Hunting for Significance

Christer Johansson[1]

[1] University of Bergen;

Abstract

The concepts of statistical significance and effect size are discussed. Three small case studies using data from Google search, and the British National Corpus are presented. The cases illustrate how some linguistic questions can be investigated quantitatively, and which measures could be interesting for the linguist to consider. Implications and a linguistic interpretation of effect size are discussed. The effect size gives a clue to how hard it is to detect a significant association, and this is information that significance alone does not provide.

Keywords: significance; effect size, ecological validity, quantitative linguistics

*** Principal contact:**
Christer Johansson, Professor
Computational Linguistics, LLE, University of Bergen, HF-bygget, 5020 Bergen, Norway
Tel.: +47 55 55 822 62
E-mail: christer.johansson@uib.no

1. Introduction

Statistical significance might qualify as one of the most misunderstood scientific concepts among non-scientists. This is partly due to the different use of the word significance in daily speech. What is the significance of dreams? In everyday use we would think of what dreams could mean in general, and to our lives in particular. Statistical significance, however, only relates to the question if an observation (of a difference) is real or if it can be explained by random chance or coincidence. Is it an observation or is it a fluke? In particular, statistical significance does not necessarily mean that our observation is important, relevant or even useful. It is not even so that the more statistically significant an observation is the more important it is. There are many cases where an observation is highly significant, but it still does not help much to explain the phenomena that we are interested in. Statistical significance simply relates to how sure we are that the observation (most often an observation of a difference, either between groups or a difference from an expectation) is truly an observation, and we are not just looking at random variation. In many fields we are satisfied with a p-value less than 0.05 giving a one chance in twenty that the observation is explained by random chance. In other fields, such as astronomy, they might not be satisfied until p is less than one in a very large number; for example when detecting the presence of a planet around a star based on regular fluctuations of light intensity observed from the star. Some of this is also a question of economy. When we make for example reaction time experiments on people, gathering measurements may take about half an hour for each subject. Lately we have built up resources in the form of very large corpora that in one sense can be compared to astronomical measurements, i.e. we are able to get very a large number of occurrences (individual observations) for most linguistically relevant patterns, and thus also very impressive numbers for significance, often with p approaching 0, suggesting that there is no possibility of mistakes, the differences are almost certainly there. The next logical question is more uncomfortable: does it really mean something? What is our observation explaining? Does it help us predict anything?

case 1: Spelling

In July 2011 Ramesh Krishnamurthy, in a follow-up on a post by Yorick Wilks, posted an interesting question, along with some data, on the Corpora list in reply to a question of what is the correct measure for measuring language change, and errors in spelling in particular. The data is reproduced in table 1.

Table 1
Reversing i and e in two words

	Spelling variant	
Word	**correct**	**reversed**
ceiling	223000000	738000
piece	1290000000	10800000

Although the exact numbers have certainly changed since then, I am convinced that the proportions are fairly much the same if you search Google for the frequencies of the correct and reversed varieties. Anyway the frequencies Google provides are estimates of document frequencies, and may vary slightly for reasons other than the fact that the number of documents double at a tremendous rate (some guess at a doubling every five years, or about 15 percent per year. The number of examples of *ceiling* has increased more than thirty percent in a year and a half, but *cieling* is up by only 15 percent. Are spelling checkers getting better, or applied more by writers or by Google?). For *ceiling* about three in a thousand have reversed the *ei* to an *ie*, and for

piece about eight in a thousand are reversed from *ie* to *ei*, almost three times the reversals for the higher frequency word. Is this observation significant? Yes, and impressively so: testing the table using a chi-square test gives an amazing chi-square statistic of 636454.7, and a p value that is approaching 0.

Does this mean that knowing the intended word gives us a better chance at predicting if it is misspelled or not? Over 99 percent of both words are correctly spelled (at least, if there are only these two options). The question is related to the question of effect size (in the case of cross-tables analysis also called strength of association). Does it give us an edge if we know which row (or column) we are looking at? The measure for the effect size is Cramer's Phi (√). If the effect size is large enough it might have some predictive value, i.e. it can be useful for a purpose. How large an effect size needs to be to be useful depends on the task, so we must proceed with caution. Note that only significant observations will have a *relevant* effect size.

Formula 1

Effect size for association between word and spelling

$$\Phi = \sqrt{\frac{\chi^2}{N}}\ ; with\ numbers\ inserted: \Phi = \sqrt{\frac{636454.7}{524538000}}$$

Note N is the sum of all entries in the table.

Phi is a small number, 0.02, and this is called a tiny effect size. It indicates that not much of the data in the table is explained by knowing which word we are looking at. It would take many observations of correct and 'incorrect' use before we could claim that the associations in the table are real observations. We would need to make many observations to report with significance ($p<0.05$) that reversals of the stem vowel are different for *ceiling* and *piece*. Significance tells us that there *are* differences between *ceiling* and *piece* for the proportion of reversed stem vowels in misspellings. The small effect size tells us that the actual difference is hard to observe.

Another difference between *ceiling* and *piece* is that misspellings of *ceiling* will most likely result in non-words, which are easy to find, whereas misspellings of *piece* might include close neighbors such as *peas*, and *peace*, that are real words and much harder to detect as misspellings without an analysis of the meaning of the word in context.

Another tricky anomaly is that *piece*, if read letter by letter, could lead us to misread the first syllable as *pie*. Maybe such false starts could explain part of why the proportion of reversed vowels is *significantly* higher for *piece*? The effect size indicates that most of the time there is no problem, and it would be unlikely that, based on these words, spelling would drift towards either *ie* or *ei*; in both cases a vast majority of examples are spelled correctly.

case 2: The Ditransitive Construction

A lately much discussed phenomena is the variation in some English verbs to either take the ditransitive construction or use a to-construction. Several articles (Bresnan et al., 2007; Stefanowitsch, 2006; Stefanowitsch and Gries,2008; Jenset and Johansson, 2013) investigate this from a statistical point of departure. Stefanowitsch and Gries (2008)use Fisher's Exact test to calculate significance. The reason for using Fisher's test is that it also handles cases where the numbers are low, and even below five for an individual cell in a table. Another reason is that they use a fairly small corpus (the British National Corpus, BNC), compared to all the data available from Google. The BNC is tagged with linguistic tags that make it possible to identify direct and indirect objects of a verb, which improves the quality of the data, but at the expense of retrieving only a few examples. Using a Google search would give many more examples, but the numbers would include some wrong ones. However, it should be noted that tagging is assigned

automatically and still may contain errors. Table 2 gives their (ibid.) data for ditransitive (e.g. *give me the money*) or to-dative (e.g. *bring the money to me*) , for *give* and other verbs. The table shows that *give* has a higher proportion of its examples in the ditransitive construction than other verbs (taken together). The first question is if this is a significant observation.

Table 2
Distribution of verbs in the ditransitive and the to-dative construction

	Verbs	
Construction	**give**	**Other**
Ditransitive	461	146
To-dative	574	1773

Instead of using Fisher's Exact test, it is entirely possible to use a chi-square test on a contingency table (i.e. a *cross-tables* test) especially since there are enough examples to find significance anyway. Remember that significance only relates to the question if we have made an observation or not. In most experiments there are still other reasons why significance is not the whole story, for example there might be hidden variables that could make the association go in the opposite direction (cf. Simpson's paradox). Bresnan et al. (2007) and Jenset and Johansson (2013) perform more careful analyses that include more factors, but in general few studies can include all relevant factors, just because the relevant factors are not known, but may emerge with more experience and more detailed investigations. One such factor is if the recipient or theme are expressed using pronouns or full noun phrases, and also how syntactically 'heavy' the noun phrases are. There are also effects for how conventional, thus expected, the act is.

The result of a cross-tables analysis is highly significant, a chi-square statistic of 559.5, and p approaches 0. I find little use for specifying exactly how low the probability that the observation can be explained by random chance is, suffice to say that it is very unlikely to happen by chance. However, it is interesting to see the effect size.

Formula 2
Effect size for association between construction and verb

$$\Phi = \sqrt{\frac{\chi^2}{N}}\ ; with\ numbers\ inserted: \Phi = \sqrt{\frac{559.5}{2954}}$$

Note N is the sum of all entries in the table.

Phi= 0.436, which is a *medium* effect size, which intuitively shows that it would not take that many observations to observe that the verb *give* has other preferences for choosing ditransitive than the other investigated verbs together. In fact, we would reach significance ($p<0.05$) with almost one hundredth of the data (shown in table 3), which means that it does not take long before we see it. Note that it is more that the *other verbs* prefer the to-dative, than that *give* prefers the ditransitive.

Table 3
Distribution of verbs in the ditransitive and the to-dative, minimal set for significance

	Verbs	
Construction	**give**	**Other**
Ditransitive	5	2
To-dative	6	20

I hope to have shown that the effect size is a very useful measure for finding out if it takes many or just a few observations to find out about associations in tables like those we have looked at. Significance tells us if we have made enough observations to claim that an observed association is real, or if it can be explained by random chance.

case 3: To compound or not to compound?

An often discussed problem in the Scandinavian countries is the splitting of compounds in writing. This is often discussed as if this were a recent phenomenon resulting from more people writing more in English. Even though compounds are more common in Scandinavian or Germanic languages in general than in English, there are compounds in English as well, such as *railroad* and *toothbrush*. Just to demonstrate, the Google frequencies corresponding to the word *railroad* were looked up, and tested if there were any significant differences in tendency for splitting between the languages. The data is given in table 4. Are the differences significant? Is there a significant relation between language and splitting?

Yes, and impressively so: testing the table (two degrees of freedom) using a cross-tables test gives an incredible chi-square statistic of 1131030, and a p value approaching 0. We can be almost certain that there are differences between the languages.

Table 4
Distribution of realization of a word for three languages

	Realization of Word	
Language	**Correct**	**Split**
Norwegian (jernbane)	2600000	9300
German (Eisenbahn)	21100000	25800
English (railroad)	144000000	7380000

If we look at the proportions in percent (table 5) for each language, we see that the main difference is that English writers tend to split more. Both Norwegian and German writers tend to compound correctly, with more than 99 percent correct, but somewhat surprisingly English writers are only about 95 percent correct. This would indicate that compounding comes less naturally to English writers (as a group, containing all writing in English on the Internet) and that splitting is more of a problem in English than in either Norwegian or German, although much less discussed in English.

Table 5
Proportions for splitting a compound word in three languages

	Realization of Word	
Language	**Correct**	**Split**
Norwegian (jernbane)	99.64	0.36
German (Eisenbahn)	99.88	0.12
English (railroad)	95.12	4.88

So why have the English not discovered the problem of splitting compounds? Again, the answer might lie in the effect size. Phi=0.08, which is a tiny effect size. (Note that adjusting Phi for the table size is not necessary when N is so much larger than number of cells in the table.) You would need to observe on average about a hundred documents about the railroad to find even five documents using "rail road". For English, splitting is the correct alternative most of the time so an erroneous split would have psychological support from a large group of other noun--noun compounds.

Formula 3

Effect size for association between language and splitting

$$\Phi = \sqrt{\frac{\chi^2}{N}}\ ; with\ numbers\ inserted: \Phi = \sqrt{\frac{1131030}{75115100}}$$

Note N is the sum of all entries in the table.

However, in this case it happens to go in the direction of the main alternative, i.e. most English compounds use splitting or noun noun compounding. It seems that expectations set us up to sometimes detecting very small effect sizes. In the case of Norwegian it amounts to a few mistakes in a thousand compounds, on average. What is surprising is how much writing has been dedicated to such a small effect.

2. Discussion

For a difference of four percent, as in splitting for Norwegian and English it would take about 400 observations (i.e. documents if we keep using Google) of compounding, 200 in Norwegian and 200 in English to discover a significant ($p<0.05$) difference in use of compounding for "jernbane/railroad" (and words like *toothbrush* I would guess work similarly), if the tendency to split is fairly uniform within each language.

Effect size is calculated using division of the chi-square statistic by the total number of observations (N), but since the \chi^2-statistic gives one chance for each individual observation to add to a deviation from the expected frequency (i.e. N chances) the N cancels out in the calculation of the effect size. So effect size is thus not dependent on the number of observations.

The effect size additionally gives an indication of how many samples it would take to notice a difference in association.

Johansson (1996) suggested that a useful measure of association for uses in computational linguistics would not depend on the total number of examples. A variant of mutual information was suggested that fulfilled that criterion, i.e. the difference in mutual information between a word sequence in order and reversed order. Such a measure would be very useful for cases like the Google frequencies we have frequently used; because in that case only the number of occurrences is given, not the size of the entire corpus. Jenset and Johansson (2013) suggest modifying mutual information so that more examples are required to report increases in saliency.

In this paper, the effect size has been promoted as a very useful measurement. The effect size also has the property that it is independent of the size of the corpus. In the three discussed cases, we have seen that even very large values for the \chi^2-statistic might be related to a tiny effect size. In fact, the smallest \chi^2-statistic (559.5) showed the largest effect size. This does not matter so much, as significance is only important for deciding if we have an observation of association or not. In fact, almost any difference could become highly significant given enough data.

However, it is conceivable that the effect size could be so tiny that it would take a too large number of individual observations for us to notice any association, despite this association being significant if we have access to enough data. The size of the entire internet would contain more written data than we are likely to read in our entire life.

Maybe this is related to what we could call grammaticality? Maybe grammaticality is not an either-or choice, but a fuzzy division between what we would have a chance to hear (or read) while learning a language? Perhaps in the first ten years of exposure? Then constructions such as those discussed by Stefanowitsch (2006) are not so much an issue of finding negative evidence in a corpus, as being able to find any positive evidence within a reasonable number of observations? Arguing that we could find negative evidence for grammaticality would accept the idea that grammaticality is an either-or choice. It might be a better idea to accept that all sentences are grammatical to some degree, only that some are extremely hard to find any supporting evidence for. Grammaticality is thus related to how readily observable an association is, i.e. the effect size *and* the significance of the observation.

Table 6
Some examples of "ditransitives" for shine

Phrase	**Document frequency of phrase**
"shone him the light"	6
"shine him the light"	1
"shone me the light"	71700
"shine me the light"	29000
"shone me the finger"	1
"shine me the finger"	868

One example discussed by Stefanowitsch (2006) is how we could find out that the verb *shine* is unlikely to involve transfer of an object in the sense of shining the object to somebody, even though we occasionally find phrases like *"he shone me the light"*, or *"shine somebody the finger"*. In table 6 are some patterns with their Google-frequencies. What stands out is perhaps how unproductive the pattern is, but that does not mean that the pattern can never be productive.

When beaming an object from a place to another was introduced in Star Trek in the 1960s it might not have had many attested examples, but with the popularity of the television series, beaming something to somewhere became an accepted pattern, without seriously affecting the main structure of English. Changes like this, however unlikely at first, may indeed gain popularity over relatively short periods of time. Likewise, even if it currently takes effort to imagine shining an object over to somebody called Bill, this is no indication that this is an eternal truth. Therefore, the ungrammaticality of a sentence today, e.g. *"He shines Tony books"* (cf. McEnery and Wilson (2001) and note 8 by Stefanowitsch (2006)) may be due just to the fact that this construction is not in fashion now, but may very well be at some other point in time, and is thus ready to be accepted as soon as the construction becomes popular enough for it to be detected within a reasonable number of observations. A similar point is made for phonological constraints by Pierrehumbert (2000), i.e. how many examples would it take to discover a phonological regularity? This question ties into the level of phonological detail it is possible to maintain in a population of speakers, where speakers have slightly different vocabularies; different in which and how many items are included for each individual. "Phonology is shaped by ecological validity and overly detailed phonological grammars are not viable in a diverse speech community." (Pierrehumbert, 2000).

3. Conclusion

Significance tells us if an observation can be explained by random variation or not. For language data a model assuming a random distribution (e.g. a normal distribution) might not be the best model. It is very rarely that we see words occurring together by random chance, and therefore it is no big surprise that we often find very high significance if only the samples are large enough, analogous to the astronomer finding planets around far away stars using billions of measurements.

Effect size tells us if it would take a lot of measurements or just a few measurements to discover the differences as significant. The higher the effect size the fewer observations we would need to find out the difference.

Acknowledgement

Thank you to Emil, who made me think harder about railroads, toothbrushes, patterns and order.

R code

The statistical package R (R Core Team, 2012) can be used for calculating the cross-tables statistics . The following provides a starting point. The reader needs to insert his or her own data, in order of the columns.

```
data <- matrix(c(1,2,3,4), ncol = 2,
            dimnames = list(c("first row","second row"),
            c("first column", "second column")))
data
chisq.test(data)
Phi = sqrt(chisq.test(data)statistic/sum(data))
Phi
```

References

Bresnan, J., A. Cueni, T. Nikitina, and R. H. Baayen (2007). Predicting the dative alternation. In G.Bouma, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation,* pp. 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Jenset, G. and C. Johansson (2013). Lexical fillers influence the dative alternation: Estimating construc-tional saliency using web document frequencies. *Journal of Quantitative Linguistics* 20(1), 13–44.

Johansson, C. (1996). Good bigrams. In *proceedings of COLING*, pp. 592–597.

McEnery, T. and A. Wilson (2001). *Corpus Linguistics. An Introduction* (second edition ed.). Edinburgh: Edinburgh University Press.

Pierrehumbert, J. (2000). Why phonological constraints are so granular. In A. Cutler, J. McQueen, and R. Zondervani (Eds.), *Proceedings of Spoken Word Access Processes*, Nijmegen, pp. 123–127. Nijmegen: Max Planck Institute for Psycholinguistics.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Ling- uistic Theory* 2(1), 61–77.

Stefanowitsch, A. and S. T. Gries (2008). Channel and constructional meaning: A collostructional case study. In G. Kristiansen and R. Dirven (Eds.), *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, pp. 129–152. Berlin and New York: Mouton de Gruyter.