International Journal of Linguistics and Discourse Analytics
Vol. 6, No. 1, September 2024
P-ISSN 2721-8899 E-ISSN 2721-8880

APPLYING CHI-SQUARE TEST IN MEASURING THE SIGNIFICANCE OF THE OCCURRENCE OF FRENCH SYNONYM IN CORPUS DATA

Putu Weddha Savitri*, Ni Luh Sutjiati Beratha, I Nengah Sudipa, I Made Rajeg
Udayana University, Bali, Indonesia, Indonesia
weddha_savitri@unud.ac.id*

ABSTRACT

This paper explains how to apply the significance test of a synonym set in French to its appearance in a data corpus. The chi-square significance test is a procedure that can be used to test the significance of quantitative data statistically. The object of this research is a series of adjective synonyms with the core meaning 'extraordinary' in French. This research uses Leipzig corpora collection as data source, and AntConc (version 4.2.0) as a tool used in searching for the frequency of occurrence of each synonymous word in the news and website data corpus. The results of the significance test show that distributions differences between the real frequency of occurrence and the expected frequency of occurrence of synonymous words in the data corpus can be considered as not just a coincidence. With a p value < 0.001, it can be concluded that there is a significant relationship between the differences in the distribution of each adjective in different types of data corpus.

Keywords: Adjective, French, Chi-Square, Corpus Data, Near-Synonym, Synonymy

I. Introduction

Synonymy is an interesting topic in the study of lexical semantics. It describes the meaning relationship between a set of vocabularies that have similar meanings to each other. Palmer (1981) states that synonymy refers to the idea of "similarity of meaning", while Lyons (1995) states that synonymy is an expression with the same meaning (including complex lexical expressions). This research follows what Cruse (1986) said, namely that synonymous forms must not only have mutually encompassing meanings (a high degree of semantic overlap) but must also have low constructive meanings.

Previously, the work of distinguishing the meaning and use of synonymous words took up quite a lot of research time for lexicographers. However, along with the development of digital technology, research on synonymy can be made easier by the existence of data corpus which can quickly provide large amounts of data and can be used as a valid basis for studying real patterns of natural use of words in texts (Biber, Conrad and Reppen, 1998).

Corpus linguistic studies are currently developing rapidly and are believed to be able to answer questions related to linguistic phenomena in a more empirical way (Rajeg, 2020). In research related to synonymy, a text corpus will be very useful for knowing the distribution of use of certain words in certain environments. Various studies on synonymy using corpus linguistic methods have proven that data in text corpus makes it possible to identify and investigate patterns of word usage and their distribution : Gries and Divjak (2006), Liu (2010), Cai (2012), are several studies that use the same method in their analysis.

In particular, the adjective with the core meaning of 'extraordinary' in French is very interesting to be analysed because of its high frequency of use in speech. There are 11 words with the core meaning of 'extraordinary' and with the large number of synonymous words, it can be assumed that each adjective carries certain semantic properties that differentiate one word from another. Like synonymous words in general, they show an interesting phenomenon, such as having varying frequency of occurrence in two different data corpora even though they are in the closest synonym series.

The corpus linguistic approach is often based on a quantitative approach that uses statistical methods in data processing to support the analysis of a linguistic phenomenon. This paper uses a quantitative approach to show differences in the distribution of synonymous French adjectives in two different types of text corpus, namely news corpus and website. The initial assumption is that certain words tend to appear in certain varieties of language (formal or informal). For this reason, a statistical method is used to measure the relationship between two variables, namely the choice of synonymous words and the type of text corpus in which the words tend to appear.

To ensure that the difference in the frequency of occurrence of each adjective is not a coincidence but rather that there is a relationship between the 2 observed variables, it is necessary to carry out a significance test called the *Chi-Square test*, denoted by $\chi 2$. The main reason for using this test is to measure the relationship between the two observed variables, as has been done in the research of Liu (2010), and Rajeg (2019) which became the guidance in this chi-square test. Therefore, the hypothesis that we want to test with the *chi-square test* is whether there is a relationship between the use or appearance of certain adjectives and the types of corpus sources available. In other words, statistical significance in this study refers to the chance of finding differences in the distribution of the use of synonymous adjectives in two types of text corpus, news and website.

II. Methods

The analysis in this paper is a corpus-based approach that uses quantitative methods to obtain answers to the questions asked. The object of this research is the synonymy of French adjectives with the core meaning "extraordinary". French adjectives themselves are quite unique

because, unlike English or Indonesian, their use is greatly influenced by the nouns they describe, especially regarding gender and number of the noun.

After searching in www.synonyms-fr.com with the keyword *merveilleux*, there are more than 20 words that are synonymous. However, this study only took 11 words randomly because they were considered representative and did not reduce the essence of this analysis. These adjectives are *Magnifique, Superbe, Merveilleux, Remarquable, Fantastique, Formidable, Extraordinaire, Fabuleux, Excellent, Incroyable,* And *Étonnant*. The corpus used as the data source for this paper is the *Leipzig Corpora Collection* (D.Goldhahn, 2012) in French language, by downloading two different types of text corpus, one sourced from news and the other sourced from websites. These two types of sources are considered different in terms of language variety, where news texts use a formal language style while texts on websites usually use less formal language. A tool, namely *AntConc* version 4.2.0, was used to search for quantitative data regarding the frequency of occurrence of each adjective.

III. Findings and Discussion

A descriptive presentation regarding the distribution of the 11 synonymous adjectives is started by finding for the frequency of occurrence to find out how often these synonymous adjectives appear in the data corpus. The search for the frequency of occurrence was carried out on *AntConc* tool by entering the two types of corpus data (news and websites) which had been previously downloaded from the *Leipzig Corpora Collection*. This was done to see whether there were differences in the frequency of occurrence for each adjective in two different types of data corpus.

The next step is to enter *keywords in context* (the words you want to search for) to get the frequency of occurrence of each adjective. In this case, it is necessary to make adjustments to the advance search feature because French adjectives have more than 1 lemma in accordance on the gender and number of the noun they describe. For example, the adjective étonnant has 4 lemmas, namely *étonnant* (masculine/singular), *étonnante* (feminine/singular), *étonnants* (masculine/plural), and *étonnantes* (feminine/plural).

Based on the frequency of occurrence of synonymous adjectives, it appears that there are variations in their occurrence in the two types of text. A comparison graph regarding the distribution of use/frequency of occurrence of synonymous adjectives in the two types of data corpus can be seen in Figure 1 below.

The adjective *magnifique* is the adjective that appears most often in website text, while the *étonnant* is the adjective that appears most often in the news corpus. It can also be seen that the adjective *épatant* has a very low frequency of occurrence when compared to other adjectives in both corpora. For more details, the frequency of occurrence of each adjective in the two types of data corpus, namely the news corpus and website, can be seen in table 1.

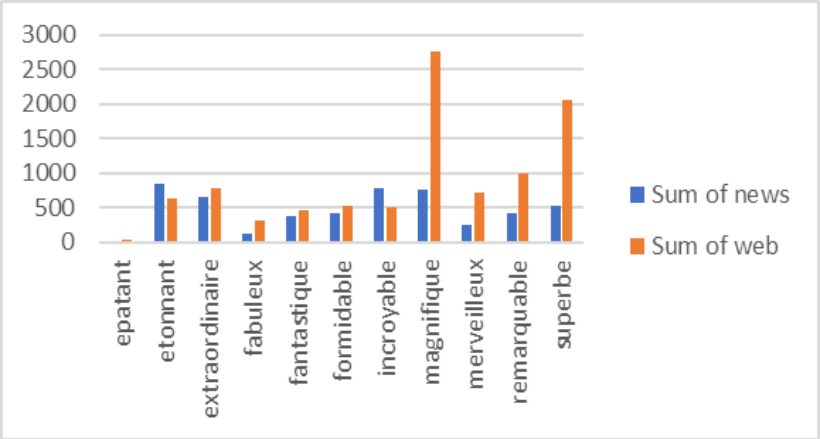

Figure 1. Graph of frequency of occurence

Table 1. Frequency of occurrence of synonymous adjectives

Word	News	Web	Total
magnifique	753	2756	3509
superbe	521	2065	2586
etonnant	850	635	1485
extraordinaire	652	788	1440
remarquable	422	987	1409
incroyable	779	510	1289
merveilleux	247	721	968
formidable	426	517	943
fantastique	366	462	828
fabuleux	127	322	449
epatant	22	24	46
Total	5165	9787	14952

From the table above it can be explained that the adjective *magnifique* appears 753 times in the news data corpus and 2756 times in the website data corpus, so that the total appearance of the adjective magnifique in the two data corpuses is 3509 times and it is also the adjective that has the highest frequency of appearance. Meanwhile, the adjective *épatant* appears 22 times in the news data corpus and 24 times in the web data corpus, so the total appearance of this adjective is only 46 times in the entire data corpus which makes it the adjective with the lowest frequency of occurrence.

Based on the data in the table above, it can be seen that there are differences in the frequency of occurrence of all adjectives so it can be assumed that there are other aspects that influence the use of certain adjectives. To ensure that the difference in the frequency of

occurrence of each adjective is not a coincidence but rather that there is a relationship between the 2 observed variables, it is necessary to carry out a significance test called the *Chi-Square test*. To start testing the *Chi-square test*, we first have to find the expected frequency of occurrence based on the null hypothesis, or what is known as the expected frequency or abbreviated as F_e. In this case, F_e is the frequency that is expected to appear if there is no relationship between the use of a particular adjective and the source of the data corpus (Rajeg, 2019). The *Chi-Square test* involves the expected frequency and the observed frequency in the data sample or what is called the real frequency (observed frequency or abbreviated as F_o). The general formula for calculating F_e from each cell (E_{ij}) is as follows

$$E_{ij} \frac{S_i \times S_j}{N}$$

S_i shows the total frequency of row *i*, while S_j shows the total frequency of column *j*, and *N* is the total observations in the sample. For example, the F_e calculation for the cell at the intersection of the *magnifique* row and the news column is 3509 x 5156, then divided by 14952 to get a result of 1212.14. Another example of calculating F_e for the intersection cell of the *extraordinaire* row and web column is 1440 x 9797, then divided by 14952 to get the result 942.57. Complete results of F_e calculations from each cell can be seen in the table 2

Table 2. Expected frequency of occurrence

Word	News	Web	Total
magnifique	1212,14	2296,86	3509,00
superbe	893,30	1692,70	2586,00
etonnant	512,98	972,02	1485,00
extraordinaire	497,43	942,57	1440,00
remarquable	486,72	922,28	1409,00
incroyable	445,27	843,73	1289,00
merveilleux	334,38	633,62	968,00
formidable	325,75	617,25	943,00
fantastique	286,02	541,98	828,00
fabuleux	155,10	293,90	449,00
epatant	15,89	30,11	46,00
Total	5165,00	9787,00	14952,00

After getting the F_e value for each cell, we can observe which cells have a lower or higher frequency of appearance than expected. For example, the frequency of occurrence of the adjective *magnifique* is 753 times based on observations in the news corpus, while the expected

frequency is 1212.14. So it can be concluded that the adjective *magnifique* appears less frequently than expected in the news corpus, and conversely, in the website corpus, the adjective *magnifique* appears more often than expected. The expected frequency of appearance was 2296.86, but in reality this adjective appeared 2756 times higher in the website corpus.

The next step is to determine the level of difference between F_o and F_e to observe that the frequency observed in the sample is not a coincidence. The chi-square statistical value (χ^2) is the sum of the contribution values of each cell to χ^2 (Gries, 2013: 168). The following is the formula for calculating the χ^2 value for each cell (Levshina, 2015; Stevanofich, 2013; Rajeg, 2019)

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

This formula shows that the χ^2 value is obtained from the sum of each cell by calculating the power of the difference between F_o dan F_e divided by F_e. or example, the contribution value to χ^2 from the *magnifique+berita* confluence cell is (753-1212,14)2/1212,14 with result 173,92. Another example for the contribution value to χ^2 from the *magnifique+website* confluence cell is (2756-2296,86)2/2296,86 which is equal to 91,78. Similar calculations are carried out on all cells and then get the sum of all cells to form the χ^2 statistical value. The total quantity of all cells in table 3 below is the *chi-square test* value, that is 1437.25

Table 3. Contribution value to χ2

Kata	*Korpus berita*	*Korpus web*
magnifique	173,92	91,78
superbe	155,17	81,89
etonnant	221,42	116,85
extraordinaire	48,03	25,35
remarquable	8,61	4,54
incroyable	250,13	132,00
merveilleux	22,84	12,05
formidable	30,85	16,28
fantastique	22,36	11,80
fabuleux	5,09	2,69
epatant	2,35	1,24

A distribution is said to be statistically significant if the significance level, called the *p-value*, is smaller than 5% or usually written with a decimal number of 0.05 (Gries, 2013:27). *P-value* is a measure used in statistics to evaluate how strong the evidence the data has in

supporting the null hypothesis (H_0). The null hypothesis is the assumption that there is no significant effect or relationship between the variables being tested. In general, the lower the *p-value*, the stronger the evidence supports rejecting the null hypothesis, conversely, if the *p-value* is large, it does not provide strong enough evidence to reject the null hypothesis. A *p-value* of less than 5% or 0.05 is often considered a general cutoff for statistical significance. This means that if the *p-value* is less than 0.05 then the results are considered statistically significant and the null hypothesis is usually rejected.

However, before determining the *p-value*, we need to find the degrees of freedom. Degrees of freedom (abbreviated *df*) is a concept in statistics that refers to the number of values that are free to vary in a particular statistical calculation after taking into account existing constraints or settings. In short, df refers to changing values (Rajeg, 2019). In general, degrees of freedom are calculated by the formula:

$$Df = (\text{N rows} - 1) \,.\, (\text{N colums} - 1)$$
$$Df = (11 - 1) \,.\, (2 - 1) = 10$$

The value of $df = 10$ with a fixed total marginal frequency shows that only 10 cells from table 3 have values that can be changed without changing the total marginal frequency. The classic way to determine the level of significance of a distribution is to compare the χ^2 value obtained, $\chi^2 = 1437.25$, with the critical value from the χ^2 value distribution table as shown in Table 4 below (Gries, 2013:184)

Table 4. χ^2 value for the *p-value* significance level

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Critical values of the Chi-square distribution with *d* degrees of freedom

Next, to find out the significance of the distribution of adjectives and text types in the sample, we can look at the row that shows *df* = 10 and see whether the value of χ^2 = 1437.25 obtained from the sample is higher than the value of χ^2 in that row (Figure 2). It can be seen that, when compared, the χ^2 value obtained in the sample is much higher than the value in the 0.1% significance column, namely p = 0.001 with χ^2 = 29.588. This shows that there is a very small chance (smaller than 0.1% or p < 0.001) to find differences in the distribution of each adjective in a particular type of text if the difference is considered a coincidence and there is no connection between the use of certain adjectives in the sample text

IV. Conclusion

As previously stated, differences in the distribution of a word or phrase in the data corpus can lead us to an answer to the linguistic phenomena observed. Corpus linguistics as an approach that is currently developing rapidly is believed to be able to answer problems in language in a more empirical way. In relation to this research, differences in the distribution of a set of adjectives that are synonymous in French in two different text corpora, namely news corpus and website, have been tested statistically using the *chi-square* test to show that there is a relationship between the two variables involved. Based on the calculation of the *chi-square* statistic test with a p-value of less than 0.1% (p < 0.001), it can be concluded that there is a very significant relationship between the two variables tested or it could be said that it is not a coincidence if there is a difference in the distribution of each adjective with different types of text (formal and less formal).

References

Biber, Douglas, Susan Conrad and Randi Reppen. (1998). *Corpus Linguistics: Investigating Languagestructure and use*. Cambridge: Cambridge University Press.

Cruse, D. Alan. (2006). *Meaning in Language: An Introduction into Semantiks and Pragmatic*. Oxford University Press.

D. Goldhahn, T. Eckart & U. (2012). Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *In: Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*

Dixon, R.M.W. (2010). *Basic Linguistic Theory: Volume 2 Gramatical Topics*. New York: Oxford University Press

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

Facchinetti, R. (2007). *Corpus Linguistisc 25 Years On*. Amsterdam – New York: Rodopi

Inkpen, D. & Hirst, G. (2002). Building and using a lexical knowledge base of near synonym differences. *Computational Lingusitics. 32*(2), 223-262

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

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge.

Moon, R. (2010). *What can a corpus tells us about lexis? In: A O'Keffe and M. McCarthy (Eds)*. The Routledges handbook of corpus linguistics. (pp.197-211). Oxford: Routledge

Palmer, F. R. (1981). *Semantics*. London: Cambridge University Press

Rajeg, G. P. W., & Rajeg, I. M. (2019). Pemahaman Kuantitatif Dasar Dan Penerapannya Dalam Mengkaji Keterkaitan Antara Bentuk Dan Makna. *Linguistik Indonesia*, *37*(1), 13–31. https://doi.org/10.26499/li.v37i1.87

Rajeg, G. P. W. (2020). Linguistik Korpus Kuantitatif Dan Kajian Semantik Leksikal Sinonim Emosi Bahasa Indonesia. *Linguistik Indonesia, 38*(2)

Sinclair, J. (1991). *Corpus, Concordance, Collocation: Describing English language*. Oxford: Oxford University Press.

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

https://synonyms-fr.com

https://synonyms-fr.com

https://corpora.uni-leipzig.de/