Linguistik Indonesia, Februari 2019, 13-31 *Volume ke-37, No.1*

ISSN cetak 0215-4846; ISSN online 2580-2429

PEMAHAMAN KUANTITATIF DASAR DAN PENERAPANNYA DALAM MENGKAJI KETERKAITAN ANTARA BENTUK DAN MAKNA

Gede Primahadi Wijaya Rajeg [1,2] *& I Made Rajeg* [1*]
Universitas Udayana[1] & Monash University[2]
primahadiwijaya@gmail.com[1,2] & made_rajeg@unud.ac.id[1]

Abstrak

Makalah ini memaparkan pemahaman kuantitatif dasar, khususnya teknik statistik analitik *Chi-Square* (χ^2), dan penerapannya untuk permasalahan linguistik terkait hubungan antara bentuk dan makna. Berdasarkan data dari *Indonesian Web as Corpus*, makalah ini menggunakan medan leksikal PANAS sebagai contoh untuk mengukur keterkaitan antara pemakaian (non-)metaforis dan bentuk morfosintaksis kata berdasarkan sampel sitiran verba berakar *panas*, dan kata *panas* itu sendiri. Analisis χ^2 menunjukkan keterkaitan yang sangat signifikan dan kuat antara realisasi morfosintaksis medan leksikal PANAS dan distribusi pemakaian (non-)metaforisnya. Efek keterkaitan yang paling kuat ditunjukkan oleh bentuk inkohatif *memanas* yang berasosiasi positif dengan pemakaian metaforis, dan bentuk *dipanaskan* dan *panas* yang berasosiasi negatif dengan pemakaian metaforis. Temuan ini berimplikasi terhadap adanya ciri semantis dominan terkait wujud morfosintaksis suatu kata, fenomena yang dapat mencerminkan keterkaitan antara bentuk dan makna pada bahasa.

Kata kunci: Chi-Square, Linguistik Korpus Kuantitatif, bahasa pemrograman R, metafora, keberpasangan bentuk dan makna, medan leksikal PANAS, bahasa Indonesia

Abstract

This contribution discusses basic concepts of Chi-Square (χ^2) test as a kind of analytical statistics and illustrates its application to one of the central issues in linguistics, namely form-meaning relationship. As a case study using Indonesian Web as Corpus from the Sketch Engine, this paper measures the association between morphosyntactic forms of words in the lexical field of PANAS *'hot' and their (non-)metaphorical usages. The χ^2 test demonstrates a highly significant and robust association between the morphosyntactic form of words with the root* panas *'hot' and their preference for (non-)metaphorical usages. The clear effects are shown by (i) the strong preference of the inchoative form* memanas *'to become hot' for metaphorical usage, and (ii) the strong dispreference of* dipanaskan *'to be caused to be hot' and panas 'hot' for metaphorical usage. This finding has the implication on the predominant semantic trait of words with certain morphosyntactic forms, thus capturing the form-meaning relationship in language.*

Keywords: Chi-Square, Quantitative Corpus Linguistics, R programming language, metaphors, form-meaning pairing, lexical field of PANAS, Indonesian

PENDAHULUAN

Memasuki awal abad ke-21, penelitian linguistik umumnya, dan Linguistik Kognitif (LKog) khususnya, semakin gencar bergerak menuju pendekatan kuantitatif menggunakan beragam metode statistik untuk pengolahan data; pergerakan ini diistilahkan dengan "*the quantitative turn*" (Janda, 2013a; Levshina, 2015, hlm. 2-3). Revolusi kuantitatif ini utamanya dipicu oleh tiga faktor, yaitu (i) pergeseran teoretis ke arah pendekatan linguistik berdasarkan pemakaian (*usage-based linguistics*), seperti halnya LKog (Tummers, Heylen, & Geeraerts, 2005), (ii) ketersediaan sejumlah bank data elektronik pemakaian bahasa (*linguistic corpora*) yang ukurannya bisa mencapai miliaran, bahkan triliunan kata, dan (iii) keberadaan peranti statistik komputasional tak berbayar, terutama R (R Core Team, 2018), yang semakin lazim, baik di kalangan peneliti bahasa (Baayen, 2008; Gries, 2013; Janda, 2013a; Levshina, 2015) maupun pakar pengolahan data (*data scientists*) (Wickham & Grolemund, 2017).

Asas pemakaian (*usage-based assumption*) yang melandasi pendekatan linguistik seperti LKog berimplikasi metodologis terhadap pentingnya keberadaan data empiris dalam kajian kebahasaan (Tummers et al., 2005). Asas pemakaian menekankan bahwa kajian linguistik harus berdasarkan atas data pemakaian nyata, seperti pengamatan pada korpus pemakaian bahasa (Glynn, 2010, hlm. 10; Janda, 2016, hlm. 130) atau yang bisa diperoleh melalui survei dan eksperimen (Tummers et al., 2005, hlm. 229). Asumsi dasar dari pendekatan linguistik berdasarkan pemakaian ialah pengetahuan dan khazanah kebahasaan penutur dibentuk oleh pemakaian bahasa (Janda, 2016, hlm. 129-131). Selain itu, pengalaman berkelanjutan seorang penutur dalam berkomunikasi dan menerima masukan kebahasaan (*linguistic input*) juga mengikutkan tersimpannya informasi kuantitatif terkait konteks pemakaian suatu kata dan kekerapan kemunculan bersama (*co-occurrence frequencies*) seperangkat kata dan konstruksi (Levshina, 2015, hlm. 3). Kedua asumsi tersebut dapat diuji secara lebih empiris, di antaranya melalui pendekatan linguistik korpus kuantitatif dan metode eksperimental.

Hal yang penting untuk dicatat bahwa revolusi kuantitatif bukan berarti mengesampingkan intuisi dalam penelitian kebahasaan (bdk. Janda, 2013a, hlm. 3). Intuisi tetap berperan penting, terutama dalam (i) memaknai/interpretasi data dan (ii) merumuskan praanggapan, hipotesis, dan permasalahan teoretis dalam linguistik. Ketersediaan bank data dan penerapan metode kuantitatif bertujuan untuk membantu perumusan dan pemaparan permasalahan linguistik secara lebih empiris dan terukur (Stefanowitsch, 2010). Dengan kata lain, salah satu keuntungan dari penerapan metode linguistik korpus kuantitatif adalah informasi kuantitatif yang dapat diolah secara statistik guna mendukung dan/atau mengajukan argumentasi teoretis terkait suatu fenomena kebahasaan. Hasil tersebut selanjutnya akan kembali menjadi hipotesis yang dapat digugurkan atau didukung oleh penelitian empiris selanjutnya. Namun, untuk menggabungkan permasalahan teoretis dengan pendekatan kuantitatif bukan hal yang mudah (Janda, 2016, hlm. 128). Penyebabnya ialah bahwa perihal teoretis tidak secara gamblang mensyaratkan penerapan metode kuantitatif tertentu, dan analisis kuantitatif tidak selalu menjamin adanya relevansi teoretis. Oleh sebab itu, peneliti berperan penting dalam merumuskan permasalahan teoretis yang tepat dan dapat dijawab dengan pendekatan kuantitatif (Janda, 2016; Stefanowitsch, 2010; Tummers et al., 2005, hlm. 238).

Makalah ini memberikan contoh sederhana, tetapi mendasar terkait bagaimana pendekatan kuantitatif dapat menerangkan permasalahan teoretis dalam linguistik (Glynn, 2010; Gries, 2013; Janda, 2016, 2013b; Kuznetsova, 2015; Levshina, 2015; Tummers et al., 2005).

Salah satu pertanyaan utama (*big questions*) (Janda, 2016, hlm. 128) dalam linguistik ialah hubungan antara bentuk (*form*) dan makna (*meaning*) (Kuznetsova, 2015). Pendekatan linguistik seperti LKog memandang bentuk dan makna sebagai kesatuan penting dari bahasa dan dapat dikaji melalui pendekatan kuantitatif (Glynn, 2010; Janda, 2016; Stefanowitsch, 2010). Mengkaji hubungan antara bentuk dan makna berdasarkan asas pemakaian mengikutkan bahwa (i) perbedaan pemakaian suatu bentuk linguistik mengindikasikan perbedaan makna dari bentuk tersebut (periksa Kuznetsova, 2015, hlm. 13-14; Stefanowitsch, 2010, hlm. 368-370), dan (ii) perbedaan bentuk tentunya mencerminkan perbedaan makna (Janda, 2016, hlm. 129).

Makalah ini menggunakan objek kajian pemakaian (non-)metaforis medan leksikal PANAS dalam bahasa Indonesia (§2) untuk menunjukkan bahwa terdapat perbedaan distribusi pemakaian (non-)metaforis berkaitan dengan wujud morfosintaksis medan leksikal PANAS. Kajian serupa telah dilakukan sebelumnya oleh Deignan (2006) dengan data korpus bahasa Inggris. Deignan (2006) mengamati bahwa suatu kata cenderung digunakan secara metaforis dalam kelas kata yang berbeda dan pola sintaksis tertentu. Contohnya, proporsi pemakaian literal untuk lema *blossom* 'mekar' secara signifikan lebih tinggi dalam fungsinya sebagai nomina, sedangkan proporsi pemakaian metaforisnya lebih tinggi ketika berfungsi sebagai verba (Deignan, 2006, hlm. 112). Contoh serupa dalam bahasa Indonesia dapat ditemui pada verba denominal, seperti *mengepalai* (dari akar nomina *kepala*) dan *menangani* (dari akar nomina *tangan*). Kedua verba tersebut memiliki makna metaforis (mis. *mengepalai* selalu berarti 'memimpin'), sedangkan akar nominanya masih bisa digunakan, baik secara literal untuk mengacu pada anggota badan maupun secara metaforis (mis. ***kepala*** *negara* 'presiden' atau ***tangan*** *kanan* 'orang kepercayaan'). Temuan ini mencerminkan bagaimana makna tertentu (misalnya metaforis dan literal) dapat berkaitan erat dengan ciri morfosintaksis suatu kata (bdk. Janda, 2016, hlm. 129; Kuznetsova, 2015).

Kandungan pokok dari makalah ini terkait permasalahan di atas ialah pemaparan mendasar terhadap pemahaman analisis statistik yang digunakan, mulai dari yang bersifat deskriptif (§3.1) hingga analitik (*inferential/analytical statistics*) (§3.2) (Gries, 2013, Bab 4). Teknik statistik yang dibahas adalah uji signifikansi dengan *Chi-Square* yang dilambangkan dengan χ^2 (§3.3). Makalah ini juga menyisipkan bagian terkait penggunaan R untuk melakukan uji statistik *Chi-Square* (§3.4).

Terdapat dua alasan utama yang melatarbelakangi penggunaan *Chi-Square* (χ^2) pada makalah ini. Alasan pertama berkaitan dengan tujuan analitis makalah ini guna mengukur keterkaitan antardua variabel yang bersifat 'kategorikal' (Gries, 2013, hlm. 16-17), yaitu BENTUK (MORFOSINTAKSIS) dan MAKNA[1]. Artinya, kedua variabel tersebut memiliki unsur/terdiri atas kategori dengan ciri yang berbeda. Misalnya, variabel BENTUK terdiri atas beragam bentuk morfosintaksis medan leksikal PANAS (§2). Sementara itu, variabel MAKNA memiliki dua unsur kategori, yaitu *LITERAL* dan *METAFORIS* (§3.2). *Chi-Square* ialah uji signifikansi statistik yang lazim digunakan untuk mengukur keterkaitan antardua variabel kategorikal berdasarkan kekerapan kemunculan suatu variabel kategorikal (mis. makna literal dan/atau metaforis) terhadap variabel kategorikal yang lain (mis. bentuk morfosintaksis kata) (Gries, 2013, hlm. 178; Janda, 2013a; Levshina, 2015, hlm. 199).

Alasan kedua berkaitan dengan aspek pemahaman kuantitatif secara lebih luas. *Chi-Square* dapat membangun pemahaman terhadap sejumlah konsep penting dalam uji signifikansi statistik secara umum. Konsep-konsep tersebut meliputi hipotesis kosong (*null hypothesis*) dan hipotesis alternatif (*alternative hypothesis*) (§3.2), kekerapan riil (*observed frequency*) dan

kekerapan harapan (*expected frequency*) (§3.3.1), serta tingkat signifikansi (*significance level*) (§3.2) dan peluang kesalahan (*probability of error* atau *p-value*) (§3.3.2). Pemahaman terhadap konsep-konsep tersebut bukan semata-mata untuk membantu peneliti (i) menerapkannya dalam suatu analisis, melainkan juga, yang tak kalah pentingnya, (ii) memahami tulisan-tulisan lain yang menggunakan uji signifikansi statistik seperti *Chi-Square*. Dengan kata lain, *Chi-Square* dapat dijadikan sebagai salah satu batu loncatan guna memahami uji signifikansi statistik lainnya untuk jenis variabel berbeda (Gries, 2013).

Berdasarkan dua latar belakang tersebut, makalah ini bertujuan untuk berbagi pemahaman mendasar terhadap uji signifikansi statistik *Chi-Square* dan menerapkannya terhadap permasalahan linguistik mendasar namun krusial, yaitu hubungan antara bentuk dan makna (Janda, 2016; Kuznetsova, 2015). Selanjutnya, makalah ini diharapkan dapat meningkatkan minat dan wawasan terhadap kajian linguistik Indonesia melalui pendekatan kuantitatif (baik berdasarkan data korpus ataupun eksperimen), yang dewasa ini masih terbilang jarang (beberapa contohnya, Arka, 2017; Aryawibawa & Ambridge, 2018; Denistia & Baayen, 2019; Denistia, Bajestan, & Baayen, 2018; Moeljadi, 2011, 2014; Musgrave, 2013; G. P. W. Rajeg, 2014, 2016, 2018; G. P. W. Rajeg & Rajeg, 2017, 2019a; G. P. W. Rajeg, Denistia, & Musgrave, 2018; G. P. W. Rajeg, Denistia, & Rajeg, 2018; Rajeg, 2014; Siahaan, 2011, 2015).

DATA

Korpus yang menjadi sumber data makalah ini ialah *Indonesian Web as Corpus* (IWaC) (109.281.359 kata) (Kilgarriff et al., 2014). IWaC tersedia daring melalui layanan korpus berbayar *Sketch Engine* (SE) (https://www.sketchengine.eu/). Penulis mencari 100 sampel sitiran acak pemakaian kata-kata dengan akar kata *panas*, yaitu *panas* itu sendiri, kemudian bentuk turunannya sebagai kata kerja, yaitu *memanas*, *memanaskan*, *dipanaskan*, *dipanasi*, dan *memanasi*. Gugusan kata tersebut tercakup dalam medan leksikal PANAS. Penulis menyadari terdapat beberapa bentuk lain yang tidak disertakan dalam makalah ini, misalnya *berpanas-panasan*, *sepanas*, *terpanas*, *kepanasan*, dan lainnya. Namun, hal tersebut dianggap tidak menghalangi tujuan utama makalah ini, yaitu untuk memperkenalkan pemahaman dasar atas uji signifikansi *Chi-Square* dan penerapannya terhadap salah satu kajian linguistik. Pencarian sampel sitiran untuk kata yang diteliti dilakukan dalam rentang waktu tiga puluh hari percobaan akun gratis untuk SE pada pertengahan tahun 2013. Tabel 1 menampilkan lima sitiran acak dari total 458 sitiran medan leksikal PANAS yang dikaji, 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 menemukan dan mengenali kedua pemakaian tersebut secara manual dengan melihat konteks pemakaian kata kunci pada tiap-tiap sitiran berdasarkan *Metaphor Identification Procedure* (MIP) (Pragglejaz Group, 2007). Pemakaian literal kata-kata tersebut secara umum mengacu pada pemakaiannya dalam ranah suhu yang bersifat badaniah (periksa tiga baris sitiran pertama pada Tabel 1). Sementara itu, pemakaian metaforisnya menunjukkan pemetaan unsur suhu badaniah ke ranah yang lebih abstrak, seperti emosi atau intensitas suatu keadaan (seperti pada dua baris terakhir pada Tabel 1). Sitiran ganda (*duplicates*) secara manual ditandai dan tidak disertakan dalam analisis kuantitatif selanjutnya. Semua analisis statistik, termasuk tabel, grafik, dan penulisan makalah ini dilakukan melalui *RStudio* menggunakan *R Markdown Notebook* dan dua modul R pendukung, yaitu *tidyverse* (https://www.tidyverse.org) (Wickham & Grolemund, 2017) dan *vcd* (Meyer, Zeileis, & Hornik, 2017; Zeileis, Meyer, & Hornik, 2007). Data sitiran yang dikaji dan berkas *R Markdown Notebook* tersedia dengan akses terbuka melalui repositori *figshare* (G. P. W. Rajeg & Rajeg, 2019b).

HASIL DAN PEMBAHASAN

Bagian ini mencakup tiga pokok pembahasan. Pertama, pembahasan analisis statistik yang bersifat deskriptif (§3.1). Kedua, pemahaman terkait signifikansi statistik dan hipotesis ilmiah (§3.2), serta uji signifikansi statistik menggunakan *Chi-Square* (§3.3). Ketiga, penggunaan R dalam melakukan *Chi-Square* (§3.4).

1.1 Memahami data melalui tabel dan grafik

Pemaparan deskriptif terkait distribusi kategori yang mencakup variabel BENTUK MORFOSINTAKSIS (*panas*, *memanas*, *memanasi*, dsb.) serta MAKNA (*literal* dan *metaforis*) dapat diawali dengan tabulasi kekerapan silang (*crosstabulation*) seperti pada Tabel 2 berikut.

Tabel 2. Kekerapan riil antara kategori MAKNA dan BENTUK MORFOSINTAKSIS

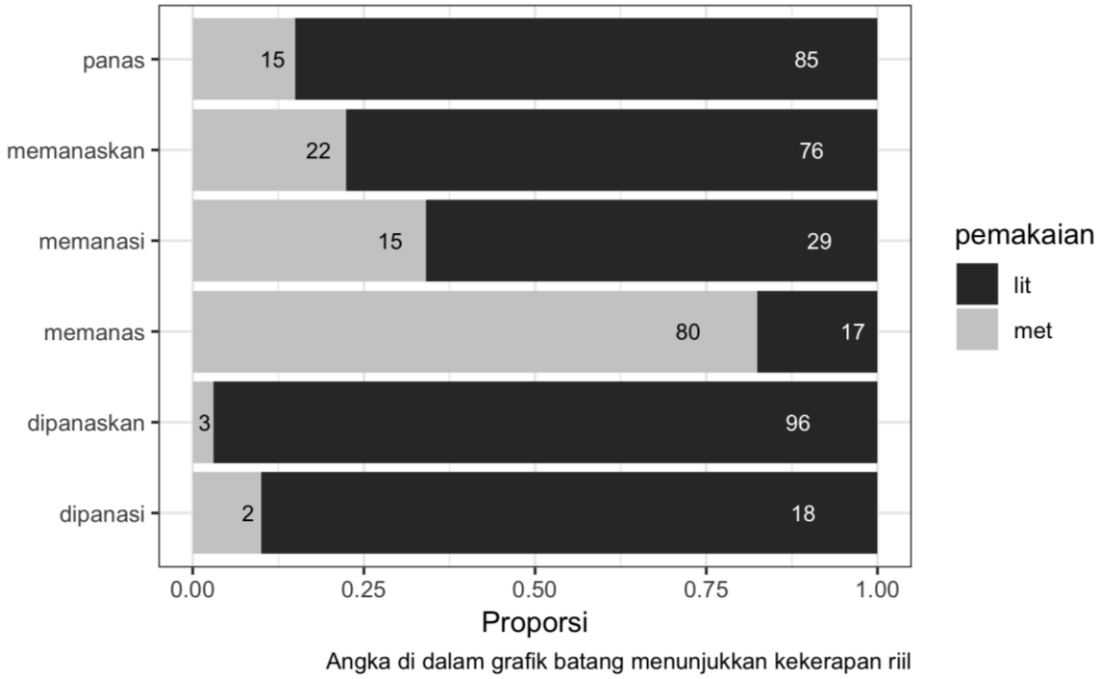
	dipanasi	dipanaskan	memanas	memanasi	memanaskan	panas	total
lit	18	96	17	29	76	85	321
met	2	3	80	15	22	15	137
total	20	99	97	44	98	100	458

Data dalam tabel di atas menunjukkan bahwa pemakaian literal bentuk pasif *dipanasi* (18 sitiran) berjumlah lebih sedikit dibandingkan dengan bentuk aktif *memanasi* (29). Namun, hal yang perlu diperhatikan ialah jumlah keseluruhan sampel dari tiap-tiap kata berbeda. Misalnya, pencarian 100 sampel untuk bentuk *dipanasi* di SE hanya menghasilkan total 20 sitiran, begitupun dengan *memanasi* (44 sitiran). Untuk dapat melakukan perbandingan yang seimbang, sebaiknya nilai kekerapan diseragamkan terlebih dahulu (Levshina, 2015, hlm. 70). Penyeragaman bisa dalam bentuk (i) kekerapan relatif (*relative frequency* atau *proportion* dalam bentuk bilangan desimal dengan rentangan antara 0 sampai 1) (Periksa Tabel 3 berikut), (ii) persentase, atau (iii) kekerapan berdasarkan suatu bilangan acuan (misalnya kekerapan per 100 atau 1000 sitiran) yang dikenal dengan istilah *normalised frequency* (Gries, 2010, hlm. 271).

Tabel 3. Kekerapan relatif antara kategori MAKNA dan BENTUK MORFOSINTAKSIS

	dipanasi	dipanaskan	memanas	memanasi	memanaskan	panas
lit	0.9	0.97	0.18	0.66	0.78	0.85
met	0.1	0.03	0.82	0.34	0.22	0.15

Kini dapat diperhatikan dalam tabel di atas 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%). Proporsi makna literal dan metaforis untuk suatu kata diperoleh dengan membagi kekerapan suatu jenis makna dengan total sitiran kata tersebut. Sebagai contoh, penghitungan proporsi makna literal *dipanasi* adalah $\frac{18}{20}$, yaitu 0.9, dan proporsi metaforisnya adalah $\frac{2}{20}$, yaitu 0.1. Informasi pada Tabel 2 dan Tabel 3 dapat dipahami secara lebih intuitif melalui tampilan visual diagram batang seperti Gambar 1, yang dihasilkan menggunakan modul *ggplot2* (Wickham, 2016) sebagai bagian dari *tidyverse*.



Gambar 1. Proporsi sitiran literal dan metaforis terkait bentuk morfosintaksis medan leksikal PANAS

Pada Gambar 1 terlihat jelas bahwa verba inkohatif *memanas* memiliki proporsi sitiran metaforis yang paling tinggi dibandingkan bentuk verba lainnya, yang didominasi oleh pemakaian literal. Ulasan deskriptif ini dapat mengindikasikan peluang adanya hubungan antara morfosintaksis kata berakar sama dan makna tertentu yang cenderung disampaikannya. Pertanyaan selanjutnya adalah bagaimana cara menentukan secara empiris bahwa perbedaan distribusi makna dari sampel pemakaian medan leksikal PANAS ini penting dan bukanlah suatu kebetulan? Dengan kata lain, bagaimana kita mengetahui bahwa distribusi tersebut signifikan secara statistik? §3.2 berikut mengulas secara ringkas pemahaman istilah signifikansi statistik.

1.2 Signifikansi statistik dan hipotesis ilmiah

Signifikansi statistik merujuk pada pertanyaan apakah perbedaan distribusi yang diamati dalam sampel hanya merupakan suatu kebetulan (*based on chance*) (Gries, 2013, hlm. 28; Stefanowitsch, 2004)[2]. Suatu distribusi dikatakan signifikan secara statistik apabila peluang (*likelihood* atau *probability*) bahwa distribusi tersebut bersifat acak/manasuka (*random*) lebih kecil dari 5%. Nilai ini disebut dengan tingkat signifikansi ($p_{critical}$) dan umumnya ditulis dalam bentuk desimal, yaitu 0.05 (hasil dari 5/100) (Gries, 2013, hlm. 27, 2014, hlm. 317; Janda, 2013a, hlm. 9-10).

Jika dikaitkan dengan contoh kajian makalah ini, signifikansi statistik merujuk pada peluang ditemukannya perbedaan persebaran pemakaian literal dan metaforis terkait bentuk morfosintaksis medan leksikal PANAS (seperti pada Tabel 2) apabila benar adanya bahwa seharusnya tidak ada perbedaan persebaran makna literal dan metaforis terkait realisasi morfosintaksis dari medan leksikal PANAS tersebut. Hal yang penting untuk diperhatikan ialah signifikansi statistik dari distribusi suatu fenomena kebahasaan tidak akan pernah menjamin signifikansinya secara linguistik, dalam arti, tidak menunjukkan karakteristik penting suatu

sistem kebahasaan (bdk. Gries, 2013, hlm. 28). Namun, jika distribusi suatu fenomena kebahasaan tidak signifikan secara statistik, hal menarik sehubungan dengan pola kebahasaan tersebut mungkin tidak akan terangkat (Stefanowitsch, 2004).

Analisis statistik terkait uji signifikansi melibatkan dua hipotesis yang diuji berdasarkan data, yaitu (i) hipotesis kosong (*null hypothesis*) (selanjutnya disebut H_0) dan (ii) hipotesis alternatif (*alternative hypothesis*) (H_1) (Gries, 2014, hlm. 317; Levshina, 2015, hlm. 8-9). H_0 umumnya menyatakan distribusi setara/acak atau ketidakterkaitan di antara variabel yang dikaji, sedangkan H_1 menyatakan adanya keterkaitan di antara variabel berdasarkan perbedaan distribusinya (periksa Gries, 2013, hlm. 10-14). Berikut dipaparkan dua hipotesis yang berkaitan dengan makalah ini:

- H_0 (hipotesis kosong): Rasio pemakaian literal dan metaforis tidak berbeda sehubungan dengan beragam bentuk morfosintaksis kata-kata pada medan leksikal PANAS. Artinya, bentuk intransitif *memanas*, transitif *memanaskan*, dan bentuk pasif *dipanaskan* seharusnya memiliki distribusi pemakaian literal dan metaforis yang setara dalam sampel.
- H_1 (hipotesis alternatif): Rasio pemakaian literal dan metaforis berbeda sehubungan dengan beragam bentuk morfosintaksis kata-kata pada medan leksikal PANAS. Dengan kata lain, diasumsikan bahwa terdapat kaitan antara bentuk morfosintaksis kata pada medan leksikal PANAS dan proporsi makna literal serta metaforis terkait pemakaian kata tersebut dalam korpus.

Hal yang tidak diperinci dari H_1 terkait perbedaan distribusi ini ialah 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 H_1 adalah adanya perbedaan distribusi. Arah dari perbedaan distribusi tersebut tidak dapat diperinci untuk saat ini karena belum ditemukannya kajian sebelumnya dalam bahasa Indonesia yang melihat kaitan antara bentuk morfosintaksis kata berakar sama serta makna literal dan/atau metaforis yang disampaikan dalam pemakaian kata tersebut. Meskipun demikian, perlu diketahui bahwa Siahaan (2015) telah melakukan kajian metafora konseptual terkait pemakaian medan leksikal TEMPERATUR dalam bahasa Indonesia, tetapi dengan menggabungkan semua realisasi morfosintaksis guna mencerminkan suatu jenis temperatur. Misalnya, pemakaian metaforis *kepanasan*, *memanaskan*, dan *panas*, digabungkan ke dalam sitiran metaforis konsep PANAS tanpa melihat perbedaan dan kaitannya dengan bentuk morfosintaksis kata-kata tersebut. Oleh sebab itu, hasil analisis dalam makalah ini akan memberikan asumsi awal terhadap perbedaan distribusi pemakaian literal dan metaforis terhadap bentuk morfosintaksis kata-kata tersebut, yang selanjutnya dapat diujikan kembali menggunakan sampel yang lebih besar, medan leksikal lainnya, dan jenis korpus yang berbeda. Lebih lanjut, tujuan utama dari uji signifikansi statistik bukan mencoba membuktikan kebenaran H_1, melainkan mencoba menggugurkan asumsi yang diajukan oleh H_0 (periksa Gries, 2013, hlm. 26-29 untuk ulasan lebih dalam).

Salah satu metode uji signifikansi statistik yang dapat digunakan untuk menguji signifikansi distribusi variabel kategorikal seperti pada Tabel 2 adalah *Chi-Square test*[3] (Gries, 2013, hlm. 178-189; Janda, 2013a, hlm. 9-14; Levshina, 2015, Bab 9; Stefanowitsch, 2004). Variabel kategorikal atau yang juga dikenal dengan variabel nominal adalah variabel yang unsur-unsurnya memiliki ciri yang berbeda. Pada kasus ini, misalnya, variabel MAKNA mengandung unsur dengan ciri berbeda, yaitu *LITERAL* dan *METAFORIS*. Jenis variabel lainnya adalah variabel ordinal (mis. *RENDAH*, *SEDANG*, *TINGGI*) dan variabel rasio/numerik (mis.

panjang kata berdasarkan jumlah suku kata atau panjang suatu kalimat berdasarkan jumlah kata), yang kaitan di antara unsurnya dapat dihubungkan dengan kalimat komparatif (mis. kata X dua kali lebih panjang dibandingkan kata Y berdasarkan jumlah suku katanya) (periksa Gries, 2013, hlm. 16-17, untuk ulasan lebih dalam).

1.3 Uji signifikansi Chi-Square (χ^2)

1.3.1 *Kekerapan harapan dan kekerapan riil*

Hal pertama yang mesti dilakukan untuk mengetahui apakah suatu distribusi muncul secara acak/kebetulan adalah dengan mengetahui distribusi yang diharapkan muncul berdasarkan H_0, atau yang dikenal dengan istilah kekerapan harapan (*expected frequency*, selanjutnya disingkat menjadi F_e) (Gries, 2013, hlm. 180; Janda, 2013a, hlm. 9; Levshina, 2015, hlm. 210). Dalam hal ini, F_e adalah distribusi yang diharapkan muncul jika memang tidak ada hubungan antara bentuk morfosintaksis medan leksikal PANAS dan proporsi pemakaian literal serta metaforis (yaitu, H_0) (periksa §3.2). Tes χ^2 melibatkan perbandingan antara F_e dan kekerapan yang diamati pada sampel data, yang disebut dengan kekerapan riil (*observed frequency*, disingkat F_o). Berikut adalah rumus umum untuk menghitung F_e tiap-tiap sel, yaitu E_{ij}, di dalam tabel layaknya Tabel 2 (bdk. Levshina, 2015, hlm. 211):

$$E_{ij} = \frac{S_i \cdot S_j}{N}$$

S_i menunjukkan kekerapan marginal total (*marginal frequency*) baris i, S_j menunjukkan kekerapan total kolom j, dan N adalah total pengamatan dalam sampel. Oleh sebab itu, F_e tiap-tiap sel ditentukan berdasarkan kekerapan total tiap-tiap kolom dan baris (ditandai dengan huruf miring pada Tabel 4, yang merupakan pengulangan Tabel 2).

Tabel 4 Kekerapan riil antara kategori MAKNA dan BENTUK MORFOSINTAKSIS

	dipanasi	dipanaskan	memanas	memanasi	memanaskan	panas	total
lit	18	96	17	29	76	85	*321*
met	2	3	80	15	22	15	*137*
total	*20*	*99*	*97*	*44*	*98*	*100*	*458*

Sebagai contoh, penghitungan F_e untuk sel pada persimpangan baris *lit* dan kolom *dipanasi* adalah $\frac{20 \cdot 321}{458}$, yaitu 14.017. Sebagai contoh tambahan, F_e untuk sel pada persimpangan *met* dan *memanas* adalah $\frac{97 \cdot 137}{458}$, yaitu 29.015. Tabel 5 menunjukkan nilai F_e untuk semua kombinasi kategori MAKNA dan BENTUK MORFOSINTAKSIS; nilai F_e tersebut telah dibulatkan menjadi 3 angka desimal setelah koma.

Tabel 5. Kekerapan harapan antara kategori MAKNA dan BENTUK MORFOSINTAKSIS

	dipanasi	dipanaskan	memanas	memanasi	memanaskan	panas
lit	14.017	69.386	67.985	30.838	68.686	70.087
met	5.983	29.614	29.015	13.162	29.314	29.913

Nilai total tiap-tiap kolom dan baris pada Tabel 5 akan sama dengan nilai total tiap-tiap baris dan kolom pada Tabel 4. Misalnya, total penjumlahan nilai F_e untuk kolom *memanas* ialah 97 (yaitu 67.985 [$F_{e\ lit}$] + 29.015 [$F_{e\ met}$] = 97), jumlah yang sama untuk *memanas* pada Tabel 4.

Setelah mengetahui bagaimana F_e diperoleh, mungkin muncul pertanyaan mengapa F_e dihitung sedemikian rupa. Penalarannya ialah sebagai berikut (periksa Gries, 2013, hlm. 182-183; Stefanowitsch, 2004). Kekerapan total tiap-tiap baris dan kolom adalah nilai yang tetap (*fixed*) dan demikian adanya (*given*) berdasarkan sampel; nilai tersebut menunjukkan (i) total sitiran literal serta metaforis dalam sampel dan (ii) total sitiran sampel untuk tiap-tiap bentuk morfosintaksis medan leksikal PANAS. Apabila nilai total tersebut terdistribusi secara acak untuk kedua belas persimpangan baris dan kolom dalam Tabel 4 (*met+dipanasi*, *lit+dipanasi*, *met+dipanaskan*, *lit+dipanaskan*, dan seterusnya), distribusinya harus proporsional, yang artinya dapat dipahami sebagai berikut. Terdapat 321 sitiran (dari total 458 sitiran) medan leksikal PANAS yang menunjukkan pemakaian literal dan 137 sisanya merupakan sitiran metaforis. Dengan kata lain, 70.09% dari total 458 sitiran medan leksikal PANAS yang dikaji digunakan secara literal (yaitu, $\frac{321}{458} \cdot 100$) dan 29.91%-nya digunakan secara metaforis (yaitu, $\frac{137}{458} \cdot 100$). Apabila persebaran pemakaian literal dan metaforis untuk keenam kata medan leksikal PANAS tersebut didasarkan atas kebetulan atau bersifat acak (*based on chance*), persebarannya harus proporsional dengan persentase pemakaian literal dan metaforis secara umum. Artinya, 70.09% dari total sitiran untuk *panas*, *memanas*, *memanasi*, dan seterusnya. seharusnya bermakna literal dan 29.91% dari total sitiran untuk tiap-tiap kata tersebut seharusnya bermakna metaforis. Sebagai contoh, 70.09% dari total 20 sitiran untuk *dipanasi* yang seharusnya bermakna literal adalah sekitar 14.017 sitiran (yaitu, $\frac{70.09}{100} \cdot 20$), nilai yang persis muncul pada sel untuk *lit+dipanasi* pada Tabel 5 di atas.

Dengan membandingkan nilai tiap-tiap sel pada Tabel 5 dan Tabel 4, kita dapat mengetahui kombinasi mana yang muncul lebih sering atau jarang dari yang diharapkan. Misalnya, *memanas* lebih sering dari yang diharapkan untuk digunakan secara metaforis. Tahap selanjutnya (§3.3.2) ialah menghitung kadar perbedaan F_o terhadap F_e untuk mengetahui bahwa kekerapan riil yang diamati dalam sampel bukanlah suatu kebetulan.

1.3.2 *Kontribusi sel terhadap nilai statistik χ^2 dan tingkat signifikansi*

Uji statistik χ^2 mensyaratkan bahwa 80% dari nilai F_e pada tiap sel suatu tabel kekerapan harus lebih besar atau sama dengan lima (Gries, 2013, hlm. 166; periksa juga Levshina, 2015, hlm. 212). Apabila persyaratan tersebut tidak terpenuhi atau nilai dalam tabulasi silang terlalu kecil, disarankan menggunakan uji signifikansi *Fisher-Yates Exact* (Levshina, 2015, hlm. 213-214). Tabel 5 menunjukkan bahwa semua (100%) sel memiliki F_e lebih besar dari lima. Nilai statistik χ^2 diperoleh dengan menghitung nilai kontribusi tiap-tiap sel terhadap χ^2 ("*contribution to chi-squared*" (Gries, 2013, hlm. 168)). Perhitungan tersebut dirumuskan sebagai berikut (Levshina, 2015, hlm. 212; Stefanowitsch, 2004):

$$\text{Pearson } \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Rumus tersebut menunjukkan bahwa nilai χ^2 merupakan hasil penjumlahan dari pemangkatan perbedaan F_o dengan F_e yang dibagi dengan F_e untuk tiap-tiap sel. Sebagai contoh, nilai kontribusi terhadap χ^2 untuk sel *lit*+*dipanasi* adalah $\frac{(18-14.017)^2}{14.017}$, yaitu 1.131; ulangi penghitungan serupa untuk sel yang lain untuk selanjutnya dijumlahkan menjadi nilai statistik χ^2, yang hasilnya adalah χ^2 = 179.31. Pertanyaan berikutnya ialah apakah nilai χ^2 ini mengindikasikan perbedaan/penyimpangan (*deviation*) yang cukup besar oleh F_o terhadap F_e sehingga distribusi riil F_o pada Tabel 4 dapat dikatakan bukan suatu kebetulan?

Peranti statistik umumnya menggunakan nilai χ^2 guna mengukur peluang kesalahan (*probability of error*), yang juga dikenal dengan sebutan *p-value* (Gries, 2013, hlm. 27, 2014, hlm. 317). *P-value* menunjukkan peluang ditemukannya (i) distribusi dalam sampel (yaitu, F_o dalam Tabel 4) dan juga (ii) penyimpangannya dari F_e yang diharapkan oleh H_0, ketika H_0 dianggap benar. Apabila peluang ini lebih kecil dari tingkat signifikansi 5% (§3.2) mengingat H_0 dianggap benar, (i) distribusi yang diamati dalam sampel dapat dikatakan signifikan secara statistik, yaitu tidak muncul berdasarkan suatu kebetulan dan (ii) kita dapat menyangkal asumsi dari H_0 dengan menunjukkan bahwa terdapat perbedaan distribusi yang tidak bisa dikatakan sebagai suatu kebetulan antardua variabel atau lebih (Gries, 2013, hlm. 27; Levshina, 2015, hlm. 12). Penyangkalan terhadap H_0 ini tidak berarti bahwa H_1 atau efek perbedaan yang ditemukan benar adanya karena masih terdapat probabilitas/peluang distribusi tersebut muncul secara kebetulan, meskipun sangat kecil (Gries, 2014). Sebelum adanya komputer, cara klasik untuk menentukan tingkat signifikansi suatu distribusi ialah membandingkan nilai χ^2 yang diperoleh berdasarkan sampel, yaitu χ^2 = 179.31, dengan tabel nilai χ^2 seperti yang ditunjukkan pada Tabel 6 (Gries, 2013, hlm. 184).

Tabel 6. Nilai χ^2 untuk tingkat signifikansi ($p_{critical}$) = 0.05 (5%), 0.01 (1%), dan 0.001 (0.1%) untuk $1 \leq df \leq 5$

df	*p* = 0.05	*p* = 0.01	*p* = 0.001	…
1	3.841	6.635	10.828	…
2	5.991	9.21	13.816	…
3	7.815	11.345	16.266	…
4	9.488	13.277	18.467	…
5	11.07	15.086	20.515	…
…	…	…	…	…

Kolom *df* menunjukkan *degree of freedom* (Levshina, 2015, hlm. 12) dari suatu tabel distribusi. *Df* secara singkat mengacu pada jumlah nilai yang dapat berubah (*vary*). Rumus pengukuran nilai *df* adalah sebagai berikut:

$$df = (\text{N baris} - 1) \cdot (\text{N kolom} - 1) = (2-1) \cdot (6-1) = 5$$

Nilai *df* = 5 dengan kekerapan marginal total yang tetap menunjukkan bahwa hanya 5 sel dari Tabel 4 yang nilainya dapat diubah tanpa mengubah jumlah kekerapan marginalnya (bdk. Levshina, 2015, hlm. 12).

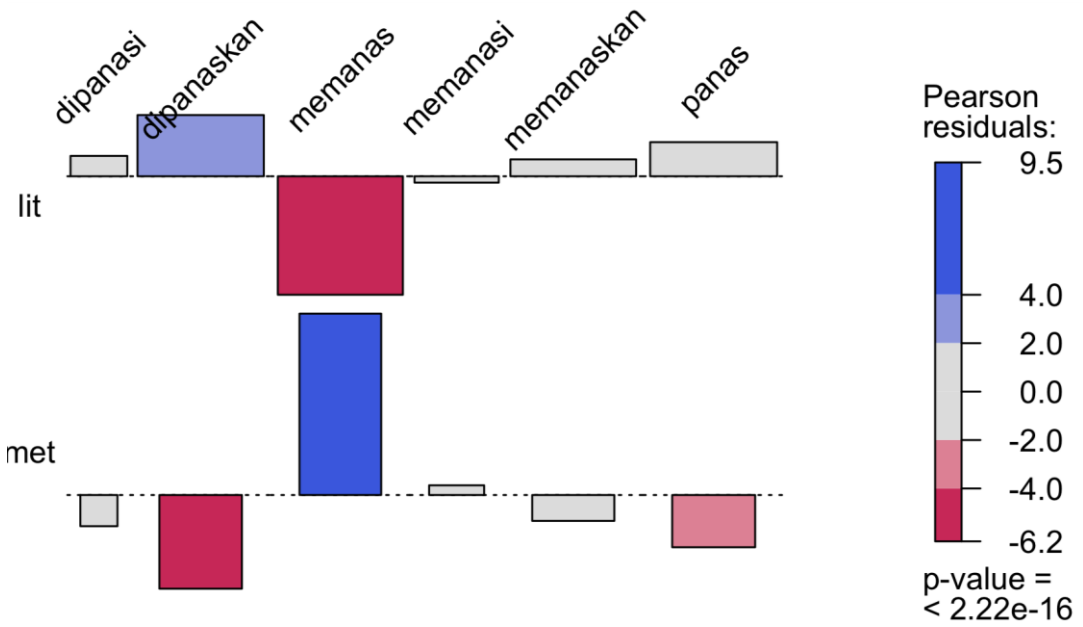
Selanjutnya, untuk mengetahui signifikansi distribusi pemakaian literal dan metaforis dari medan leksikal PANAS, dapat dimulai dengan baris yang menunjukkan nilai *df* = 5 dan

memeriksa apakah nilai χ^2 = 179.31 yang diperoleh lebih besar dari nilai χ^2 yang ada pada baris tersebut pada Tabel 6. Terlihat jelas bahwa nilai χ^2 yang diperoleh dalam sampel hampir sekitar sembilan kali lebih besar dari nilai χ^2 yang ada pada kolom signifikansi 0.1%, yaitu kolom p = 0.001 dengan χ^2 = 20.515. Hal ini menunjukkan bahwa terdapat peluang yang sangat kecil (lebih kecil dari 0.1% atau $p < 0.001$) untuk menemukan perbedaan distribusi antara bentuk morfosintaksis dan tipe makna pada Tabel 4 apabila (i) perbedaan tersebut dianggap sebagai suatu kebetulan dan (ii) tidak ada keterkaitan antara bentuk morfosintaksis medan leksikal PANAS dan makna (non-)metaforis.

Nilai statistik lain yang mesti dilaporkan selain *p-value* ialah kadar efek (*effect size*) (Janda, 2013a, hlm. 11). Berbeda dengan *p-value* yang menunjukkan peluang acak tidaknya kemunculan suatu distribusi, kadar efek menunjukkan seberapa penting dan kuat efek hubungan/korelasi di antara kedua variabel. Untuk tabel kekerapan dengan jumlah baris dan/atau kolom di atas dua, kadar efek yang dilaporkan ialah *Cramér's V* (Gries, 2013, hlm. 185-186; Janda, 2013a, hlm. 10; Levshina, 2015, hlm. 217). Rentangan nilai *Cramér's V* ialah antara 0 (tidak ada hubungan) dan 1 (hubungan sempurna): (i) 0.1 adalah ambang batas untuk kadar efek kecil yang dapat dilaporkan; (ii) 0.3 untuk kadar efek cukup (*moderate*); dan (iii) 0.5 untuk kadar efek besar/kuat (*robust*) (Janda, 2013a, hlm. 10-11; bdk. Levshina, 2015, hlm. 209). Berikut ini adalah rumus penghitungan *Cramér's V* (Gries, 2013, hlm. 186):

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{N \cdot (min[N_{baris}, N_{kolom}] - 1)}} = \sqrt{\frac{179.31}{458 \cdot (2-1)}} = 0.626$$

Jadi, hasil analisis statistik *Chi-Square* menunjukkan perbedaan distribusi yang sangat signifikan dan keterkaitan yang kuat antara bentuk morfosintaksis kata dalam medan leksikal PANAS dan pemakaian literal serta metaforisnya (N = 458; χ^2 = 179.31; df = 5; $p < 0.001$; *Cramér's V* = 0.626) (bdk. Deignan, 2006). Gambar 2 berikut, yang dihasilkan melalui modul R *vcd*, merupakan bagan asosiasi yang secara lebih jelas menunjukkan arah dan besar nilai penyimpangan F_o terhadap F_e tiap-tiap sel terkait dua variabel yang dikaji (Gries, 2013, hlm. 188; Levshina, 2015, hlm. 220).



Gambar 2. Bagan asosiasi kategori MAKNA dan BENTUK MORFOSINTAKSIS medan leksikal PANAS

Nilai *Pearson residuals* pada gambar di atas menunjukkan besaran nilai efek perbedaan antara kekerapan riil (F_o) dan kekerapan harapan (F_e) tiap-tiap sel (Gries, 2013, hlm. 187-188; Levshina, 2015, hlm. 218-219). Semakin besar nilai absolut residual tiap-tiap sel, maka semakin besar pula penyimpangan F_o terhadap F_e, dan kontribusi sel tersebut terhadap nilai statistik χ^2 keseluruhan. Efek yang berkontribusi kuat terhadap keterkaitan signifikan antara variabel MAKNA dan BENTUK MORFOSINTAKSIS ditunjukkan melalui dua unsur, yaitu (i) panjang-pendeknya diagram batang yang dapat menjulang ke atas (residual positif, yaitu $F_o > F_e$) atau terbalik ke bawah (residual negatif, yaitu $F_o < F_e$), dan (ii) kegelapan warna diagram batang tiap-tiap sel (semakin gelap warnanya, maka semakin besar pula efek perbedaan F_o terhadap F_e) (bdk. Levshina, 2015, hlm. 220). Berdasarkan informasi ini, tampak bahwa verba inkohatif *memanas* adalah satu-satunya bentuk yang berasosiasi paling kuat dengan pemakaian metaforis dibandingkan pemakaian literalnya. Sementara itu, bentuk pasif kausatif *dipanaskan* dan bentuk akar *panas* berasosiasi negatif dengan pemakaian metaforis dalam sampel. Selain itu, *dipanaskan* juga satu-satunya bentuk yang berasosiasi cukup kuat dengan pemakaian literal dalam sampel. Hasil ini menunjukkan bahwa (i) makna tertentu bisa memiliki rasio yang lebih tinggi untuk diungkapkan dalam bentuk morfosintaksis kata tertentu, dan (ii) keberpasangan bentuk dan makna tersebut dapat ditentukan secara lebih terukur dengan merumuskannya secara kuantitatif (bdk. Janda, 2016).

1.4 Uji signifikansi χ^2 dengan R

Pemaparan yang terperinci dan bertahap sebelumnya terkait χ^2 dan uji signifikansi secara umum penting untuk dipahami sebelum dipermudah dengan bantuan peranti komputasional seperti R. Nukilan kode berikut menunjukkan tahapan analisis χ^2 dengan R. Pertama-tama, data mentah makalah ini, yaitu berkas `panas_raw.txt` yang disimpan dalam folder `data`, mesti dimuat

ke dalam R dan disimpan ke dalam objek bernama `panas` (lihat kembali Tabel 1). Data ini selanjutnya digunakan untuk membuat tabulasi kekerapan seperti Tabel 4.

```
# muat data konkordansi ke dalam R dan simpan ke objek `panas`
panas <- read.table(file = "data/panas_raw.txt",
                    header = TRUE,
                    sep = "\t",
                    quote = "",
                    comment.char = "",
                    stringsAsFactors = FALSE)

# buat tabel kekerapan antara kolom use & kolom node
panas_tbl <- table(panas$use, panas$node)

# periksa tampilan tabel
panas_tbl
```

```
##
##       dipanasi dipanaskan memanas memanasi memanaskan panas
##   lit       18         96      17       29         76    85
##   met        2          3      80       15         22    15
```

Analisis χ^2 pada R dilakukan melalui fungsi `chisq.test()` dengan masukan berupa tabulasi data `panas_tbl` yang dihasilkan sebelumnya.

```
# Uji signifikansi dengan fungsi `chisq.test()`
# dan simpan hasil ke objek `panas_chi`
panas_chi <- chisq.test(panas_tbl)

# tampilkan hasil
panas_chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  panas_tbl
## X-squared = 179.31, df = 5, p-value < 2.2e-16
```

Keluaran dari `chisq.test()` terdiri atas beberapa unsur, di antaranya (i) `statistic` yang menunjukkan nilai χ^2; (ii) `parameter` untuk nilai *df*; (iii) `p.value`, yaitu p < 2.2e-16[4] pada tampilan hasil di atas; (iv) `observed` (atau F_o); (v) `expected` (atau F_e); (vi) `residuals`, yang merupakan nilai *Pearson residuals* (Gambar 2). Isi dari tiap-tiap unsur tersebut dapat diperoleh dengan menggunakan tanda mata uang dolar (`$`) sebagai berikut:

```
# tampilkan nilai chi-square
panas_chi$statistic
```

```
## X-squared
##  179.3101
```

```
# tampilkan nilai probabilitas
panas_chi$p.value
```

```
## [1] 7.512214e-37
```

```
# tampilkan tabel kekerapan harapan
panas_chi$expected
```

```
##
##       dipanasi dipanaskan  memanas memanasi memanaskan    panas
##   lit 14.017467   69.38646 67.98472 30.83843   68.68559 70.08734
##   met  5.982533   29.61354 29.01528 13.16157   29.31441 29.91266
```

```
# tampilkan tabel nilai Pearson residuals
panas_chi$residuals
```

```
##
##         dipanasi dipanaskan    memanas   memanasi memanaskan      panas
##   lit  1.0637133  3.1949585 -6.1834999 -0.3310553  0.8825646  1.7812934
##   met -1.6282340 -4.8905472  9.4651302  0.5067488 -1.3509484 -2.7266392
```

Nilai kontribusi terhadap χ^2 untuk tiap-tiap sel merupakan hasil pangkat 2 dari nilai *Pearson residuals* tiap-tiap sel:

```
# hitung kontribusi terhadap chi-square
panas_chi$residuals^2
```

```
##
##         dipanasi dipanaskan    memanas   memanasi memanaskan      panas
##   lit  1.1314859 10.2077600 38.2356716  0.1095976  0.7789203  3.1730060
##   met  2.6511459 23.9174522 89.5886904  0.2567943  1.8250615  7.4345616
```

Selanjutnya, nilai kadar efek *Cramér's V* dihitung dengan kode berikut:

```
# ambil nilai chi-square
# dan hilangkan label nama
chistats <- panas_chi$statistic
chistats <- unname(chistats)

# ambil nilai terkecil dari jumlah baris dan kolom `panas_tbl`
min_dim <- min(nrow(panas_tbl), ncol(panas_tbl))

# total pengamatan "N"
N <- sum(panas_tbl)

# hitung Cramér's V berdasarkan rumus pada §3.3.2 di atas
# lalu bulatkan hasil menjadi 3 angka desimal
cramersv <- sqrt(chistats/(N * (min_dim - 1)))
round(cramersv, digits = 3)
```

```
## [1] 0.626
```

PENUTUP

Makalah ini telah mengulas pemahaman kuantitatif mendasar terhadap uji signifikansi *Chi-Square* dalam mengkaji hubungan antara bentuk dan makna pada kasus distribusi makna (non-)metaforis medan leksikal PANAS dalam bahasa Indonesia. Dalam kajian ini ditemukan bahwa terdapat hubungan yang sangat signifikan dan kuat antara realisasi morfosintaksis medan leksikal PANAS dan makna literal serta metaforisnya (§3.3.2). Dengan menggunakan metode kuantitatif, ditunjukkan pula bahwa hubungan antara bentuk dan makna bukan merupakan soal hitam-putih, tetapi bersifat gradien. Sifat ini sebaiknya dicirikan berdasarkan distribusi statistik. Contohnya, meskipun *memanas* bisa digunakan dengan makna literal, tetapi bentuk inkohatif ini lebih condong dari yang diharapkan untuk digunakan secara metaforis. Sebaliknya, bentuk pasif kausatif *dipanaskan* jauh lebih lazim digunakan secara literal (Gambar 2). Kecenderungan statistik seperti ini akan sulit untuk diungkap dengan mengandalkan metode introspeksi. Pertanyaan terkait mengapa bentuk-bentuk tersebut memiliki perilaku semantis yang berbeda belum dapat dijawab kali ini. Setidaknya, hasil analisis ini dalam bahasa Indonesia mendukung temuan serupa dalam bahasa Inggris (Deignan, 2006) terkait hubungan bentuk gramatikal suatu kata dengan perbedaan proporsi pemakaian metaforis dan literalnya. Kajian ini juga telah (i) memberikan bukti empiris terkait asumsi bahwa bentuk yang berbeda memiliki batasan semantis yang berbeda (§1) dan (ii) menunjukkan bagaimana asumsi tersebut dapat dirumuskan berdasarkan distribusi kuantitatif dalam korpus.

Penerapan uji statistik analitik (*analytical statistics*) dapat menentukan tingkat keumuman (*generalisation*) temuan berdasarkan sampel terhadap keseluruhan populasi, dalam hal ini semua populasi pemakaian medan leksikal PANAS yang dikaji, khususnya pada jenis teks daring yang mendasari korpus IWaC. Namun, jenis teks dari korpus yang digunakan merupakan salah satu faktor yang belum dapat dipertimbangkan dengan saksama. Hasil berbeda dari makalah ini bisa muncul jika menggunakan korpus dengan tema berbeda (misalnya novel). Selain perihal korpus, keterkaitan signifikan dan kuat yang ditemukan juga terbatas pada kata-kata dari satu jenis medan leksikal semantik, yaitu PANAS; kajian sistematis terhadap jenis medan leksikal lainnya masih perlu dilakukan. Terlepas dari batasan-batasan tersebut, makalah ini sedikitnya telah mencontohkan bagaimana salah satu pertanyaan mendasar dalam linguistik dapat dikaji menggunakan metode kuantitatif, khususnya *Chi-square*. Peta kelinguistikan setakat ini telah bergerak ke arah komputasional, kuantitatif, dan mengandalkan data tekstual yang besar sehingga pemahaman terhadap metode kuantitatif dan peranti komputasional, seperti R, menjadi investasi berharga bagi generasi peneliti bahasa abad ke-21.

CATATAN

[*]Penulis berterima kasih kepada mitra bebestari atas dukungan dan masukan yang baik untuk meningkatkan kualitas makalah ini. Penulisan makalah ini didukung oleh dana penelitian Doktoral dari Monash University, Australia, yang diberikan kepada Gede Primahadi Wijaya Rajeg: *Monash International Postgraduate Research Scholarships* (MIPRS) & *Monash Graduate Scholarships* (MGS).

DAFTAR PUSTAKA

Arka, I. W. (2017). The core-oblique distinction in some Austronesian languages of Indonesia and beyond. *Linguistik Indonesia*, *35*(2), 101–144. doi:10.26499/li.v35i2.58

Aryawibawa, I. N., & Ambridge, B. (2018). Is Syntax Semantically Constrained? Evidence From a Grammaticality Judgment Study of Indonesian. *Cognitive Science*, 1–14. doi:10.1111/cogs.12697

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK ; New York: Cambridge University Press.

Deignan, A. (2006). The grammar of linguistic metaphors. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 106–122). Berlin: Mouton de Gruyter.

Denistia, K., & Baayen, R. H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*. https://doi.org/10.1007/s11525-019-09340-7

Denistia, K., Bajestan, E. S., & Baayen, R. H. (2018, September). *A semantic vector model for the Indonesian prefixes pe- and peN-*. Presented at the 11th International Conference on the Mental Lexicon, Edmonton, Alberta, Canada. Retrieved from https://mentallexicon2018.ca/wp-content/uploads/2018/09/MenLex2018_AbstractBooklet.pdf

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

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*.

Moeljadi, D. (2011). Possessive verbal predicate constructions in Indonesian. *Tokyo University Linguistic Papers*, *31*, 117–133. Retrieved from http://compling.hss.ntu.edu.sg/who/david/papers/TULIP31_davidmoeljadi.pdf

Moeljadi, D. (2014). Usage of Indonesian possessive verbal predicates: A statistical analysis based on storytelling survey. *Tokyo University Linguistic Papers*, *35*, 155–176. Retrieved from http://compling.hss.ntu.edu.sg/who/david/papers/TULIP35_davidmoeljadi.pdf

Musgrave, S. (2013). Functional categories in the syntax and semantics of Malay. *NUSA*, *55*, 135–152. Retrieved from http://hdl.handle.net/10108/74330

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/

Rajeg, G. P. W. (2014). Metaphorical profiles of five Indonesian quasi-synonyms of ANGER: Multiple distinctive collexeme analysis. In *Proceedings of the International Congress of the Linguistic Society of Indonesia 2014* (pp. 165–170). Bandar Lampung, Sumatra, Indonesia: Masyarakat Linguistik Indonesia (MLI). doi:10.4225/03/58578ddba1fd2

Rajeg, G. P. W. (2016). Exploring the semantics of near-synonyms via metaphorical profiles: A quantitative corpus-based study of Indonesian words for HAPPINESS. In *Proceedings of the International Congress of The Linguistic Society of Indonesia* (pp. 261–265). Universitas Udayana, Bali-Indonesia: Masyarakat Linguistik Indonesia (MLI) & Universitas Udayana. doi:10.4225/03/5913aec719240

Rajeg, G. P. W. (2018). Happyr: The accompanying R package for Rajeg's (2018) PhD thesis titled "Metaphorical profiles and near-synonyms: A corpus-based study of Indonesian words for Happiness" (Version 0.1.0). doi:10.5281/zenodo.1436331

Rajeg, G. P. W., & Rajeg, I. M. (2017). Mempertemukan morfologi dan linguistik korpus: Kajian konstruksi pembentukan kata kerja [*per*-+Ajektiva] dalam Bahasa Indonesia. In I. N. Sudipa & M. S. Satyawati (eds.), *Rona Bahasa: Buku persembahan kepada Prof. Dr. Aron Meko Mbete memasuki masa purnatugas* (pp. 288–327). Denpasar, Bali, Indonesia: Swasta Nulus. doi:10.4225/03/5a0627de02453

Rajeg, G. P. W., & Rajeg, I. M. (2019a). Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. *Etika Bahasa: Buku Persembahan Menapaki Usia Pensiun I Ketut Tika*. forthcoming. doi:10.31227/osf.io/uwzts

Rajeg, G. P. W., & Rajeg, I. M. (2019b, February 28). Data and R Markdown Notebook for *Pemahaman kuantitatif dasar dan penerapannya dalam mengkaji keterkaitan antara bentuk dan makna* (Version 2). figshare. doi:10.26180/5c6e1160b8d8a

Rajeg, G. P. W., Denistia, K., & Musgrave, S. (2018, May). *Semantic vector space model and the usage patterns of Indonesian denominal verbs with meN-, meN- -Kan, and meN- -i affixes*. Presented at the Twenty-Second International Symposium on Malay/Indonesian Linguistics (ISMIL 22), The University of California, Los Angeles. doi:10.4225/03/5acffc60eb649

Rajeg, G. P. W., Denistia, K., & Rajeg, I. M. (2018). Working with a linguistic corpus using R: An introductory note with Indonesian negating construction. *Linguistik Indonesia*, *36*(1), 1–36. doi:10.4225/03/5a7ee2ac84303

Rajeg, I. M. (2014). Metafora spesifik emosi Bahasa Indonesia: Kajian linguistik korpus. In *Proceedings of the International Congress of the Linguistic Society of Indonesia 2014* (pp. 209–213). Bandar Lampung, Sumatra, Indonesia: Masyarakat Linguistik Indonesia (MLI). Retrieved from https://lib.atmajaya.ac.id/default.aspx?tabID=61&src=a&id=278166

Siahaan, P. (2011). HEAD and EYE in German and Indonesian figurative uses. In Z. A. Maalej & N. Yu (Eds.), *Embodiment via Body Parts: Studies from various languages and cultures* (pp. 93–114). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Siahaan, P. (2015). Why is it not cool? Temperature terms in Indonesian. In M. Koptjevskaja-Tamm (Ed.), *The Linguistics of Temperature* (pp. 666–699). Amsterdam: John Benjamins Publishing Company.

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

Stefanowitsch, A. (2010). Empirical cognitive semantics: Some thoughts. In Dylan Glynn & Kerstin Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 355–380). Berlin: Mouton de Gruyter.

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, *1*(2), 225–261.

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

Wickham, H., & Grolemund, G. (2017). *R for Data Science*. Canada: O'Reilly. Retrieved from http://r4ds.had.co.nz/

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] Nama variabel atau kategori ditulis dengan huruf KECIL CAPITAL.

[2] Pranala untuk tutorial oleh Stefanowitsch (2004) sudah tidak aktif sejak tahun 2012, namun penulis masih menyimpan pindaian PDF laman tersebut.

[3] Unsur *chi* dilafalkan *ky*, yang berima sama dengan *high* atau *fly* dalam Bahasa Inggris (Stefanowitsch, 2004).

[4] Nilai 2.2e-16 merupakan tampilan ilmiah untuk bilangan desimal 0.00000000000000022 yang mengandung 15 angka 0 setelah koma dan diikuti angka 22.