(jumlah baris - 1) * (jumlah kolom - 1)
(2 - 1) * (2 - 1)

$X^2 = 3.97$

(0.27 + 0.31 + 1.58 + 1.81)

🤔 🤔 Apakah nilai ini mengindikasikan suatu perbedaan yang sangat besar antara frek. pengamatan dan harapan sehingga distribusi ini bisa dikatakan **signifikan secara statistik**?

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis statistik *Chi-Square* (lihat MS Excel sheet)

	LOB	KOLHAPUR
***Will*+infinitive**	(Obs-Exp)²/Exp (2316-2291.25)²/2291.25 **0.27**	(Obs-Exp)²/Exp (1974-1998.75)²/1998.75 **0.31**
***Shall*+infinitive**	(Obs-Exp)²/Exp (363-387.75)²/387.75 **1.58**	(Obs-Exp)²/Exp (363-338.25)²/338.25 **1.81**

Degree of Freedom (df)

X^2 / Nilai Chi-square

p-value

Degree of freedom (df) = 1

(jumlah baris - 1) * (jumlah kolom - 1)
(2 - 1) * (2 - 1)

X^2 = 3.97

(0.27 + 0.31 + 1.58 + 1.81)

Peranti statistik akan mengubah nilai ini menjadi nilai **probabilitas kesalahan** (***probability of error***) a.k.a ***p*-value**

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis statistik *Chi-Square* (lihat MS Excel sheet)

	LOB	KOLHAPUR
***Will*+infinitive**	(Obs-Exp)²/Exp (2316-2291.25)²/2291.25 **0.27**	(Obs-Exp)²/Exp (1974-1998.75)²/1998.75 **0.31**
***Shall*+infinitive**	(Obs-Exp)²/Exp (363-387.75)²/387.75 **1.58**	(Obs-Exp)²/Exp (363-338.25)²/338.25 **1.81**

Degree of Freedom (df)

X^2 / Nilai Chi-square

p-value

Degree of freedom (df) = 1

(jumlah baris - 1) * (jumlah kolom - 1)
(2 - 1) * (2 - 1)

X^2 = 3.97

(0.27 + 0.31 + 1.58 + 1.81)

p-value adalah ambang batas nilai untuk dapat menentukan apakah kita sedang/tidak sedang berhadapan dengan distribusi yang bersifat acak/kebetulan

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis statistik *Chi-Square* (lihat MS Excel sheet)

	LOB	KOLHAPUR
***Will*+infinitive**	(Obs-Exp)²/Exp (2316-2291.25)²/2291.25 **0.27**	(Obs-Exp)²/Exp (1974-1998.75)²/1998.75 **0.31**
***Shall*+infinitive**	(Obs-Exp)²/Exp (363-387.75)²/387.75 **1.58**	(Obs-Exp)²/Exp (363-338.25)²/338.25 **1.81**

Degree of Freedom (df)

X^2 / Nilai Chi-square

p-value

Degree of freedom (df) = 1

(jumlah baris - 1) * (jumlah kolom - 1)
(2 - 1) * (2 - 1)

X^2 = 3.97

(0.27 + 0.31 + 1.58 + 1.81)

Distribusi dikatakan signifikan (i.e. bukan acak) jika probabilitas (i.e. *p*-value) bahwa distribusi tersebut merupakan distribusi acak/kebetulan lebih kecil dari 5% (atau $p < 0.05$)

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Cara lama menentukan signifikansi melalui Tabel nilai Chi-Square

Table 5: *Old-fashioned Table of Chi-Square Values*

df	5.0%	1.0%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
...	...	...	...

Tentukan *degree of freedom* distribusi

Lihat nilai Chi-Square (X^2) di baris yang sesuai dengan nilai *degree of freedom*

Lihat nilai probabilitas yang sesuai dengan nilai Chi-Square (X^2) distribusi

Nilai X^2 distribusi modal adalah 3.97 (yaitu lebih besar dari 3.841)

Jadi *p*-value distribusi modal < 0.05

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis statistik *Chi-Square* (lihat MS Excel sheet)

	LOB	KOLHAPUR
***Will*+infinitive**	(Obs-Exp)²/Exp (2316-2291.25)²/2291.25 **0.27**	(Obs-Exp)²/Exp (1974-1998.75)²/1998.75 **0.31**
***Shall*+infinitive**	(Obs-Exp)²/Exp (363-387.75)²/387.75 **1.58**	(Obs-Exp)²/Exp (363-338.25)²/338.25 **1.81**

Degree of Freedom (df)

X^2 / Nilai Chi-square

p-value

Degree of freedom (df) = 1

(jumlah baris - 1) * (jumlah kolom - 1)
(2 - 1) * (2 - 1)

X^2 = 3.97

(0.27 + 0.31 + 1.58 + 1.81)

p-value < 0.05

Satu nilai lagi yang wajib dilaporkan adalah nilai efek (*effect size*): Cramér's *V*/ϕ (*Phi*) (Gries 2009: 173-175)
Signifikansi BUKANLAH nilai efek

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).
Gries, Stefan Th. 2009. Statistics for linguistics with R: A practical introduction. Berlin: Mouton de Gruyter.

Pemahaman kuantitatif dasar untuk linguistik korpus

Menghitung nilai koefisien korelasi/nilai efek (Cramér's *V*) (lihat MS Excel sheet)

$$\text{Cramér's V} = \sqrt{\frac{\chi^2}{N \cdot \left(min[N_{baris}, N_{kolom}] - 1\right)}} = \sqrt{\frac{3.96}{5016 \cdot (min[2,2] - 1)}} = 0.028$$

Rentang dan besaran nilai efek:
$0.1 \leq \phi < 0.3$ = efek kecil
$0.3 \leq \phi < 0.5$ = efek sedang
$\phi \geq 0.5$ = efek kuat/besar
(Levshina 2015: 209)

Perbedaan distribusi modal mungkin tidaklah acak/bukan kebetulan ($p < 0.05$) namun hal itu efeknya (mis. secara praktis) sangat kecil (ϕ/Cramér's V = 0.028)

Gries, Stefan Th. 2009. Statistics for linguistics with R: A practical introduction. Berlin: Mouton de Gruyter.
Levshina, Natalia. 2015. How to do Linguistics with R: Data exploration and statistical analysis. John Benjamins Publishing Company. https://doi.org/10.1075/z.195.

Pemahaman kuantitatif dasar untuk linguistik korpus

Melaporkan hasil analisis statistik Chi-Square

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316 > (2291.25)	1974 < (1998.75)	4290
Shall+infinitive	363 < (387.75)	363 > (338.25)	726
TOTAL	2679	2337	5016

Narasi yang bisa dilaporkan dari analisis statistik distribusi modal ini misalnya:

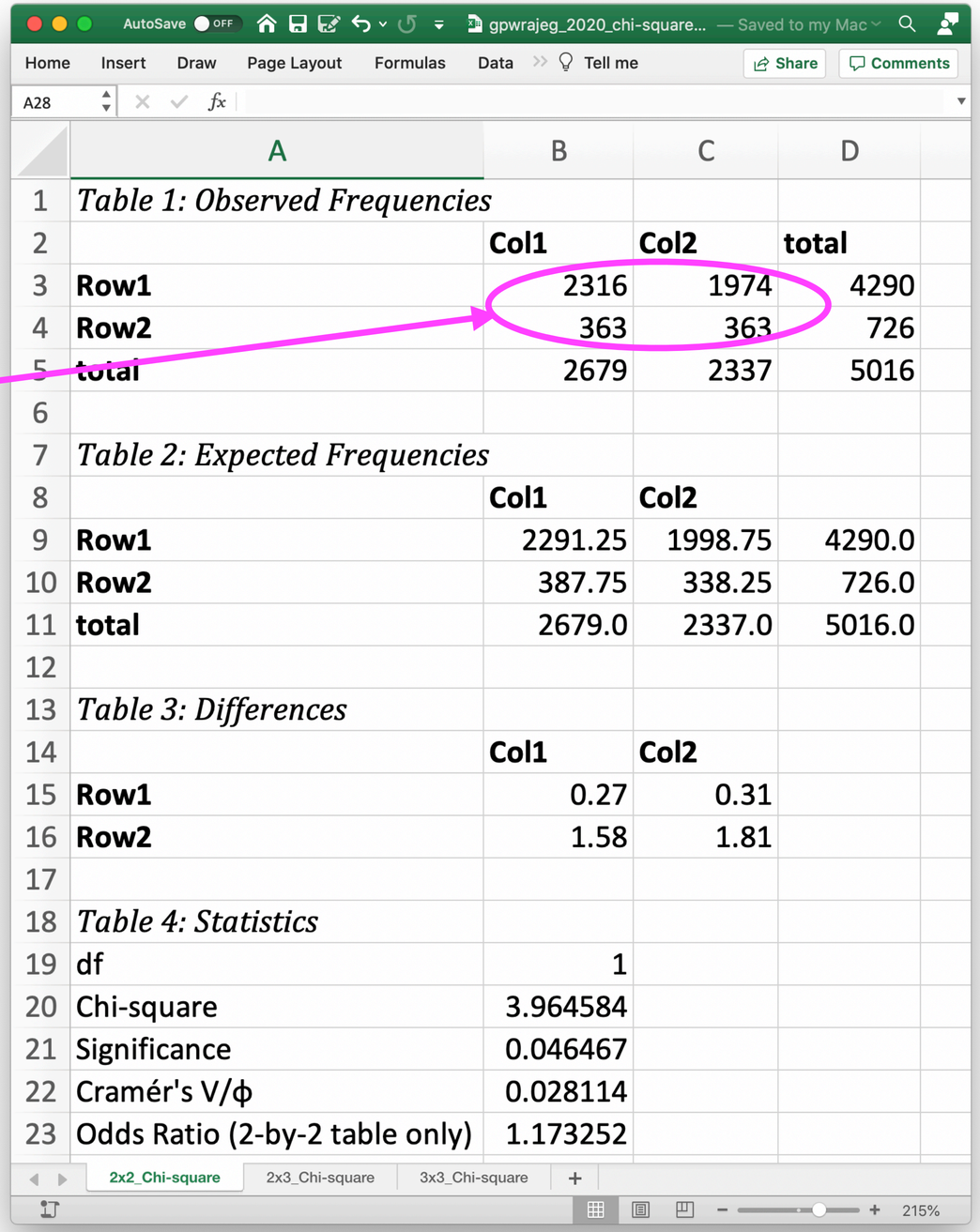
"Kajian ini menemukan bahwa kata kerja modal *shall* secara signifikan lebih sering digunakan (daripada yang diharapkan) di korpus bahasa Inggris ragam India (*Indian English*) daripada di korpus bahasa Inggris ragam British ($X^2 = 3.97$; $df = 1$; $p < 0.05$) namun efek tersebut sangat kecil (ϕ/Cramér's V=0.028)."

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis Chi-Square dengan MS Excel

	LOB	KOLHAPUR	TOTAL
Will+infinitive	2316	1974	4290
Shall+infinitive	363	363	726
TOTAL	2679	2337	5016



Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

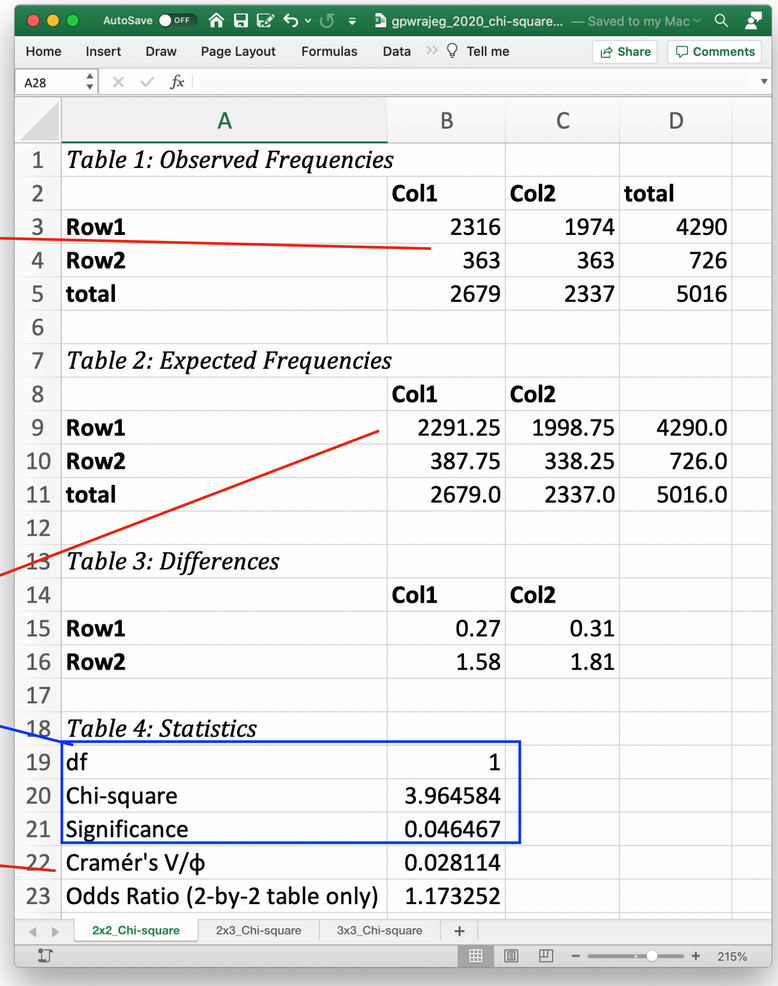
Melakukan analisis Chi-Square di R

```
> # buat tabel distribusi modal di korpus KOLHAPUR dan LOB
> crosstab <- matrix(c(2316, 363, 1974, 363),
+                    nrow = 2,
+                    byrow = FALSE,
+                    dimnames = list(modal = c("will", "shall"),
+                                    corpus = c("LOB", "KOLHAPUR")))
> crosstab
       corpus
modal    LOB KOLHAPUR
  will  2316     1974
  shall  363      363
> # chi-square test dengan fungsi `chisq.test()` dengan tabel `crosstab` sbg input
> chisq.test(crosstab, correct = FALSE)

        Pearson's Chi-squared test

data:  crosstab
X-squared = 3.9646, df = 1, p-value = 0.04647

> # hitung nilai efek (Phi/Cramér's V)
> sqrt(chisq_out$statistic / sum(crosstab) * (min(dim(crosstab)) - 1))
 X-squared
0.02811383
> # lihat frekuensi harapan (expected frequency)
> chisq_out$expected
       corpus
modal       LOB KOLHAPUR
  will  2291.25  1998.75
  shall  387.75   338.25
```



	A	B	C	D
1	*Table 1: Observed Frequencies*			
2		**Col1**	**Col2**	**total**
3	**Row1**	2316	1974	4290
4	**Row2**	363	363	726
5	**total**	2679	2337	5016
6				
7	*Table 2: Expected Frequencies*			
8		**Col1**	**Col2**	
9	**Row1**	2291.25	1998.75	4290.0
10	**Row2**	387.75	338.25	726.0
11	**total**	2679.0	2337.0	5016.0
12				
13	*Table 3: Differences*			
14		**Col1**	**Col2**	
15	**Row1**	0.27	0.31	
16	**Row2**	1.58	1.81	
17				
18	*Table 4: Statistics*			
19	df	1		
20	Chi-square	3.964584		
21	Significance	0.046467		
22	Cramér's V/ϕ	0.028114		
23	Odds Ratio (2-by-2 table only)	1.173252		

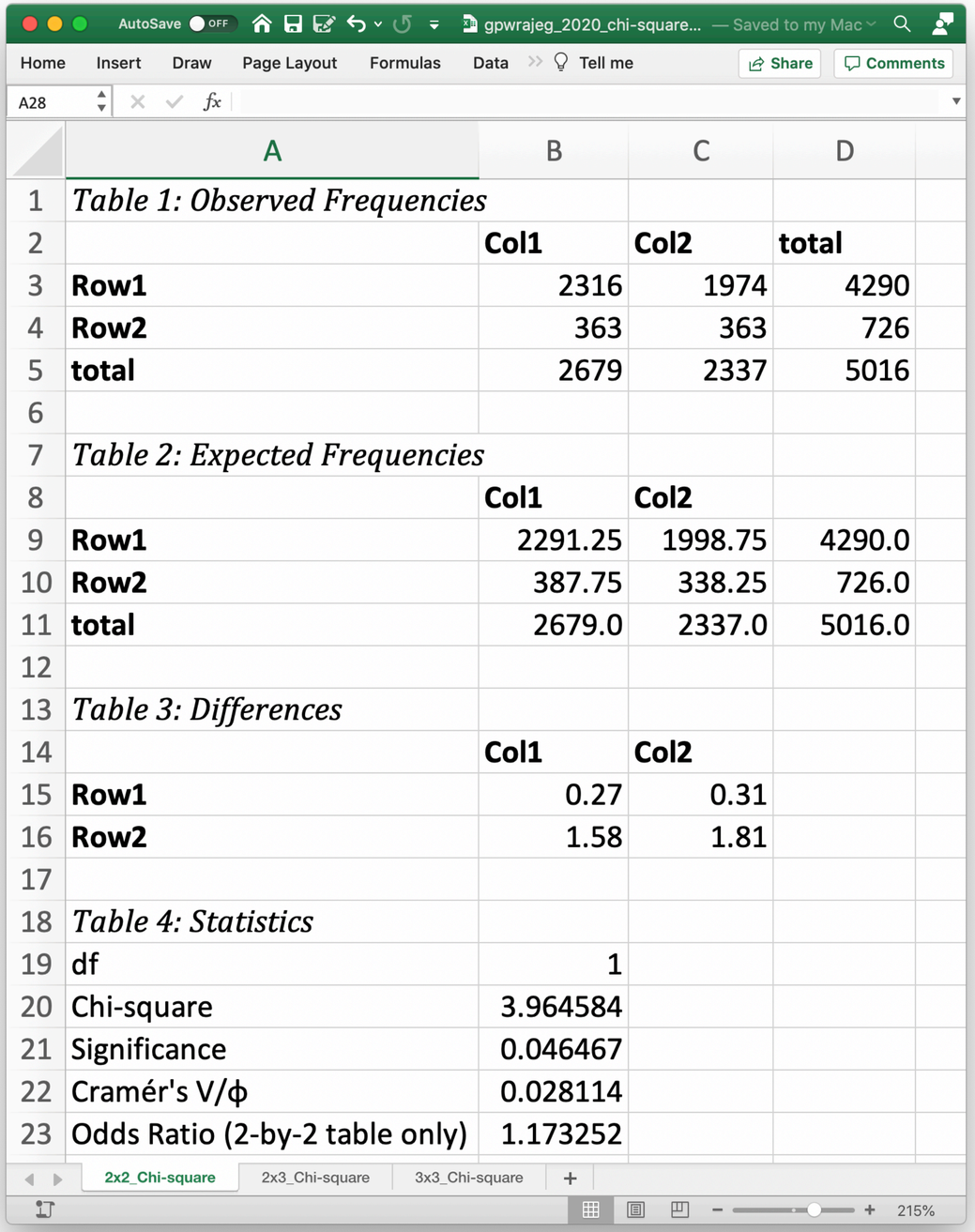
Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Pemahaman kuantitatif dasar untuk linguistik korpus

Melakukan analisis Chi-Square dengan MS Excel

⚠️ Chi-Square dapat digunakan jika 80% Frek. Harapan di tabel lebih besar dari 5

Jika frekuensi di dalam tabel terlalu kecil, bisa gunakan Fisher Exact test (belum terimplementasi di Excel, namun dengan mudah dilakukan di R)



AutoSave OFF gpwrajeg_2020_chi-square... — Saved to my Mac

Home Insert Draw Page Layout Formulas Data Tell me Share Comments

A28

	A	B	C	D
1	*Table 1: Observed Frequencies*			
2		**Col1**	**Col2**	**total**
3	**Row1**	2316	1974	4290
4	**Row2**	363	363	726
5	**total**	2679	2337	5016
6				
7	*Table 2: Expected Frequencies*			
8		**Col1**	**Col2**	
9	**Row1**	2291.25	1998.75	4290.0
10	**Row2**	387.75	338.25	726.0
11	**total**	2679.0	2337.0	5016.0
12				
13	*Table 3: Differences*			
14		**Col1**	**Col2**	
15	**Row1**	0.27	0.31	
16	**Row2**	1.58	1.81	
17				
18	*Table 4: Statistics*			
19	df	1		
20	Chi-square	3.964584		
21	Significance	0.046467		
22	Cramér's V/ϕ	0.028114		
23	Odds Ratio (2-by-2 table only)	1.173252		

2x2_Chi-square 2x3_Chi-square 3x3_Chi-square

215%

Stefanowitsch, Anatol. 2004. Quantitative thinking for corpus linguists. Tutorial. http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html (4 August, 2011).

Tabel 2 dimensi dan untuk analisis statistik kunci di linguistik korpus

Tabel 2 dimensi dan untuk analisis statistik kunci di linguistik korpus

- Kolokat signifikan suatu kata
- Kolokat signifikan khas suatu kata
  - membandingkan dua kata (yang bisa berupa [sin/ant]onim)

Kolokat signifikan

	Kata target **X**	Kata-kata lainnya	TOTAL
Kolokat Potensial **C**	Frekuensi **C** di rentang lingkup kolokasi di sekitar **X**	Frekuensi **C** di lingkup lain di korpus	Total frekuensi **C** di korpus
Kata-kata lainnya	Frekuensi dari kata-kata lainnya di rentang lingkup kolokasi di sekitar **X**	Frekuensi dari kata-kata lain di konteks lain di korpus	Total frekuensi kata-kata lainnya di korpus
TOTAL	Total frekuensi kata target **X** di keseluruhan korpus	Total frekuensi kata-kata lainnya di korpus	Total frekuensi semua kata di korpus (i.e. ukuran korpus)

Kolokat signifikan

	Kata target **X** (***melangkah***)	Kata-kata lainnya	TOTAL
Kolokat Potensial **C** (***maju***)	9	1395-9 = 1386	1395
Kata-kata lainnya	141-9 = 132	5,761,654-132 = 5,761,522	5,763,049-1395 = 5,761,654
TOTAL	141	5,763,049-141 = 5,762,908	5,763,049

Kolokat signifikan

Situasi di mana Chi-Square tidak bisa dilakukan

	Kata target **X** (***melangkah***)	Kata-kata lainnya	TOTAL
Kolokat Potensial **C** (***maju***)	9 (exp: 0.034)	1386 (exp: 1394)	1395
Kata-kata lainnya	132 (exp: 140.965)	5,761,522 (exp: 5,761,513)	5,761,654
TOTAL	141	5,762,908	5,763,049

Kolokat signifikan

```
> chisq.test(matrix(c(9, 132, 1386,5761522), nrow = 2), correct = F)

   Pearson's Chi-squared test

data:  matrix(c(9, 132, 1386, 5761522), nrow = 2)
X-squared = 2355.9, df = 1, p-value < 2.2e-16

Warning message:
In chisq.test(matrix(c(9, 132, 1386, 5761522), nrow = 2), correct = F) :
  Chi-squared approximation may be incorrect
> fisher.test(matrix(c(9, 132, 1386,5761522), nrow = 2), alternative = "greater")

   Fisher's Exact Test for Count Data

data:  matrix(c(9, 132, 1386, 5761522), nrow = 2)
p-value < 2.2e-16
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 144.2053      Inf
sample estimates:
odds ratio
  283.0834
```

Warning untuk Chi-Square

Gunakan
Fisher Exact Test

Kolokat signifikan

	Kata target **X** (***melangkah***)	Kata-kata lainnya	TOTAL
Kolokat Potensial **C** (***maju***)	9 (exp: 0.034)	1386 (exp: 1394)	1395
Kata-kata lainnya	132 (exp: 140.965)	5,761,522 (exp: 5,761,513)	5,761,654
TOTAL	141	5,762,908	5,763,049

Collocation	Word 1	Word 2	Freq Word 1 & Word 2	Freq Word 1	Freq Word 2	Word in Corpus	MI	T-Score	Z-Score
melangkah maju	melangkah	maju	9	141	1395	5763049	8.043	2.989	48.531

$$\text{Pointwise MI} = \log 2 \frac{Frek.Pengamatan}{Frek.Harapan} = \log 2 \frac{9}{0.034} = 8.043$$

$$\text{T-score} = \log 2 \frac{Frek.Pengamatan - Frek.Harapan}{\sqrt{Frek.Harapan}} = \log 2 \frac{9 - 0.034}{\sqrt{0.034}} = 2.989$$

Kolokat signifikan

	Kata target **X** (***melangkah***)	Kata-kata lainnya	TOTAL
Kolokat Potensial **C** (***maju***)	9 (exp: 0.034)	1386 (exp: 1394)	1395
Kata-kata lainnya	132 (exp: 140.965)	5,761,522 (exp: 5,761,513)	5,761,654
TOTAL	141	5,762,908	5,763,049

```
> -log10(fisher.test(matrix(c(9, 132, 1386,5761522), nrow = 2), alternative = "greater")$p.value)
[1] 18.89831
```

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Kolokat signifikan khas

	Kata target **X**	Kata target **Y**	TOTAL
Kolokat Potensial **C**	Frekuensi **C** di rentang lingkup kolokasi di sekitar **X**	Frekuensi **C** di rentang lingkup kolokasi di sekitar **Y**	Total frekuensi **C**
Kata-kata lainnya	Frekuensi dari kata-kata lainnya di rentang lingkup kolokasi di sekitar **X**	Frekuensi dari kata-kata lainnya di rentang lingkup kolokasi di sekitar **Y**	Total frekuensi kata-kata lainnya di korpus
TOTAL	Total frekuensi kata target **X** di rentang lingkup kolokasi	Total frekuensi kata target **Y** di rentang lingkup kolokasi	Total frekuensi semua kata di rentang lingkup kolokasi **X** & **Y**

Kolokat signifikan khas

Situasi di mana Chi-Square tidak bisa dilakukan

	Kata target **X** (***memperbesar***)	Kata target **Y** (***membesarkan***)	TOTAL
Kolokat Potensial **C** (***jumlah***)	9 (exp: 5.77)	1 (exp: 4.23)	10
Kata-kata lainnya	408 (exp: 411.23)	305 (exp: 301.77)	713
TOTAL	417	306	723

Collocation	Word 1	Word 2	Freq Word 1 & Word 2	Freq Word 1	Freq Word 2	Word in Corpus	MI	T-Score	Z-Score
memperbesar jumlah	memperbesar	jumlah	9	417	10	723	0.642	1.077	1.346
mebesarkan jumlah	membesarkan	jumlah	1	306	10	723	-2.081	-3.232	-1.571

Rajeg, Gede Primahadi Wijaya & I Made Rajeg. 2019. Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. In I Nengah Sudipa (ed.), *ETIKA BAHASA Buku persembahan menapaki usia pensiun: I Ketut Tika*, vol. 1, 65–83. Denpasar, Bali, Indonesia: Swasta Nulus. https://doi.org/10.26180/5bf4e49ea1582. https://osf.io/preprints/inarxiv/uwzts/ (30 January, 2019).

Kolokat signifikan khas

```
> chisq.test(matrix(c(9, 408, 1, 305), nrow = 2), correct = F)

   Pearson's Chi-squared test

data:  matrix(c(9, 408, 1, 305), nrow = 2)
X-squared = 4.3402, df = 1, p-value = 0.03722

Warning message:
In chisq.test(matrix(c(9, 408, 1, 305), nrow = 2), correct = F) :
  Chi-squared approximation may be incorrect
> fisher.test(matrix(c(9, 408, 1, 305), nrow = 2), alternative = "greater")

   Fisher's Exact Test for Count Data

data:  matrix(c(9, 408, 1, 305), nrow = 2)
p-value = 0.03306
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.137856      Inf
sample estimates:
odds ratio
  6.715127
```

Warning untuk Chi-Square

Gunakan
Fisher Exact Test

Kolokat signifikan khas

Situasi di mana Chi-Square tidak bisa dilakukan

	Kata target **X** (***memperbesar***)	Kata target **Y** (***membesarkan***)	TOTAL
Kolokat Potensial **C** (***PAN***)	1 (exp: 4.04)	6 (exp: 2.96)	7
Kata-kata lainnya	416 (exp: 412.96)	300 (exp: 303.04)	716
TOTAL	417	306	723

Rajeg, Gede Primahadi Wijaya & I Made Rajeg. 2019. Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. In I Nengah Sudipa (ed.), *ETIKA BAHASA Buku persembahan menapaki usia pensiun: I Ketut Tika*, vol. 1, 65–83. Denpasar, Bali, Indonesia: Swasta Nulus. https://doi.org/10.26180/5bf4e49ea1582. https://osf.io/preprints/inarxiv/uwzts/ (30 January, 2019).

Kolokat signifikan khas

```
> fisher.test(matrix(c(1, 416, 6, 300), nrow = 2), alternative = "less")

   Fisher's Exact Test for Count Data

data:  matrix(c(1, 416, 6, 300), nrow = 2)
p-value = 0.02508
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.7908206
sample estimates:
odds ratio
 0.1205022

>
> log10(fisher.test(matrix(c(1, 416, 6, 300), nrow = 2), alternative = "less")$p.value)
[1] -1.600684
```

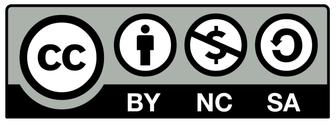
	Kata target X (*memperbesar*)	Kata target Y (*membesarkan*)	TOTAL
Kolokat Potensial C (*PAN*)	1 (exp: 4.04)	6 (exp: 2.96)	7
Kata-kata lainnya	416 (exp: 412.96)	300 (exp: 303.04)	716
TOTAL	417	306	723

Rajeg, Gede Primahadi Wijaya & I Made Rajeg. 2019. Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. In I Nengah Sudipa (ed.), *ETIKA BAHASA Buku persembahan menapaki usia pensiun: I Ketut Tika*, vol. 1, 65–83. Denpasar, Bali, Indonesia: Swasta Nulus. https://doi.org/10.26180/5bf4e49ea1582. https://osf.io/preprints/inarxiv/uwzts/ (30 January, 2019).

Alur presentasi

- Pengantar ✅
  - LKorp sebagai bidang ilmu distribusional ✅
- Pemahaman kuantitatif dasar untuk linguis korpus ✅
  - Uji signifikansi dengan chi-square ✅
  - Praktik

Latihan dengan MS Excel sheet untuk Chi-Square



Statistik untuk Linguistik Korpus

Gede Primahadi Wijaya RAJEG
Universitas Udayana, Bali
https://orcid.org/0000-0002-2047-8621
@PrimahadiWijaya

5 November 2020