EasyChair Preprint

№ 418

Modeling Non-Compositional Expressions using a Search Engine

Cheikh Bamba Dione and Christer Johansson

August 9, 2018

Modeling Non-Compositional Expressions using a Search Engine

Cheikh Bamba Dione
Dept. of Linguistic, Literary and Aesthetic Studies
University of Bergen, Norway
Dione.Bamba@uib.no

Christer Johansson
Dept. of Linguistic, Literary and Aesthetic Studies
University of Bergen, Norway
Christer.Johansson@uib.no

***Abstract*— Non-compositional multi-word expressions present great challenges to natural language processing applications. In this paper, we present a method for modeling non-compositional expressions based on the assumption that the meaning of expressions depends on context. Therefore, context words can be used to select documents and separate documents where the expression has different meanings. Deviation from a baseline is measured using serendipity (i.e. the pointwise effect size). We used this statistical measure to mark which patterns are over- and under-represented and to take a decision if the pattern under scrutiny belongs to the meaning selected by the context words or not. We used the Google search engine to find document frequency estimates. When used with Google document frequency estimates, the serendipity measure closely mirrors some human intuitions on the preferred alternative.**

I. Introduction

Multiword expressions (MWEs) like idioms and collocations co-occur so often that they are perceived as a linguistic unit. MWEs typically have a non-compositional meaning, i.e. their meanings usually cannot be determined compositionally from the meanings of the individual words. Therefore, they present a great challenge to natural language processing applications and are sometimes referred to as "a pain in the neck" [1]. The identification of non-compositional MWEs is a crucial subtask for many computational systems. For instance, machine translation systems need to know if a sequence of words can be translated word by word or if it has a special meaning, which requires a particular translation. Their ubiquity in language may affect the performance of tasks like parsing and word sense disambiguation [2].

MWEs "are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" [3]. This definition is useful, as "it states that some form of idiomaticity is a necessary feature of MWEs" [4]. The term idiomaticity expresses the markedness or deviation from the linguistic properties of the component words. This notion applies at the lexical, syntactic, semantic, pragmatic and/or statistical levels, as layed out in [5, p.269ff].

For a MWE, *lexical idiomaticity* occurs when one or more of its lexical components are not part of the conventional lexicon of the language in question. For instance, the expression *ad hoc* is lexically marked since none of its components (*ad* and *hoc*) are standalone English words. *Ad hominem*, and *post hoc (ergo procter hoc)* are examples of Latin expressions that are used with specific meanings in English, without fully combining with other words.

Syntactic idiomaticity means that the syntax of the MWE is not derived from that of its components. For example, the adverbial phrase "by and large" is made up of the anomalous coordination of a preposition (*by*) and an adjective (*large*).

Semantic idiomacity occurs when the meaning of an MWE is not fully derivable from its component parts. For instance, the expression *middle of the road* typically refers to "non-extremism, especially in political views"[5]. This expression is semantically marked in the sense that its meaning could not be readily predicted from the meaning of its constituent words (*middle* and *road*).

Pragmatic idiomaticity means that a MWE is associated with particular communicative situations or discourse contexts. For example, the expression *good morning* is a greeting used without irony at the start of the day and it does not necessarily mean that the morning is good in any of the senses of "good".

Statistical idiomacity refers to the situation where particular combinations of words occur with markedly higher frequency in comparison to the component words or alternative phrasings of the same concept. For instance, it would be perfectly correct (linguistically speaking) to say *computer translation* instead of *machine translation*. However, statistically we find the particular lexicalisation *machine translation* far more frequent than *computer translation*. *Statistical idiomacity* includes notions such as collocations (e.g. *powerful car*) vs. anti-collocations (e.g. *strong car*), as can be noticed reversed in *strong breath* vs. *powerful breath*.

In this paper, we propose a statistical method that identifies non-compositional expressions based on the assumption that these expressions deviate from some expectations. Such a deviance is measured using serendipity, as introduced by [6]. Serendipity is the pointwise effect size (i.e. the effect size per cell), and signals that particular observations or events deviate from expectations.

A. Related Work

In recent past, several attempts have been made to address the automatic classification of MWEs. One method [7] proposed to compare the mutual information between the constituents of a non-compositional phrase and that of a phrase created by substituting the constituents of that phrase with their similar words. The assumption is that the mutual information score of the former phrase is significantly different from that of the latter phrase. However, this system [7] scored low on both precision (39%) and recall (21%).

Another attempt [8] used distributional semantics and latent semantic analysis (LSA) to developed a model that uses the local linguistic context to discriminate the non-compositional (or idiomatic) from its literal use of a MWE. Their model also analyses the contexts of a MWE and the context of the MWE components, which reveals that such a context comparison can be an important factor for distinguishing idiomatic from literal use of an expression.

Our final example is a framework [9] based on distributional vector-space models ("machine learning") developed to learn and detect semantic non-compositionality of English noun compounds. Non-compositional compounds are defined in that work (ibid.) as those compounds that are not well modeled (i.e., marked as outliers) by the learned semantic composition function. It is argued (ibid.) that polynomial projection and neural networks can model semantic composition more effectively compared to the previous approaches based on multiplicative or additive functions.

B. *Overview*

The paper is structured as follows. Section II presents the *serendipity* statistical measure as used in this article. Section III discusses issues related to the use of Google frequency estimates. Section IV discusses some examples used in this study. Section V discusses the results achieved and puts them in context. Section VI concludes the discussion.

II. Serendipity

A. *Cross table analysis*

For a better understanding of the serendipity measure, we need to briefly discuss some prerequisites related to statistical independence, including cross tables, statistical significance and effect size.

Cross tables group variables to understand the correlation between the different variables. A cross table analysis, also known as contingency table analysis, is an analysis of frequency that tests whether rows and columns are statistically independent of each other. The most common type of cross table is a 2 × 2 table:

TABLE I

Example of a cross-table

a	c	R_1
b	d	R_2
C_1	C_2	T

R_1 and R_2 represent the total frequencies for row 1 and row 2. R_1 is obtained by adding a to c. Likewise, R_2 is obtained by adding b and d. C_1 (=a+b) and C_2 (=c+d) represent the total frequencies for column 1 and column 2, and T is the sum total of all the cells. The null hypothesis (i.e. the "expectation") is that the rows and the columns are independent of each other. Consequently, the expected probability of belonging to row 1 is R_1/T. Likewise, belonging to row 2 is R_2/T, belonging to column 1 is C_1/T, and belonging to column 2 is C_2/T. Simultaneously belonging to row 1 and column 1 would be $\frac{R_1}{T}\frac{C_1}{T} = \frac{R_1 C_1}{T^2}$

B. *The chi-square Statistic*

If the assumption of independence holds, the probability of belonging to a cell can be obtained by multiplying the probabilities of the corresponding column and row. Once we have the cell probabilities we get the expected frequencies by distributing the total across all cells. Table II shows the expected frequencies of each cell, after multiplying with the grand total T, and after simplification.

TABLE II

The expected frequencies

$E_{11} = \frac{R_1 \cdot C_1}{T}$	$E_{12} = \frac{R_1 \cdot C_2}{T}$
$E_{21} = \frac{R_2 \cdot C_1}{T}$	$E_{22} = \frac{R_2 \cdot C_2}{T}$

Significant deviations from independence can be assessed by the difference between the observed and expected frequencies. A positive difference indicates that the cell is *over-represented.* A negative difference indicates that the cell is *under-represented.* The test of significance works as follows: for each cell, compute the difference between observed and expected frequencies, square that difference and divide it by the expected frequency. Finally, the values for all cells are added together. This gives the χ^2 statistic, which tells if the observed deviances from the expected deviances are explained by random chance or not.

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{1}$$

C. *Effect Size*

The χ^2 statistic answers the question of how likely the rows and columns deviate from statistical independence. How large the effect is has to be calculated by the effect size (ϕ). For a 2 × 2 contingency table, the ϕ coefficient is calculated as $\phi = \sqrt{\frac{\chi^2}{N}}$. If either the rows or the columns are greater than 2, then the ϕ coefficient can be generalized using Cramér's $v = \sqrt{\frac{\chi^2}{df * N}}$, where *df* (the degree of freedom) is the smallest number of either $rows - 1$ or $columns - 1$.

Any observed difference is assumed to be detected by sampling variability [10], and with a sufficiently large sample, a statistical test will find a significant difference unless the effect size is exactly zero. But very small differences, even if significant, are often meaningless. For example, if a sample size is 10 000, we can detect, with significance, a 1% difference from the expected frequencies for two groups. If that difference is important or not depend on what we measure and which questions we ask. The effect size is related to the question of how many observations we need in order to find out if the observed deviances are significant. If we only need to make a few observations this could be an important difference.

Unlike significance, effect size is independent of sample size. ϕ is an important tool in reporting and interpreting the importance of a found difference.

D. Serendipity

Serendipity is the contribution of each cell to the effect size. Following [6], it can be computed by comparing the cell's contribution to the χ^2 statistic with the overall χ^2. This will give, for each cell, its proportional contribution to the effect size. Serendipity in each cell i can be computed using the formula given in (2), where ϕ and χ^2 are numbers and $\frac{O_i-E_i}{E_i}$ are terms in a series, and n is the number of cells. The proof involves multiplying ϕ with 1, and rewriting that 1 as $\frac{\chi^2}{\chi^2}$, and expanding the upper χ^2 using the definition formula. Obviously, $\frac{\phi}{\chi^2}$ can be multiplied into the sum, to give the effect contribution per cell.

$$\phi * 1 = \frac{\phi\chi^2}{\chi^2} = \frac{\phi}{\chi^2}\sum_{i=1}^{n}(\frac{(O_i - E_i)^2}{E_i}) \qquad (2)$$

An add-one solution can be applied to handle low cell frequencies: add $\frac{1}{n}$ to each cell (this will sum to 1) and divide by (χ^2+1) rather than χ^2. This gives some probability space for the non-observed.

The definition [6] of the (signed) serendipity function in the R programming language is given below. Note the multiplication with 100 for readability.

```
serendipity <- function (x){
  df <- min(nrow(x),ncol(x))
  if (df>1) df <- df - 1
  model <- chisq.test(x,correct=F)
  phi <- sqrt(model$statistic/(df*sum(x)))
  o <- model$observed
  e <- model$expected
  s <- sign(o-e)
  phi2 <- phi*((1/prod(dim(x))+(o-e)^2/e)
            /(1+model$statistic) )
  return ( round(100*s*phi2, 2) )
}
```

III. Google Frequency Estimates

We need to note that the frequencies we obtain from Google are *estimated frequencies*. The estimation procedure is controlled by Google and may change. However, in our experience the estimated frequencies are often very humanlike in that estimates may be higher in a more specific context. [6] gives the example of an unusual compounding (*Slotts gate*) that has a selectively higher frequency if it is combined by a high frequency word (*Oslo*) that it is associated with. Johansson discuss this as a "machine version of the ... Conjunction Fallacy". It can be argued that this could be a feature rather than a bug, as it makes (relative) frequency estimates that are closer to how humans judge such frequencies in a given context. Other reasons for using Google estimates are: coverage, and up-to-dateness. "As a comparison, Kuperman and Bertram (2013)[11] finds only 27 examples each of *apple sauce* and *applesauce* in a controlled corpus, whereas Google finds 20 million estimated documents for *applesauce* versus 28 million for *apple sauce*..."[6].

Uncertainties of using Google frequencies are connected to the fact that we do not know the algorithm that is used, and that algorithm can change at any time. Since it can be argued that lexical frequencies in context are extremely useful information for linguistic research it would be highly desirable to have a machine that can deliver frequencies over the net with the same or better coverage, and can provide possibilities for specifying estimation and smoothing functions. (Google do provide word frequencies (rather than document frequencies) as a separate resource [12], but this does not have the same coverage for longer phrases and we do not get the boost from term specificity, and crucially it does not allow us to search for context words within the same document.)

Any search engine will utilize the frequency of a term in a document in relation to the expected frequency of that term in general. For example, function words such as "the" will be present in almost all English documents. This has been elaborated into a measure of *term specificity* [13], or *inverse document frequency*, that gives a measure relative to the proportion of all documents containing the specific term. This can be further elaborated into measures that are proportional to the probability of a term occurring in a document. We do not know exactly which algorithm is used by Google, but the inverse of that measure is likely used in Google's reported frequencies. One consequence of this is that document frequencies of highly correlated and somewhat unusual (selective) terms may be overestimated and such terms that are anti-correlated may similarly be under-reported. One further factor to consider is that the number of documents that Google may retrieve is not fixed, but rather fast growing. As the standard user is mostly interested in the highest ranking documents for a search query there is typically no need for an exhaustive search, and the depth of the search may be limited by bandwidth capacity, which in turn may vary across the day. This means that Google frequency estimates may vary slightly depending on other factors than the search query itself.

A. Conjecture Fallacy

It has been argued [6] that Google frequency estimates may show a "machine version of the [...] Conjunction Fallacy", i.e. that the inclusion of an extra search term may result in higher, not lower, estimated frequencies. An analogy with the famous "feminist bank teller" [14] is that documents about feminists are rarely also about bank tellers, and vice versa, so they are highly anti-correlated. If a woman is described in terms associated with feminist activists and if we for a moment think of that description as a search query then such a query would heavily select documents where the term *bank teller* is relatively rare and *feminist* relatively frequent, and more so than in the general corpus. One of the context words used in the original experiment [14] was *outspoken*. Consider this as the context word, and regard Table III. The serendipity measure is given in brackets. As can be seen, "bank teller" in the context of *outspoken* is considered under-represented (-2.78) and should thus be avoided, and *feminist* + *"bank teller"* is over-represented (0.6) in that context and could be chosen as it is more frequent than expected in that context. Thus the serendipity measure used with Google frequency (sampled June 9, 2018) closely matches human intuitions [14]. Of course, before the publication [14] the number of *outspoken bank tellers* may have been much lower. However, the example shows that our human reliance on context rather than logic may be accurately mirrored in our procedure, though partly depending on how frequencies are estimated, and affected by term specificity.

TABLE III

GOOGLE CONJUNCTION FALLACY (FREQUENCIES IN THOUSANDS).

pattern	alone	+outspoken
feminist	94000 (-0.01)	1800 (0.22)
"bank teller"	8500 (0.05)	23 (-2.78)
feminist "bank teller"	131 (-0.01)	10 (0.60)

IV. EXAMPLES

Our sample data contain expressions that can be both compositional and non-compositional [8]. For instance, the German expression *ins Wasser fallen* has a non-compositional interpretation on which it means 'to fail to happen' (1) and a compositional interpretation on which it means 'to fall into water' (2):

(1) *Das Kind ist ins Wasser gefallen.*
'The child has fallen into the water.'

(2) *Die Eröffnung/(Einweihung) ist ins Wasser gefallen.*
'The opening/inauguration is cancelled.'

Our assumption is that the compositional or non-compositional meaning of an expression can be modelled in terms of the words with which it co-occurs (i.e. "das Kind", "die Eröffnung", or "die Einweihung" and the context words it may be associated with. The context words are those words that can function as an "implication" (i.e. effects or consequences) of the multiword expression. For instance, in its compositional meaning the MWE *ins Wasser gefallen* may entail actions like *schwimmen* "swim", *baden* "bathe" and even *ertrinken* "drown". To determine whether an expression has a compositional or non-compositional meaning, we compute the effect size for different configurations where an argument (referred to as *X*) co-occurs with the MWE with and without context. For instance, for the expression *ins Wasser gefallen*, the value of *X* can be *das Kind* "the child" or *die Eröffnung* "the opening", and the CONTEXT=(schwimmen|baden|ertrinken). When using Google Search, we basically get the two following configurations:

1) *X + MWE*: the phrase *X* and the MWE occur independently in the same document.
2) *X + MWE + CONTEXT*: Like the previous case, except we select documents containing specific context words.

We expect compositional MWEs to appear in contexts more similar to those in which their "implications" appear than do non-compositional MWEs.

A. Example 1: ins Wasser gefallen

Let X be a variable that can take one of the two values: x_1="*Das Kind*" or x_2="*Die Eröffnung*". Also, let us consider MWE be a variable referring to the multiword expression *ins Wasser gefallen*, and CONTEXT be a set that contains one or more of the context words *schwimmen, baden* and *ertrinken*. The Google frequency (June 6, 2018) of x_1 and MWE occurring independently (i.e. $x_1 + MWE$) is 10200 documents against 7320 for x_2 and MWE. If we add the context words, there are 22800 documents for x_1 + MWE + *CONTEXT* against 3580 documents for x_2 + MWE + CONTEXT. Table IV shows the frequency and effect size (in parenthesis) for the phrases "das Kind" and "die Eröffnung" in combination with the multiword expression "ins Wasser gefallen" with and without context.

TABLE IV

FREQUENCY AND (EFFECT SIZE) FOR *Das Kind/Die Eröffnung*

X	X + MWE	X + MWE + CONTEXT
x_1=*Das Kind*	10200 (-4.77)	*22800* (**3.17**)
x_2=*Die Eröffnung*	7320 (**14.44**)	3580 (-9.59)
x_1=*Das Kind*	10200 (-5.93)	22800 (**4.35**)
x_2=*Die Eröffnung*	7320 (**9.37**)	3580 (-6.86)
x_3=*Die Einweihung*	2140 (**5.82**)	455 (-4.27)

The effect size values provided in Table IV suggest that for MWE only we should select the non-compositional meaning (*die Eröffnung*) if there is no other information (positive effect size = 14.44). However, the compositional meaning should be chosen given the context words *schwimmen*, *baden* or *ertrinken* (positive effect size = 3.17), because it is much more frequent than expected. Recall that the context words are chosen to select documents with the combinatorial meaning. If we analyze simultaneously "die Einweihung" as well, we see that "das Kind" (4.35) is still the over-represented alternative in the context of water activities, and both the non-compositional examples behave similarly for proportions of document frequencies with and without context. Note also the above mentioned machine version of the Conjunction Fallacy: there are more documents (*22800*) found for "das Kind" with the context words than without them! This might be because the term is frequent by itself, and that may lead to a more shallow machine search as enough high ranking documents are found quickly without the context words, which in turn forces a deeper search when context is added. The compositional meaning has a stronger association with the context words than the non-compositional meaning, which is strongly negatively associated with those words. Most of the documents that contain the non-compositional meaning of the phrase *ins Wasser gefallen* do not mention words like *schwimmen*, *baden* or *ertrinken*.

B. *Example 2:* auf dem Tisch liegen

One complication is that MWEs can occur in many grammatical varieties. The following example shows the active form of the MWE *auf dem Tisch liegen* "to lie on the table", meaning either to be on the table physically or out in the open, and thus also open for negotiation. Sometimes the two meanings may coincide, for example when there is a written proposal that is physically on the table, see example 3 and 4. As Katz & Giesbrecht noted "[...] in the newspaper genre, highly idiomatic expressions [...] were often used in their idiomatic sense [...] particularly frequently in contexts in which elements of the literal meaning were also present"[8, p.17]. The selection of context words to separate a *dining situation* from a *negotiation* demands specific context words, and there could be entire procedures for finding context words that separate two situations.

(3) *Der Teller liegt auf dem Tisch.*
'The plate is/lies on the table.'

(4) *Der Vorschlag liegt auf dem Tisch.*
'The proposal is in the open / on the table.'

The results in Table V (frequencies estimated June 14, 2018) show that it is possible to separate these two situations from each other using the right context words. Context A selects documents that contain *Gabel*/fork, OR *Löffel*/spoon, OR *Messer*/knife. Context B selects documents that contain the word *verhandelt*/negotiated. In context A the literal, combinatorial, meaning (*der Teller*) is over-represented (2.64). In context B, the non-combinatorial, idiomatic, meaning (*der Vorschlag*) is over-represented (1.17). Combining context A and B shows that context A over-represents the literal meaning (5.71) and context B the idiomatic meaning (1.58).

The context words may come from the document that is being analyzed – we would look for terms that are more specific to our document than to the average document, and select those that co-occur with the MWE under scrutiny. There is obviously a need for meta-information about the MWE-expression to know which is the combinatorial and non-combinatorial. We need to know one of them to be able to compare them, and determine if they are the same or different. This could be accomplished by having stored prototypical subjects when the MWE is compositional or non-compositional. In this example, we used "Vorschlag", which is one of the non-compositional examples tested previously for the phenomena [8].

V. Discussion

As we have seen, the use of search engine frequency estimates are not only useful for coverage, but also when combined with the serendipity measure they closely mirror some human intuitions on the most "likely" (i.e. over-represented) alternative. Processing multiword frequencies is normally hard due to low coverage even in very large corpora. However, moving to frequency estimates rather than actually retrieving examples in context may be one step on the way. Such estimates may be based on how terms co-occur within the frame of a document, as documents have themes and aboutness. This may help to select different aspects of words and expressions that can be teased apart by adding context words that will select for one theme or meaning over other alternatives. For example, the word *outspoken* selected documents in line with how people often perceive the most likely outcome, and we saw the machine version of the *Conjunction Fallacy*. We have also shown how some German multiword expressions could be identified by noting that the compositional meaning is more likely when added context is congruent and collocates with that meaning, such as adding verbs like "swimming" or "drowning" as context to "falling into the water". The context words can be selected to collocate with the more literal meaning of the phrase, which can be checked by some excellent available resources that find relations between content words. One such resource is the *word2vec* function [15] as implemented in the *Gensim* project [16]. The main obstacles for the scientific use of search engine frequencies are related to the nature of commercial search engines. The goal of Google is not to provide linguistically motivated frequency estimates that can be replicated independently and understood using open mathematical principles of how the estimation is performed. However, they provide a useful *blackbox* for frequency estimation that can be used for proof of concept testing, as is done in the present article. Furthermore, since it is only the frequency estimates we are interested in, the links to found documents are not necessary.

TABLE V
FREQUENCY AND (EFFECT SIZE) FOR *Der Teller / Der Vorschlag*

X	X + MWE	X + MWE + CONTEXT **A**
x_1=*Der Teller*	5110 (-1.30)	2990 (**2.64**)
x_2=*Der Vorschlag*	13500 (0.54)	6110 (-1.10)
X	X + MWE	X + MWE + CONTEXT **B**
x_1=*Der Teller*	5110 (2.20)	2410 (-3.51)
x_2=*Der Vorschlag*	13500 (-0.74)	9240 (**1.17**)
X	X + MWE + CONTEXT **A**	X + MWE + CONTEXT **B**
x_1=*Der Teller*	2990 (**5.71**)	2410 (-4.46)
x_2=*Der Vorschlag*	6110 (-2.01)	9240 (**1.58**)

VI. Conclusion

Human intuitions are not necessarily consonant with formal logic, and may often be prone to logical fallacies such as the *Conjunction Fallacy*. We have argued that such fallacies are part of how we think and reason, and in communication such fallacies may aid communication when we assume that words and sentences in a context were indeed said with an intention. Previously [6] the serendipity measure has been used to investigate the process of compounding (and decompounding) when words are written together (e.g. "toothbrush" and "jaywalker") or apart as in "apple sauce". When we evaluate evidence in communication we may "think more like gamblers, in that we value information that changes probabilities more than we value absolute probabilities" [6]. This may even help communication if we all have this tendency to use context for inferring meaning in a communicative situation, provided that the more helpful context can be detected, for example by deviations from general expectations. The modeling presented in this short article is certainly helped by the effect from term specificity as modeled in search engines. We have used frequency estimates to find estimates that consider correlations and anti-correlations between terms, and thus provide an increased contrast between background frequencies and contextual frequencies. We believe that this feature of search engines is a useful feature. We have used the serendipity measure to mark which patterns are over- and under-represented, and we use that measure to take decisions. We suggest that automatic finding of context words can be achieved using available resources, for example the word similarity functions in the Gensim project[16]. Handling ambiguous expressions in general is a difficult problem for translation, as discourses within a language often play with triggering more than one meaning of an expression. However, in translation we are forced to select one meaning of an expression. Serendipity may help to indicate several relations between meanings and the available context words. Handling of multi-word expressions here also assumes that individual words can be found with ease. This situation is more complicated in for example Asian scripts, where space and punctuation cannot be assumed. However, calculation of the statistics is language independent once the relevant units are found, and we suspect that people react similarly to over- and under-representation independently of their languages.

Acknowledgements

We thank three anonymous reviewers for helpful comments.

References

[1] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for NLP," in *International Conference on Intelligent Text Processing and Computational Linguistics*, A. Gelbuk, Ed. Springer-Verlag, 2002, pp. 1–15.

[2] C. Ramisch and A. Villavicencio, "Computational treatment of multiword expressions," in *The Oxford Handbook of Computational Linguistics 2nd edition*, 2016.

[3] T. Baldwin and S. N. Kim, "Multiword Expressions," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. Boca Raton, USA: CRC Press, 2010, pp. 267–292.

[4] G. I. Lyse and G. Andersen, "Collocations and statistical analysis of n-grams," in *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, ser. Studies in Corpus Linguistics. Amsterdam: John Benjamins Publishing, 2012, pp. 79–109.

[5] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed. Chapman & Hall/CRC, 2010.

[6] C. Johansson, "A word or two?" *Bergen Language and Linguistics Studies*, vol. 8, no. 1, 2017. [Online]. Available: https://bells.uib.no/index.php/bells/article/view/1329

[7] D. Lin, "Automatic identification of non-compositional phrases," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 317–324.

[8] G. Katz and E. Giesbrecht, "Automatic identification of non-compositional multi-word expressions using latent semantic analysis," in *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, 2006, pp. 12–19.

[9] M. Yazdani, M. Farahmand, and J. Henderson, "Learning semantic composition to detect non-compositionality of multiword expressions," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1733–1742.

[10] G. M. Sullivan and R. Feinn, "Using effect size—or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.

[11] V. Kuperman and R. Bertram, "Moving spaces: Spelling alternation in English noun-noun compounds," *Language and Cognitive processes*, vol. 28, no. 7, pp. 939–966, 2013.

[12] M. Brysbaert, E. Keuleers, and B. New, "Assessing the Usefulness of Google Books' Word Frequencies for Psycholinguistic Research on Word Processing," *Frontiers in Psychology*, vol. 2, no. 27, 2011. [Online]. Available: http://doi.org/10.3389/fpsyg.2011.00027

[13] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[14] A. Tversky and D. Kahneman, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement," *Psychological Review*, vol. 90, no. 4, pp. 293–315, October 1983.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[16] R. Řehůřek and P. Sojka, "Software Framework for Topic Modeling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en