S Significant

Tom Ruette

(1) important, considerable, with impact
(2) not by coincidence, after careful statistical testing

In the too short period during which I worked at the Humboldt University of Berlin, I had the pleasure to work with Frau Donhauser on a significant—in the first sense—project that had the vision to enable historical linguists to find significant—in the second sense— results in a corpus of Old German texts. The idea was to linguistically annotate Old German texts comprehensively in a highly formalized and computer-readable format, so that morphological, lexical and syntactic phenomena could be retrieved efficiently by means of a search interface on a website. The results were to be available in such a way that philologists could do a qualitative in-depth analysis, whereas corpus linguists could unleash their toolkits for more quantitative analyses. My role in the project was primarily to build that web-based search portal and to interface between the two target groups.

Let me briefly sketch the consequences of having a readily searchable linguistic database of Old German texts from a corpus linguistic perspective. First, we can peruse ancient texts and have the linguistic annotations right at our fingertips. Second, we can practically instantaneously retrieve the text passages in the corpus that contain the phenomenon under investigation—a feature that previously required intimate knowledge of the primary and secondary texts. These two possibilities alone constitute a tremendous step forward, in terms of efficiency, for the historical linguistic community. However, the main achievement from a corpus linguistic point of view is the relative ease with which now the occurrence frequencies of phenomena can be looked up.

Frequencies open up a whole new research dimension, because they allow the researcher to make quantitative claims that can be backed up by well-established statistical techniques that capture the uncertainty in the comparison of numbers. This uncertainty, which is generally known as *statistical significance*, is the main topic of this contribution. Before I go into details, however, a caveat concerning the uncertainty of the numbers themselves, i.e. the frequencies, needs to be expressed. Indeed, the frequencies that fall out of a corpus query are not to be taken at face value.

They are nothing more than a quantitative representation of the qualitative work that has been invested in the linguistic annotations of the texts by the corpus makers. As such, an analyst that directly works on the frequencies without verifying the individual findings in the corpus implicitly subscribes to the interpretation of the corpus maker. The corpus maker therefore has to offer linguistic annotations that are in line with the current linguistic consensus, in case she wants the corpus to be used by the targeted community. One can immediately appreciate that this is, especially in a philological context, a very difficult exercise. By insisting to adopt state-of-the-art methodology in corpus compilation, Frau Donhauser has secured a prosperous and dynamic future for the Old German Reference Corpus.

After this lengthy introduction, let me demonstrate how some of the most basic, but also most insightful inferential statistics can be used to reject a null hypothesis: the Fisher Exact Test and Cramer's V Effect Size. Null hypothesis is a statement that claims that there is no relationship between two observations. As a running example to clarify things, I take an example which Frau Donhauser and I discussed regularly, concerning the dative plural ending, for which Braune (2004: 186) observed that endings in *–m* are older than endings in *–n*. This claim of Braune can be packaged as a null hypothesis:

> H_0: nouns in dative plural do not end more frequently in *–m* than in *–n* if these nouns occur in the early Old German texts, in contrast to the younger texts.

For a statistician, this reads just the same as "test subjects do not recover faster from a cold if they are exposed to an actual medicine, in contrast to test subjects in a control group that received a placebo". So, why now not apply the tested methods from other fields of science in linguistics? The task is thus to decide, on the basis of observed frequencies, if a null hypothesis holds, or if we can reject it. The Fisher Exact test will give you the chance – in the form of a *p*-value – that you observe the same (relative) frequencies, were you to count dative plurals in other Old German texts. The Cramer's V test tells you—in the form of an effect size—the extent of the difference between the levels of the time periods.

For the sake of the argument—which focuses on explaining basic statistical methods, and not a fully fledge historical linguistic investigation—I have counted some observations of dative plural nouns in six smaller Old German texts.* I end up with the following frequencies for dative plural nouns that end in *–m* or in *–n*:

	–m	*–n*
Freisinger Paternoster I (early)	3	0
Fränkisches Taufgelöbnis (early)	5	1
Sangaller Glaube und Beichte I (late)	0	5
Jüngere Bairische Beichte (late)	0	2
Muspilli (early)	0	2
Weissenburger Katechismus (early)	5	0

Table 1: Absolute frequencies for dative plural nouns ending in *-m* or *-n*.

The time indication *early* versus *late* that I added in the table above is based on the Paderborner Repertorium and the *Verfasserlexikon*. Texts that are attributed to a period before the tenth century are called *early* texts, the other texts are called *late*. If we now collapse these frequencies in a two-by-two confusion matrix, we get the following table:

	–m	*–n*
early	13	3
late	0	7

Table 2: Confusion matrix of the dative plural noun ending and the age attribution of the texts.

Obviously, this table cannot be held as representative for the whole of the Old German period, with only six texts being considered. Nonetheless, with the purpose of this text being a basic explanation of the concept of *statistical significance*, this confusion table will do nicely to support the argument. Foreshadowing the remainder of the text, I will now explain why a Fisher Exact test yields a *p*-value of less than 0.0005—giving confidence that the null hypothesis can be rejected—and a Cramer's V test yields a value of 0.75—indicating a substantial effect size and a strong association between text period and dative plural noun ending.

Explaining the mathematics behind these two tests is not relevant. Rather, it is important to convey the intuition behind these tests and their outcomes, and how to interpret them. Let me start with the Fisher Exact test. This test—which is in intuition quite similar to the more widely known Chi Squared test—will return the probability that the ratio of the frequencies in the confusion matrix is due to chance. This probability is known as the *p*-value, and is commonly required to be less than 0.05,

because then the chances that the finding of a correlation between the factors is accidental are less than 5% and can be neglected.

The intuition behind an effect size, measured by the Cramer's V test, is that it returns the extent of the association between the two factors. If one were to quantify the amount of days it takes to heal from a cold with the help from an actual medicine and a placebo, the effect size would indicate the difference in days to get over the cold with medicine versus placebo. The Cramer's V test returns a number between 0 and 1, with zero indicating a practically non-existing effect size. A Cramer's V over 0.25 is typically already considered to be an indication that one factor has a considerable effect on the other.

Notice that significance is different from effect. Since a *p*-value is a calculation of chance, it is inversely related to the amount of observations one has. Therefore, the more observations one has, the higher the chance to find smaller *p*-value. The effect size does not rely on the amount of observations. Yes, one is more certain about the accuracy of the effect size if the amount of observations is larger, but one can—but shouldn't—calculate an effect size for a single observation. Vice versa, a highly significant finding can yield a ridiculously small effect size. Tersely, one could say that any finding can be made significant ($p < 0.05$) if one has enough observations; to get a grasp on the impact of a finding, one should report its effect size.

So, let me wrap up this contribution by relating the project that I worked on under the supervision of Frau Donhauser with the concept of significance. It will take many years to accumulate enough evidence to label this project as statistically significant. However, the effect size of this project, and by proxy, of the scientific vision of Frau Donhauser, is instantaneously established.

* For this example, we rely on the frequencies that were gathered in Ruette and Speelman (2015). There, the precise method for obtaining these frequencies from the reference corpus of Old German is offered, including an explanation to gain access to the corpus and how to perform queries.