A word or two?

Christer Johansson

Abstract. The tendency for people to write compounds as two separate words, i.e. decompounding, is, for Scandinavian, often attributed to influence from English. However, English writers also both accidentally compound and decompound words. This article introduces *serendipity* as a statistical signal of surprise, i.e. deviance from expectations. Examples show that this measure can decide many cases of accidental compounding or decompounding by estimating which alternative is over-represented. Interestingly, the least frequent alternative can be clearly over-represented, thus providing a signal that is different from probability estimates, and linked to change in probability.

1 Introduction

People sometimes accidentally miss a key when writing on a *typewriter* (which is one example of an English compound), resulting in 'words' such as *isthere*, and sometimes people hit a space, where it should not be, resulting in words such as *tooth brush* instead of *toothbrush.* Missing a space between two words that are not part of a compound, i.e. *accidental compounding*, can be viewed as a more or less random process that happens at some rate of error. Insertion of an extraneous space is often related to a morpheme boundary, i.e. *accidental decompounding*. This latter process could be a step in approximating where to put a space. Both accidental compounding and decompounding can be a problem for writers, who risk mockery from ungenerous readers.

Accidental compounding: A compound phrase consisting of two separate words are sometimes wrongly written as one word. This could be an accident, such as when two words that often occur together are written as one word: for example *is there* written as *isthere.* Some compound phrases such as *power station* that should be written as two words, are fairly often wrongly written as one word. One explanation is that they are accidental compounds, just like *isthere.*

Accidental decompounding: One explanation for decompounding words such as *lighthouse* is that it is simply a matter of an accidental insertion of a space at a morpheme boundary. There is a background rate of how often we insert a space by mistake at

The very model of a modern linguist edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1329.

such a boundary. However, it is clear that this rate is affected by how often we see the word written correctly and incorrectly. *Football* is almost never decompounded, and *muscle car* is rarely written as one word.

There is less compounding in English than in German or Scandinavian, with many exceptions such as *firefighter*, *football* and *toothbrush*. Sometimes, an English compound begins as a two-word phrase (e.g. *jay walking*, which is now *jaywalking* and the meaning of *jay* in the compound is more or less lost). Such phrases tend to drift towards a one-word compound with increased usage, and correspondingly more specific meaning.

Decompounding words may lead to misunderstandings, as it can affect lexical choice. For example, in Swedish *kassa apparater* are *faulty machines*, but *kassaapparater* are *cashier's registers*. In Swedish it is not common for two vowels to clash, thus an inserted space might reflect the lower transition probability between two vowels inside a word, compared to between words. Language-specific letter transition probabilities have been shown to affect reaction times and accuracy for decision tasks on Norwegian Bokmål and English (Van Kesteren et al. 2012).

The proportion between writing in one or two words, could be more or less surprising. The relative frequency of a one-word compound is typically higher than accounted for by the proportion of words that are accidentally compounded, i.e. the components are not statistically independent.

The new measure of serendipity, as it is introduced in this article, is interesting as an alternative to thinking in the absolute probabilities that most people are not good at estimating. For example, we have a tendency to overestimate the probability of two events occurring together (the Conjunction Fallacy, see section on prerequisites) and also attribute causation to rare events that happen in close sequence (*post hoc ergo propter hoc*, or Causation Fallacy) or repeatedly occur together (correlation implies Causation Fallacy).

What we need is a measure that is insensitive to the number of examples, for example by putting more emphasis on *effect size* rather than *significance* (Johansson 2013). I will also argue that we think more like gamblers, in that we value information that *changes* probabilities more than we value absolute probabilities. If we get information that makes a horse ten times more likely to win, wouldn't we put some money on it even if it still is an unlikely winner? Serendipity is a measure of surprise, and surprise is a good trigger for learning.

In a famous review, Chomsky (1959) draws a caricature of Skinner's research on verbal behavior, by more or less equating the approach with reinforcement learning in animal studies (MacCorquodale 1970). Chomsky's review gave the impression that statistics was not very useful for investigating language structure, and put focus on how improbable a simplistic probabilistic calculation of recursive structures would be.

The question of which signals we use to find linguistic information is not resolved, except by the assumption of innate structures; i.e. we do not find it because it is already there. Optimizing the probability of a sentence is obviously hard given a limited sample, and the infinite possibilities of language to form new sentences and new words. This article will show how the change in probability when comparing alternatives can be used for a seemingly simple task of deciding whether (any) two consecutive words should be written as one or two words. This expands on work by Rømcke and Johansson (2008), where frequencies from a search engine were used to decide categories for named entities, by comparing frequency responses to the words in contexts such as *Hotel in Bergen* or *Her name is Bergen*. Search engine frequencies have also been used to investigate the dative alternation, using frequency responses to the two versions of the dative construction for a set of different dative verbs (Jenset and Johansson 2013). The aim of the examples is to illustrate the signal surprise and expectation, as measured by a new measure.

This article will introduce *serendipity* as the pointwise effect size, by showing how to distribute effect size over the contributions to significance of the individual cells. Crucially, effect size and serendipity are insensitive to how much data we use. This article will begin with some prerequisites, related to statistical independence, cross table testing, significance, effect size and what could be called *serendipity* (i.e., the effect size of a single cell in a cross table), and using Google as a quick and dirty source of observed frequencies.

2 Prerequisites

2.1 Cross tables

A cross table analysis is an analysis of frequencies that tests whether rows and columns are statistically independent of each other. The most basic case is the 2 rows by 2 columns, and it is certainly the easiest cross table to interpret. The null hypothesis is that the rows and columns are independent of each other. Let us consider a simple cross table and calculate the expected independent frequencies of each cell. Let $a + c = R_1$ and $b + d = R_2$ be the total frequencies for row 1 and 2, and $a + b = C_1$ and $c + d = C_2$ the total frequencies of column 1 and 2, and $a + b + c + d = T$ is the total (cf. Table 1).

a	c	R_1
b	d	R_2
C_1	C_2	T

Table 1: A cross table

If the rows and columns are independent then the probability of belonging to row 1 is R_1/T, and R_2/T is the probability of belonging to row 2. The probability of belonging to column 1 is C_1/T and C_2/T is the probability of belonging to column 2.

Assuming independence, the probability of belonging to a cell that is the combination of a row and a column, is simply the multiplication of the row and column probabilities. We can calculate the expected frequencies in each cell (cf. Table 2) by distributing the total by the proportion in each cell.

In order to test for significant deviation from independence, we should look at the difference between observed and expected frequencies (e.g., $O_{11} - E_{11}$); if this difference is positive, that cell is *over-represented* and if it is negative, it is *under-represented.* The test sums up the square of these differences, each one compared with its expected frequency: $\chi^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$; this positive sum indicates how rare it would be to find so much deviance if the rows and columns are indeed independent. In order to look it up in the χ^2-distribution, you need to know how many ways this could happen (i.e. the degrees of freedom). The 2-by-2-table has one degree of freedom, since if you know the value in one cell, you can easily calculate the rest from the row and column sums. This is called degrees of freedom (*df*), because the rest of the cells are uniquely determined by the row and column sums if we know the value for $R-1$ row cells and $C-1$ column cells, simultaneously, i.e. $df = (R-1) * (C-1)$, where R and C are the number of rows and columns, respectively. It should not be a big surprise if there are significant deviations from independence for language data; after all the process that generated the frequencies (for example, writing an essay) is not a random process. Significance only tells us if it is likely or not that the rows and columns are independent. It does not tell how large or how relevant the effect is.

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

Table 2: The expected frequencies

2.1.1 Pearson residuals

If we want to say which cells contribute more to *significance*, then we should look at the *signed* Pearson residuals, which measure the signed contribution to significance in each cell: $\frac{(O_{ij}-E_{ij})}{\sqrt{E_{ij}}}$ from the term $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ in the χ^2-formula, before squaring. One reason for squaring is to sum up deviance in absolute values, and use this to decide the fit of the observations to a model of independence.

2.1.2 Association plots

An association plot is a tool to graphically explore and visualize the effects of each cell in a table. The R command *assocplot* can be used, but the function *assoc* from the

Visualizing Categorical Data (vcd) package (Meyer et al. 2016) provides more possibilities, for example visualizing the Pearson residuals range with a color gradient. The association plot provides bars, whose base is proportional to $\sqrt{E_{ij}}$, and the height is proportional to $O_{ij} - E_{ij}$ and thus the width of the base tells about expectations in that cell, and the height of the bar tells about the deviance from expectations.

2.2 Effect size

Effect size for a χ^2 test is calculated as $\Phi = \sqrt{\frac{\chi^2}{N}}$. Cramér's Φ can be generalized to larger tables, using Cramér's $\nu = \sqrt{\frac{\chi^2}{df*N}}$, where df is the smallest number of the number of either $rows - 1$ or $columns - 1$. The effect size is nearly independent of how many observations the table represents, whereas almost any real difference between observed and expected can be detected with significance by sampling large enough samples from the population. Therefore, if we want to compare results, we should include the effect size. Significance is just a receipt that we have observed a deviance from independence that cannot be explained by random chance.

2.3 Serendipity or the effect size per cell

We are interested in each cell's contribution to the effect size. One way is to note each cell's contribution to the χ^2 statistic compared to the overall χ^2, and this tells each cell's proportional contribution to the effect size measure. The effect size in each cell i, distributed by each cell's contribution to significance (i.e. to χ^2) is given in formula (1).

$$\Phi \sum_{i=1}^{n} (\frac{(O_i - E_i)^2}{E_i})/\chi^2 => \frac{\Phi}{\chi^2}(\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} ... \frac{(O_n - E_n)^2}{E_n}) \tag{1}$$

where Φ and χ^2 are numbers and $\frac{(O_i - E_i)^2}{E_i}$ are terms in a series, which can be ordered in a table.

Proof of correctness for $\chi^2 > 0$: Note $1 = \frac{\chi^2}{\chi^2}$ and rewrite according to definition formula (2).

$$\sum_{i=1}^{n} (\frac{(O_i - E_i)^2}{E_i}/\chi^2) = \frac{1}{\chi^2} \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

which are the terms in the series we need to distribute Φ over all cells.

There are still some problems with the desired function. First of all, division is sensitive when χ^2 is close to zero. This can be fixed by adding one to the numerator and the denominator, and we can let the one in the numerator divide up equally on all the cells. Finally we may scale the value by multiplying by 100 and rounding to two decimals. The scaling is just cosmetic, and makes it easier to read the output. If you sum up the absolute values of all cells you will get back the overall Φ, but remember we have scaled the value by multiplying with 100. The effect size is actually generalized to any size of table by using Cramér's ν to correct for increased degrees of freedom (df).

In the programming language R we define the function in (3) that returns a table with the signed effect size for each cell:

```
(3)   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) )
        }
```

2.4 Conjunction Fallacy

The probability of two events occurring together (in *conjunction*) is always less than or equal to the probability of either one occurring alone (cf. Wikipedia on Conjunction Fallacy). Tversky and Kahneman (1983) investigated some conditions under which we are more likely to estimate the conjunction as more probable than just one of the events. The classic example presents Linda as having several characteristics of a feminist activist, but nothing to suggest that she is a bank teller. In that situation, most subjects would state that Linda is more likely a feminist and a bank teller, than a bank teller. One possible flaw, is that the alternatives are read in contrast to each other and therefore the alternative that she is a bank teller is actively read as: *she is a bank teller but not a feminist.* This was controlled for in several follow up experiments (ibid.). One version had two explicit arguments to choose from, either argument *a)* "Linda is more likely to be a bank teller than she is to be a feminist bank teller, because every feminist bank teller is a bank teller, but some women bank tellers are not feminists, and Linda could be one of them" (ibid., p. 299) or argument *b)* "Linda is more likely to be a feminist bank teller than she is likely to be a bank teller, because she resembles an active feminist more than she resembles a bank teller" (ibid.). A majority of subjects (65%, 58) preferred alternative *b*, which is still better than the 85% that preferred the conjunction, if there were no explicit arguments for the alternatives. Gould (1988) presents a popular text on this and other statistical fallacies.

2.5 Google frequency estimates

Google provides frequency estimates of search phrases. In my experience, it works better for short phrases, and not too many context words, where examples could be retrieved from big tables rather than estimated. Remember that the frequencies are *estimated*, and may not be accurate. For our purposes we are more interested in the proportions than in the absolute frequencies. Google frequencies give an estimate of the number of *documents* that contain the search terms, which in itself points to conservative estimates. However, the collection of documents may contain many duplicates. Furthermore, the estimates may have uncertainties that vary non-linearly with the size of the frequency estimate, but there are very few other sources that covers so much of the lower frequency words, which is especially important when we look for low frequency compounds. Obviously, there are data sources of much better quality. There are, however, many reasons why we could prefer Google frequencies. First, when looking for words in context it is crucial to have as many examples as possible. As a comparison, Kuperman and Bertram (2013) finds only 27 examples each of *apple sauce* and *applesauce* in a controlled corpus, whereas Google finds 20 million estimated documents for *applesauce* versus 450 thousand for *apple sauce* (some documents may mention both variants). Second, Google gives document frequencies, which may actually lessen the bias of individual writers, who may overuse certain patterns. Third, Google does not (always) normalize words, so misspelled words or compounds can be represented. Fourth, search through the Google search engine makes replication widely available for almost anybody with Internet access. Finally, Google is updated more frequently than most corpora, which is important when we are interested in contemporary usage. However, it would obviously be good to have access to a linguistic search engine, since Google's search engine is not tailored for the needs of linguists and therefore may prioritize other issues such as bandwidth capacity. The algorithms that are used by Google may also change without notice. It is also an idea to build future applications, where part of the computing is done on the Internet as a distributed system.

In my experience, Google frequencies may often display a machine version of the above mentioned Conjunction Fallacy, i.e. a more specific search may very well indicate more documents rather than fewer. For the normal user of the search engine, this is not a problem as long as the highest ranking documents are the most relevant. For serious research this means that the frequencies should be seen as illustrations rather than hard facts. As will be clear from the examples, in practice the proportions are often very clear, which means that only large errors will affect the decisions based on the effect size measure introduced in this article.

2.5.1 Is it wiki or kiwi?

Table 3 works as an illustration: When searching for the words *kiwi* and *wiki* with and without context words, millions of documents were indicated. Note that *kiwi* is the least frequent alternative both with and without context words. However, when we put on the effect-size goggles, it is clear that we could choose *wiki* if there is no other information (positive effect size = 0.03) and we should choose *kiwi* given the context words *banana* and *fruit* (positive effect size = 15.22), because it is much more frequent than expected. More specifically, we could even program a computer to use the effect-size measure, as one of many measures, to take decisions between alternatives. It should be clear that the measure is not only probability of occurrence, but directed deviance from expectations, where expectations are set up by statistical independence of (structured) alternatives. Whether people similarly use effect size to guide intuitions about probabilities is a research question that has been hinted at previously, when discussing the Conjunction Fallacy. It should also be noted that part of the work is done by selecting alternatives to compare with, which is one way to establish a baseline for expectations.

	word	+fruit +banana
kiwi	71.2 (-0.21)	6.0 (15.22)
wiki	905.0 (0.03)	6.7 (-1.30)

Table 3: Frequency and (effect size) for *kiwi*/*wiki*.

In probabilistic terms, we would always have a larger *probability* of finding a document containing the word *wiki* than one containing *kiwi*, but *kiwi* has a much stronger association with *banana* and *fruit* than the word *wiki*, which is weakly negatively associated with those words (i.e., most documents that contain *wiki* do not mention *fruit* and *banana*). The **Pearson residual** of *kiwi* in context is 15.9, but if the table is divided by 10 then the cell's effect size is still 14.88, but the Pearson residual is just 5.03, illustrating that the effect size is roughly constant, but the contribution to *significance* varies. Effect size is more relevant than significance, *when we are looking for associations.*

In terms of **Bayesian probability**, given that we have a choice between *kiwi* and *wiki*, the *prior probability*, from the column cells and the column totals, of *kiwi* is $712/(9050 + 712) = 0.073$, and of *wiki* $9050/(9050 + 712) = 0.927$; the probability of *kiwi* given *fruit* and *banana* is $60/(60 + 67) = 0.472$; the probability of *wiki* (i.e., not *kiwi*) given *fruit* and *banana* is $67/(60+67) = 0.528$. The *adjusted probability* of *kiwi* is 0.066, which is lower than its prior probability, and the probability of *wiki* is correspondingly 0.934. This shows that the effect size as an association measure is not just Bayesian probability.

In relation to the Conjunction Fallacy, the association between *kiwi* and *fruit* and *banana* is clearly shown by the ratio between odds with and without information. With the contextual information, 6.0 out of 12.7 (i.e., $6.0 + 6.7$) is for *kiwi*, which gives an odds of 0.472, or expressed as a percentage: 47.2% gives *kiwi* in the specific comparison. Without the contextual information, 71.2 out of 976.2 (i.e. $71.2+905.0$) is for *kiwi*, which gives an odds of 0.072. The odds ratio for *kiwi* is $0.472/0.072 = 6.56$, and similarly for *wiki* $0.528/0.927 = 0.570$. Thus, knowing *fruit* and *banana* increases the chances that it is *kiwi* more than 6 times, while it almost halves the odds that it is *wiki*. So even if it is still unlikely that it is *kiwi*, it would be tempting to choose it — if it was a bet and the payout is set by the prior probability. This looks similar to how the Conjunction Fallacy works in that having the extra information gives an advantage, in the example above knowing about Linda increases the chances that she is both a bank teller and a feminist much more than it increases the chances that she is a bank teller. What looks like a fallacy might in fact be a more or less innate tendency to value *change in probability* much more than the absolute probability. People could also use this as a communicative strategy: mention only information that changes background knowledge.

2.6 Summary

Effect size distributed per cell is a convenient way of investigating associations in a table. It works well with frequencies, and it is intuitive to understand the concepts in terms of over- and under-represented compared to estimates based on statistical independence of rows and columns. Cells that deviate from expectations are marked clearly.

3 Examples

3.1 To compound or not to compound?

The Google frequency (February 6, 2017) of *there is* is 2390 million against 2.720 million for *thereis*. For *is there* there are 460 million documents and for *isthere* 0.500 million documents. The ratio is between $878 : 1$ and $920 : 1$, respectively. The rounded ratio of $1000 : 1$ will do fine as a baseline, which we can call *is there*. This ratio is likely similar in other languages as well, since it is motivated by the same process of missing a keystroke. However, with less material there is a higher risk that the ratio will be more off. Also, increasing use of better spelling correction may influence the frequencies.

Obviously, for most real applications the missing keystroke rate will have to be estimated for the individual for increased precision, and luckily this should not be very hard to do. It could even be a good idea to estimate more precise measures such as the rate of missing a space between any two specified characters.

3.1.1 Is it firefighter or fire fighter?

Table 4 shows the table for deciding between *firefighter* and *fire fighter*, which has a ratio of $65 : 1$ in favor of being written as one word. Incidentally, the ratio is almost the same for *fireman*, at $68 : 1$, it just looks like *firefighter* has a 45% higher frequency. Note that *is there* is preferred as two words, even though we have not explicitly said that it should never be written as one word, and therefore it could *adapt* (by lower effect size) to words like *firefighter*. The important part for the decision is merely which way the effects go.

However, can we be sure that it is not two words? Reasoning from statistical hypothesis testing, we would have to *disprove*, or at least make it unlikely, that the word has *not* been split by mistake. This is a bit trickier. One of the most frequent compound

	word	two words
firefighter	51.5 (88.61)	0.795 (-4.67)
is there	1 (-4.65)	1000 (0.27)

Table 4: Frequency in millions and (effect size) for *firefighter*.

words in English is probably *football*. We could compare that with *firefighter*. Table 5 shows that *firefighter* is indeed not as strong a compound as *football*. Compared to *football*, *fire fighter* cannot be completely ruled out as a two word compound that has undergone *accidental compounding*. For a decision, we then have to compare the effect sizes of *firefighter* as one word in Table 4 (88.6) and *fire fighter* as two words in Table 5 (8.8), and 88.6 clearly wins over 8.8. The first conclusion is that *firefighter* is not likely to be explained as *accidental compounding* and the second is that *firefighter* is a weaker compound than *football*. If we are still unsure we could compare it with a reference word such as *banana peel*, see Table 6. The word *firefighter* is one word, but *banana peel* is strongly a two-word compound, when compared to each other.

	word	two words
firefighter	51.5 (-0.18)	0.795 (8.77)
football	1330 (0.17)	0.408 (-0.51)

Table 5: Frequency in millions and (effect size) for *firefighter* vs. *football*.

	word	two words
firefighter	51500 (0.01)	795 (-0.43)
banana peel	4.2 (-1.29)	403 (55.51)

Table 6: Frequency in thousands and (effect size).

3.1.2 Is it Slotts gate or Slottsgate?

One famous example of decompounding in Norwegian is *Øvre Slottsgate* 'Upper Castle Street'. In Oslo, the street sign actually reads *Øvre Slotts gate*. Investigating the web finds that there is indeed a tentative association between decompounding this street name and documents that also contains the word Oslo, see Table 7.

This is also an example of the statistical Conjunction Fallacy for Google frequencies — adding a demand for an extra keyword ought to give fewer documents, but the search engine has detected a strong association between Oslo and this street name, and the estimate is higher, possibly because the search engine has performed a deeper search. This seems to affect the rarer variant more. For comparison, see Table 8 where the address is made more specific by adding *Øvre* to the street name. From the table we see very small effect sizes in all the cells, but a small preference for associating *Øvre Slotts gate* with Oslo. The decompounded version is associated with Oslo, and we also know that in Oslo there are street signs that show the decompounded version. Since effect sizes are so small for all versions of *Øvre Slotts gate* the decision could be to trust as it was written. In a practical application, the false alarm rate also need to be kept low, and setting an individual threshold for when to suggest an edit makes sense.

	all	oslo
slotts gate	1520 (-3.93)	6420 (4.17)
slottsgate	256000 (0.06)	236000 (-0.07)

Table 7: Frequency and (effect size) for *Slottsgate* ±Oslo.

	all	oslo
øvre slotts gate	3890 (-0.19)	3890 (0.20)
øvre slottsgate	150000 (0.02)	142000 (-0.02)

Table 8: Frequency and (effect size) for *Øvre Slottsgate* ±Oslo.

4 Analysis

The decision for one word over two words is affected by the baselines for the comparisons. From the examples, we have seen that there are two baselines: one for our expectations for accidental compounding and one for accidental decompounding. In order to show how this works for different proportions, two extremes were chosen as a graphical illustration. One baseline proportion is the accidental compound *thereis*, which occurs once for every thousand occurrences of the correct *there is*. The other extreme is a hypothetical word that should be written as two words but often ends up as one word (e.g., *musclecar*) and that proportion is set at 3 incorrect *onewords* to 1 correct *two word*, which is a pessimistic estimate of accidental decompounding. Note that *musclecar* has the same structure as *football*, i.e., a body part and an object. Most non-compounds have detectably more support as non-compounds, but Figure 1 illustrates that two words could be favored, even under conditions where the baseline itself favors one word $3:1$.

	one word	two words
candidate	a	b
baseline	c	d

Table 9: Baseline matrix

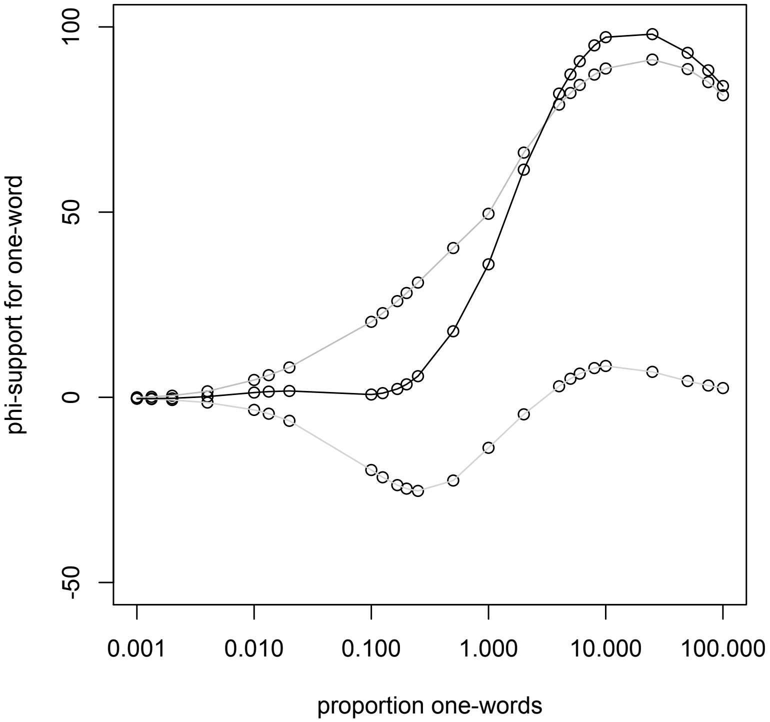

Figure 1: Support for one word. The Y-axis shows the serendipity score and the x-axis shows the proportion $a : b$ (cf. Table 9)

Table 9 shows a matrix for comparing a candidate with a baseline. The frequency for a one-word interpretation is given in the cell marked *a*, and *b* is the two-word frequency. The letters of the table also mark the position for the serendipity scores.

Figure 1 shows how much *phi-support* (i.e., the serendipity score, or pointwise effect size) a new candidate has for being *oneword.* Two different baselines are illustrated: The upper gray line shows a baseline at $c = 1 : d = 1000$ and the lower gray line shows a baseline at $c = 3 : d = 1$.

The score for supporting compounding is calculated for candidate proportions that range from $1 : 1000$ to $100 : 1$. Note that these proportions are plotted on a logarithmic scale on the x-axis.

The lower gray line shows the accumulated serendipity score *against* the one-word interpretations, which accounts for support for two words as well, i.e. the support for a oneword interpretation minus the support for a two word interpretation. The dark line shows the *net support* for one word after accounting for both baselines. The line shows that given these baselines, the positive net support for oneword starts before a $1 : 1$ proportion, and it quickly gains positive support. After $100 : 1$ there will be increasingly smaller relative differences between observed and expected frequencies, and less signal for learning, and one way to view this is that there are fewer alternatives and no need to learn as expectations are as observed.

5 Discussion

Are there more linguistically relevant problems, where a measure of association such as the serendipity measure can be used? One possible example is Anaphora Resolution, which is hard to resolve using empirical methods (Nøklestad and Johansson 2005). We found that candidate antecedents can be at long text distances, and most potential candidates are *not* coreferent with the pronoun to resolve. As Nøklestad (2009, p. 215) notes: "Thus, the tendency of the system to classify a candidate as non-antecedent is so strong that a single feature is rarely able to overcome it. This is hardly surprising, given the overwhelming majority of negative examples in the training data (...)".

An idea introduced in this article is that informativeness, and surprise, are related to how much probabilities *change* in a new context, and that this can be used as a trigger for learning. This idea could be applied to coreference resolution: Which antecedent candidate will change the background probability the most? Such an approach has the possibility to find associations that are not the most objectively probable. However, if we take into account that other people may react on, and use, change in probability, this has a good chance to be a relevant signal. Just as the solution to the famous Monty Hall problem (Rosenhouse 2009) lies in realizing that the objective situation that there are two boxes to choose from, has a context and a history that makes it highly rational for a participant to change to the other box, thus changing the initial risk of losing to a chance of winning.

In this article, examples have shown that the risk of both accidental compounding and accidental decompounding has to be taken in account. The serendipity measure that was introduced here reacts on an effect size that is crucially *insensitive* to, or near independent of, the size of the data sample. This means that the measure can be compared, even if we do not know the size of the population. When we compare one-word and two-word 'compounds' with each other, we find that, for English, there seems to be a gliding scale from preferring one word to preferring two words for compounds.

Note that the decision space has not been optimized. For a spelling application, information on *how* something was written should be taken into account. Was the word written fluently, without major hesitations as noticed by time between key presses (key latencies), or were there several attempts at writing the word? The attempts may have information about the intended word. Are there similar words in the text? It is common to find the intended word correctly spelled in the same text. Keeping track of how often the different kinds of errors occur for a writer could help us discover the optimal point at which more errors are fixed than created. Uncertainty in search engine frequencies is thought to be handled by comparing close examples, with the same number of words in the patterns. The main reason to use search engines is their coverage. Any controlled source with better coverage should be preferred.

In relation to a model of how people handle compounding (Kuperman and Bertram 2013) it is interesting to note that frequency of use, and familiarity, seems to play an important part. As noted previously there is a tendency for unfamiliar or new compounds to start out as spaced compounds (e.g. *jay walker*) and drift towards a fully compounded unit, such as *jaywalker*. Kuperman and Bertram (2013) provide further examples and notice "going against […] orthographic preferences in production comes with a high cost in recognition", which creates a pressure towards adapting to the expectations of readers. They (ibid.) also mention that the strategy for selecting the best alternative form of compounding evolves, as various processes such as morphemic segmentation, semantic integration and visual recognition are influenced by frequency of usage and familiarity. Additionally, there are effects that could be characterized as related to balance between the constituents of the compound; in length, and syllable structure. Such effects may counteract, or support, a transition to more compounding in usage.

Acknowledgement

The author would like to thank two anonymous reviewers for comments that increased the readability, and made some arguments clearer.

References

Chomsky, Noam (1959). "Review of Skinner's *Verbal Behavior*". In: *Language* 35, pp. 26–58.

Gould, Stephen Jay (1988). "The Streak of Streaks". In: *The New York Review of Books.*

Jenset, Gard Buen and Christer Johansson (2013). "Lexical fillers influence the dative alternation: Estimating constructional saliency using web document frequencies". In: *Journal of Quantitative Linguistics* 20.1, pp. 13–44.

Johansson, Christer (2013). "Hunting for significance". In: *The many facets of corpus linguistics in Bergen – In honour of Knut Hofland.* Ed. by Lidun Hareide, Michael Oakes, and Christer Johansson. Vol. 3. Bergen Language and Linguistics Studies (BeLLS) 1. University of Bergen, pp. 211–220.

Kuperman, Victor and Raymond Bertram (2013). “Moving spaces: Spelling alternation in English noun-noun compounds”. In: *Language and Cognitive processes* 28.7, pp. 939–966.

MacCorquodale, Kenneth (1970). “On Chomsky’s review of Skinner’s *Verbal Behavior*”. In: *Journal of the Experimental Analysis of Behavior* 13.1, pp. 83–99.

Meyer, David, Achim Zeileis, Kurt Hornik, Florian Gerber, and Michael Friendly (2016). *Package ‘vcd’ (Visualizing Categorial Data).* Tech. rep. CRAN.

Nøklestad, Anders (2009). “A machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection”. PhD thesis. University of Oslo.

Nøklestad, Anders and Christer Johansson (2005). “Detecting Reference Chains in Norwegian”. In: *Proceedings of the 15th Nodalida Conference.* Juensuu, Finland: University of Joensuu electronic publications in linguistics and language technology.

Rømcke, Audun and Christer Johansson (2008). “Named Entity Recognition using the Web”. In: *Proceedings of the Second Workshop on Anaphora Resolution.* Ed. by Christer Johansson. Northern European Association for Language Technology (NEALT). Bergen, Norway.

Rosenhouse, Jason (2009). *The Monty Hall problem.* Oxford: Oxford University Press.

Tversky, Amos and Daniel Kahneman (1983). “Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement”. In: *Psychological Review* 90.4, pp. 293–315.

Van Kesteren, Ron, Ton Dijkstra, and Koenraad De Smedt (2012). “Markedness effects in Norwegian–English bilinguals: Task-dependent use of language-specific letters and bigrams”. In: *The Quarterly Journal of Experimental Psychology* 65.11, pp. 2129–2154.