Pemahaman kuantitatif dasar untuk kajian linguistik korpus kuantitatif

1, 2 Gede Primahadi Wijaya Rajeg & 1 I Made Rajeg

1 Universitas Udayana, Indonesia & 2 Monash University, Australia

[primahadiwijaya@gmail.com](mailto:primahadiwijaya@gmail.com) & [made\_rajeg@unud.ac.id](mailto:made_rajeg@unud.ac.id)

Abstrak

Makalah ini membahas beberapa unsur mendasar dari pemahaman kuantitatif untuk kajian kebahasaan. …

# Pendahuluan

# Data

Korpus yang menjadi sumber data makalah ini adalah *Indonesian Web as Corpus* (IWaC) (109.281.359 juta kata) (Kilgarriff et al., [2014](#ref-kilgarriff_2014_sketch)). IWaC tersedia daring melalui layanan korpus berbayar *Sketch Engine* (SE)[[1]](#footnote-1). Penulis mencari 100 sampel sitiran acak pemakaian kata-kata dengan akar kata *panas*, yaitu *panas* itu sendiri, kemudian bentuk turunannya sebagai kata kerja, utamanya *memanas*, *memanaskan*, *dipanaskan*, *dipanasi* dan *memanasi*[[2]](#footnote-2); kata-kata tersebut selanjutnya akan disebut medan leksikal panas. Pencarian dilakukan dalam rentang waktu percobaan gratis untuk akun SE selama 30 hari di pertengahan tahun 2013. Data sitiran kata-kata tersebut dan berkas *R Markdown*, yang mengandung kode pemrograman R untuk menghasilkan analisis kuantitatif, berpadu dengan narasi dari makalah ini, akan diterbitkan daring dan terbuka di laman *figshare* penulis.

Tabel [1](#conc) menampilkan lima sitiran acak dari total 458 sitiran konkordansi medan leksikal panas, dengan tampilan konteks di sebelah kiri (*left*) dan kanan (*right*) dari kata kunci (*node*).

Tabel 1 Lima sitiran konkordansi acak medan leksikal panas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | left | node | right | use |
| 7173 | sitas ( kegemukan ) akibat gesekan lapisan lemak yang | memanasi | testis , dan kelompok pria yang dalam pekerjaannya duduk ter | lit |
| 3197 | ransel yang ditingkatkan . Dikarenakan | panas | yang dihasilkan dari perlengkapan RGM -79 SC | lit |
| 74310 | n-bahan lain yang dapat terbakar untuk | memanaskan | air . Pemanasan air ini juga dapat ditempuh denga | lit |
| 23156 | Bambang Trihatmojo dan Halimah kembali | memanas | . Pascapertengkarannya beberapa waktu lalu , tern | met |
| 53466 | dengan baik sebagai satu senjata untuk | memanaskan | rapat dengan pidato yang hebat . Pidato bertu | met |

Kolom *id* merupakan penanda sitiran yang dikeluarkan dari pencarian konkordansi di SE. Kemudian, kolom *use* merupakan variabel yang menandai apakah baris sitiran untuk suatu kata kunci menunjukkan pemakaian “lit(eral)” atau “met(aforis)”. Peneliti menemukenali kedua pemakaian tersebut secara manual dengan melihat konteks pemakaian kata kunci di masing-masing sitiran berdasarkan *Metaphor Identification Procedure* (MIP) (Pragglejaz Group, [2007](#ref-pragglejaz_group_mip:_2007)), sembari menandainya melalui *Excel*. Pemakaian literal kata-kata tersebut secara umum mengacu pada pemakaiannya pada ranah suhu yang bersifat badaniah (periksa sitiran tiga baris pertama di Tabel [1](#conc)). Sedangkan pemakaian metaforisnya menunjukkan pemetaan unsur suhu badaniah ke ranah yang lebih abstrak, seperti emosi atau intensitas suatu keadaan (seperti pada dua baris terakhir di Tabel [1](#conc)). Sitiran ganda (*duplicates*) secara manual ditandai dan tidak diikutkan dalam analisis kuantitatif selanjutnya.

# Pembahasan

Bagian ini meliputi tiga hal terkait analisis kuantitatif. Hal pertama mengacu pada analisis statistik yang bersifat deskriptif (§[3.1](#deskripsi)). Hal kedua mencakup pemahaman terhadap analisis statistik yang bersifat analitis, yang meliputi signifikansi statistik dan hipotesa ilmiah (§[3.2](#signifikansi)), serta uji signifikansi statistik terhadap suatu distribusi menggunakan *chi-square* (§[3.3](#chisquare)). Hal terakhir adalah memperkenalkan cara melakukan uji *chi-square* menggunakan bahasa pemrograman R (R Core Team, [2018](#ref-rcore_2018)) (§[3.4](#chisqr)).

## Memahami data melalui tabel dan grafik

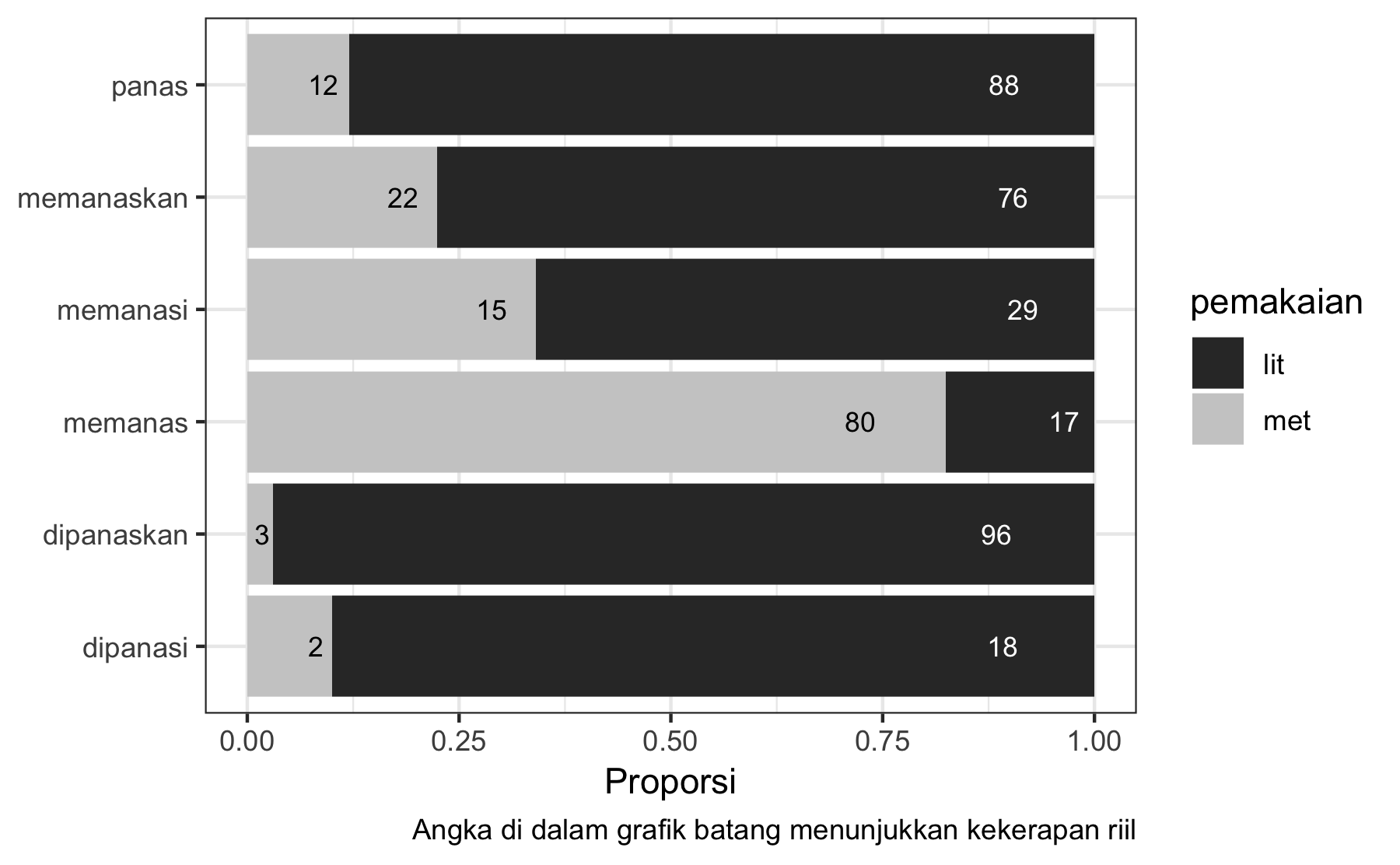
Tabel 2 Kekerapan riil antara kategori pemakaian (baris) dan kata (kolom)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | dipanasi | dipanaskan | memanas | memanasi | memanaskan | panas | total |
| lit | 18 | 96 | 17 | 29 | 76 | 88 | 324 |
| met | 2 | 3 | 80 | 15 | 22 | 12 | 134 |
| total | 20 | 99 | 97 | 44 | 98 | 100 | 458 |

Tabel 3 Kekerapan relatif/proporsi pemakaian literal dan metaforis pada medan leksikal panas

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | dipanasi | dipanaskan | memanas | memanasi | memanaskan | panas |
| lit | 0.9 | 0.97 | 0.18 | 0.66 | 0.78 | 0.88 |
| met | 0.1 | 0.03 | 0.82 | 0.34 | 0.22 | 0.12 |

Kini dapat diperhatikan bahwa proporsi pemakaian literal dari total sitiran *dipanasi* relatif lebih tinggi (yaitu 0.9 atau 90%) dibandingkan dengan *memanasi* (sekitar 0.66 atau 65.91%). Informasi pada Tabel [2](#cross-tab) dan Tabel [3](#prop-table) dapat dipahami secara lebih intuitif melalui tampilan visual seperti Gambar [1](#barplot)[[3]](#footnote-3).



Gambar 1 Proporsi sitiran literal dan metaforis kata-kata pada medan leksikal panas

## Signifikansi statistik dan hipotesa ilmiah

* H0 (hipotesa kosong): Rasio pemakaian literal dan metaforis **tidak berbeda** sehubungan dengan beragam bentuk morfosintaksis kata-kata medan leksikal panas. Artinya, bentuk intransitif *memanas*, transitif *memanaskan* dan bentuk pasif-nya *dipanaskan* seharusnya memiliki distribusi pemakaian literal dan metaforis yang setara dalam sampel.
* H1 (hipotesa alternatif): Rasio pemakaian literal dan metaforis **berbeda** sehubungan dengan beragam bentuk morfosintaksis kata-kata di medan leksikal panas tersebut. Dengan kata lain, diasumsikan bahwa terdapat kaitan antara bentuk morfosintaksis kata di medan leksikal panas dengan proporsi makna literal dan metaforis terkait pemakaian kata tersebut dalam korpus.

Hal yang tidak diperinci dari H1 terkait perbedaan distribusi ini adalah **arah** perbedaannya. Yaitu, apakah misalnya *memanaskan*, dibandingkan dengan *memanasi*, lebih sering (atau lebih jarang) digunakan secara metaforis (dibandingkan secara literal) dalam sampel korpus. Inti yang disampaikan oleh H1 adalah adanya perbedaan distribusi. Arah dari perbedaan distribusi tersebut tidak dapat diperinci untuk saat ini karena belum ditemukannya kajian sebelumnya di Bahasa Indonesia yang melihat kaitan antara perbedaan bentuk kata berakar sama dengan makna literal dan/atau metaforis yang disampaikan dalam pemakaian kata tersebut di sampel korpus. Hasil analisis dalam makalah ini akan memberikan asumsi awal terhadap perbedaan distribusi tersebut yang selanjutnya dapat diujikan kembali menggunakan sampel yang lebih besar dan jenis korpus yang berbeda. Selanjutnya, tujuan utama dari uji signifikansi statistik adalah untuk tidak mencoba membuktikan kebenaran H1, melainkan mencoba menggugurkan asumsi yang diajukan oleh H0, yaitu bahwa tidak terdapat perbedaan distribusi, asosiasi, dsb. (periksa Gries, [2013](#ref-gries_statistics_2013), hlm. 26-29 untuk ulasan lebih dalam).

Terdapat sejumlah metode uji signifikansi statistik. Yang umum digunakan untuk menguji signifikansi distribusi variabel nominal/kategorikal[[4]](#footnote-4) seperti pada Tabel [2](#cross-tab) adalah *Chi-Square test*[[5]](#footnote-5), yang dilambangkan dengan (Gries, [2013](#ref-gries_statistics_2013), hlm. 178-189; Janda, [2013a](#ref-janda_quantitative_2013), hlm. 9-14; Levshina, [2015](#ref-levshina_how_2015), Bab 9; Stefanowitsch, [2004](#ref-stefanowitsch_quantitative_2004)). §[3.3](#chisquare) akan mengulas tahapan penting dalam melakukan uji signifikansi dengan .

## Uji signifikansi Chi-Square

### Kekerapan Harapan dan Kekerapan Riil

:

menunjukkan kekerapan marjinal total (*marginal frequency*) baris , menunjukkan kekerapan total kolom , dan adalah total observasi atau sampel; jadi, Fe masing-masing sel ditentukan berdasarkan kekerapan total masing-masing kolom dan baris (ditandai dengan huruf miring pada Tabel [4](#observed) berikut, yang merupakan pengulangan Tabel [2](#cross-tab)).

Tabel 4 Tabulasi kekerapan antara kategori pemakaian (baris) dan kata (kolom)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | dipanasi | dipanaskan | memanas | memanasi | memanaskan | panas | total |
| lit | 18 | 96 | 17 | 29 | 76 | 88 | *324* |
| met | 2 | 3 | 80 | 15 | 22 | 12 | *134* |
| total | *20* | *99* | *97* | *44* | *98* | *100* | *458* |

Sebagai contoh, penghitungan Fe untuk sel dari persimpangan baris *lit* dan kolom *dipanasi* adalah , yaitu 14.1484716. Satu contoh lagi, Fe untuk sel dari persimpangan *met* dan *memanas* dihitung dengan cara yang sama, yaitu , yang menghasilkan nilai 28.3799127.

Tabel 5 Distribusi yang diharapkan atas dasar H0 antara kategori pemakaian (baris) dan kata (kolom)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | dipanasi | dipanaskan | memanas | memanasi | memanaskan | panas |
| lit | 14.148472 | 70.03493 | 68.62009 | 31.12664 | 69.32751 | 70.74236 |
| met | 5.851528 | 28.96507 | 28.37991 | 12.87336 | 28.67249 | 29.25764 |

### Kontribusi sel terhadap nilai statistik dan tingkat signifikansi

:

.

Rumus mengukur nilai *df*, dan nilai *df* untuk tabel distribusi pada makalah ini, adalah sebagai berikut:

Nilai *df* = 5, dengan kekerapan marjinal total yang tetap, menunjukkan bahwa hanya 5 sel dari Tabel [4](#observed) yang nilainya dapat diubah tanpa mengubah jumlah kekerapan marjinalnya (bdk. Levshina, [2015](#ref-levshina_how_2015), hlm. 12)..

## Uji signifikansi Chi-Square melalui R

* memperoleh nilai *Pearson Residuals*

# keluarkan tabel nilai residual  
panas\_chi$residuals

##   
## dipanasi dipanaskan memanas memanasi memanaskan panas  
## lit 1.0239490 3.1026448 -6.2315072 -0.3811776 0.8013737 2.0518315  
## met -1.5922030 -4.8244984 9.6897642 0.5927171 -1.2461065 -3.1905224

# Penutup

* Implikasi yang menunjukkan bahwa terdapat kaitan antara perbedaan bentuk morfosintaksis suatu kata (yang dalam hal ini berakar kata *panas*) dengan kecenderungan makna yang disampaikannya ketika kata tersebut digunakan di dalam suatu ujaran. Kaitan ini hanya dapat diangkat dan diukur keberadannya dengan menerapkan kajian berbasis korpus dan menganalisis data kuantitatif tersebut menggunakan statistik.
* Makalah ini sedikitknya mencontohkan bagaimana suatu fenomena kebahasaan, yang secara intuitif mungkin sulit diterangkan atau bahkan belum disadari sebelumnya, dapat diungkap secara lebih empiris menggunakan metode korpus kuantitatif. Melalui makalah ini pula penulis berharap bahwa pembaca dapat melihat sekelumit nilai penting yang ditawarkan oleh metode kuantitatif terhadap fenomena humanistik, seperti halnya bahasa, khususnya makna, yang sering dipandang bersifat “kualitatif” sehingga tidak membutuhkan sentuhan kuantitatif untuk menjelaskannya. Penulis melihat bahwa pemahaman mendasar terhadap statistik seharusnya lazim dimiliki oleh generasi peneliti bahasa di abad ke-21.

# Daftar pustaka

Glynn, D. (2010). Corpus-driven cognitive semantics: Introduction to the field. In Dylan Glynn & Kerstin Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 1–41). Berlin: Mouton de Gruyter.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Frankfurt am Main: Peter Lang. Retrieved from <http://www.linguistics.ucsb.edu/faculty/stgries/research/2010_STG_UsefulStats4CorpLing_MosaicCorpLing.pdf>

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction* (2nd ed.). Berlin: Mouton de Gruyter.

Gries, S. T. (2014). Basic significance testing. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (1st ed., pp. 316–336). Cambridge University Press. doi:[10.1017/CBO9781139013734.017](https://doi.org/10.1017/CBO9781139013734.017)

Janda, L. A. (2013a). Quantitative methods in *Cognitive Linguistics*: An introduction. In L. A. Janda (Ed.), *Cognitive Linguistics: The quantitative turn* (pp. 1–32). Berlin: Mouton de Gruyter.

Janda, L. A. (2016). Linguistic profiles: A quantitative approach to theoretical questions. *Język I Metoda*, 127–145. Retrieved from <http://www.ejournals.eu/pliki/art/6710/>

Janda, L. A. (Ed.). (2013b). *Cognitive linguistics: The quantitative turn*. Berlin: Mouton de Gruyter.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

Kuznetsova, J. (2015). *Linguistic profiles: Going from form to meaning via statistics*. Berlin: de Gruyter Mouton.

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Meyer, D., Zeileis, A., & Hornik, K. (2017). *Vcd: Visualizing categorical data*.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*(1), 1–39.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Stefanowitsch, A. (2004). Quantitative thinking for corpus linguists [Tutorial]. Retrieved August 4, 2011, from <http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html>

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, *16*(3), 507–525.

1. Pranala untuk *Sketch Engine* adalah <https://www.sketchengine.eu/> [↑](#footnote-ref-1)
2. Penulis menyadari terdapat beberapa bentuk kata lain yang tidak diikutkan dalam makalah ini (mis. *sepanas*, *terpanas*, *kepanasan*, dsb.). Namun, hal tersebut dianggap tidak menghalangi tujuan utama makalah ini untuk memperkenalkan pemahaman dasar atas kajian linguistik kuantitatif. [↑](#footnote-ref-2)
3. Gambar [1](#barplot) dihasilkan melalui R menggunakan modul “ggplot2” (Wickham, [2016](#ref-wickham_ggplot_2016)). [↑](#footnote-ref-3)
4. **Variabel Nominal/Kategorikal** menunjukkan bahwa unsur-unsur variabel tersebut memiliki ciri yang berbeda. Pada kasus kali ini, misalnya, variabel pemakaian/makna mengandung unsur dengan ciri berbeda, yaitu **literal** dan **metaforis**. Jenis variabel lainnya adalah **variabel ordinal** dan **variabel rasio/numerik** (periksa Gries, [2013](#ref-gries_statistics_2013), hlm. 16-17, untuk ulasan lebih dalam). [↑](#footnote-ref-4)
5. Unsur *chi* dilafalkan *ky*, yang memiliki rima sama dengan *high* atau *fly* di Bahasa Inggris (Stefanowitsch, [2004](#ref-stefanowitsch_quantitative_2004)). [↑](#footnote-ref-5)