DISCUSSION NOTE

Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff*

STEFAN TH. GRIES

1. Introduction

In this issue of *Corpus Linguistics and Linguistic Theory*, Adam Kilgarriff discusses several issues concerned with the role of probabilistic modelling and statistical hypothesis testing in the domain of corpus linguistics and computational linguistics. Given the overall importance of these issues to the above-mentioned fields, I felt that the topic merits even more discussion and decided to add my own two cents with the hope that this discussion note triggers further commentaries or even some lively discussion and criticism.

The points raised in Kilgarriff's paper are various and important and considerations of space do not allow me to address all of them in as great detail as they certainly deserve. I will therefore concentrate on only one particular aspect of the paper which I find − given my own research history and subjective interests − particularly important, namely the issue of statistical hypothesis testing. More precisely, I will address one of the central claims of Kilgarriff's paper. Kilgarriff argues − apparently taking up issues from methodological discussion in many other disciplines (cf. section 2) − that the efficiency of statistical null-hypothesis testing is often doubtful because (i) "[g]iven enough data, H_0 is almost always rejected however arbitrary the data" and (ii) "true randomness is not possible at all". In information-retrieval parlance, null-hypothesis significance testing when applied to large corpora yields too many false hits.

In this short discussion note I would like to do two things. First, I would like to make a few suggestions as to what I think are the most natural methodological consequences of Kilgarriff's statement and several other points of critique concerning null-hypothesis significance testing raised in other disciplines.[1] Second, I would like to revisit one of the examples Kilgarriff discusses in his paper to exemplify aspects of these proposals and show how the results bear on corpus-linguistic issues.

Corpus Linguistics and Linguistic Theory 1−2 (2005), 277−294 1613-7027/05/0001−0277

2. On the utility of null hypothesis significance testing

In several scientific disciplines such as, for example, statistics, medicine, educational science, psychology, psycholinguistics etc., the paradigm of null-hypothesis significance testing has already been criticized vehemently for many decades; cf. Cohen (1994) and Loftus (1996) for some overview and insightful discussion. For reasons of space I can only present a list of the points of critique that are frequently leveled at null-hypothesis significance testing rather than discuss them at any length (cf., e. g., Loftus [1991, 1996] for details):

- H_0's cannot really be true in the first place if only because, e. g., at some decimal place there will be a difference (e. g., between means); from this is follows that
- a large enough sample will always generate a significant result;
- null-hypothesis testing as such does not provide evidence about the pattern that was actually obtained;
- null-hypothesis testing dichotomizes the hypothesis space into 'true' and 'false' as if 0.05 was more than just some convention;
- assumptions underlying null-hypothesis significance testing may in fact influence theory formation (cf. Loftus 1995, section 5);
- null-hypothesis significance testing does not distinguish p(observed data | H_0) from p(H_0 | observed data).[2]

The above-mentioned authors and others have proposed a variety of suggestions as to how a more appropriate way of data analysis may look like, their proposals include, but are not limited to, the following (cf. the above studies for references and discussion):

- do not take conventional threshold values (p = 0.05, etc.) too seriously;
- provide measures of the relative size of an effect;
- provide confidence intervals;
- use techniques from exploratory data analysis (plotting, lattice/Trellis graphs, etc.).[3]

Many of these proposals of course also apply to the domain of corpus linguistics. In the following section I will revisit Kilgarriff's experiment concerning the application of null-hypothesis testing to (i) try to replicate his results and (ii) determine whether adopting the latter three of the above proposals yields interesting findings and more adequate ways of representation.

3. Revisiting word frequency tests

3.1. Methods

One of the examples Kilgarriff discusses in section 4 of his paper is concerned with comparing the frequencies of words from two randomly sampled subsets of the British National Corpus (BNC). The figures he reports boil down to the fact that the null hypothesis is in fact more often rejected than the random sampling of the subcorpora would lead one to believe. While I do not wish to challenge this point in general and will actually even substantiate parts of it below, I have some quibbles with some aspects of the experiment and the conclusions he draws.

First, I am not quite certain how much one would want to claim that because many word frequencies differ significantly across the two subcorpora, this shows that null-hypothesis testing does not make sense. The fact that there is one random sampling of two subcorpora from a larger corpus does not necessarily entail that one would not expect any significant differences. For example, there is a probability that the single sampling of the two subcorpora is skewed just by chance in much the same way that even a fair coin may yield heads up six times in a row, which is why a larger number of sample comparisons should have been performed to determine whether the distribution he observed in his single comparison can be replicated. In fact and as Kilgarriff is certainly aware, from the characteristics of the 820 files he used it would be possible to compute the chances of a relatively unbiased sampling by means of the multinomial distribution.

Another point, however, is more relevant here. Kilgarriff uses a measure of statistical significance, more specifically the contributions to chi-square values. As is well-known, chi-square values, from which one can compute *p*-values (at, in this case, one degree of freedom) are strongly correlated with the sample size *n*, i. e. with the sum of all observed values in the 2×2 table. This is not problematic here as such since all 2×2 tables from word-vs.-word comparisons per pair of corpus files entering into the simulation have the same *n*. Nevertheless as Kilgarriff points out, the large amounts of data often available in corpus linguistics increase the chi-square values, and thus decrease the *p*-values, but they do not tell us what we are really interested in, especially if one believes that the null hypothesis is not a realistic alternative to the alternative hypothesis anyway. What we should be more interested in, and here is the obvious connection to the preceding section, is whether the effect we doubtlessly expect given the sample size is one that matters practically, i. e., an effect size.

In order to determine whether Kilgarriff's results can be replicated and whether a different approach would do away with the undesirably large

number of false hits, I conducted a replication of Kilgarriff's study, going beyond it in a few ways. I generated word frequency lists of the ten largest files in the BNC World Edition.[4] Then, I compared each frequency list with each other frequency list such that for *every* word that occurred in at least one of the two lists, I checked whether the word occurred in one of the two frequency lists significantly more often than in the other (cf. panels 11 and 12 of Figure 1 in the appendix for the graphical representation of the frequencies of all words included). The advantage of this permutation approach is that whatever conclusions we arrive at, they are based on 45 comparisons of pairs rather than just a single one and thus potentially more reliable. To that end I computed

(i) a chi-square test on the same kind of 2×2 table as Kilgarriff;
(ii) the *p*-values following from the chi-square values plus a correction for multiple *post-hoc* tests;[5]
(iii) Cramer's *V*, a measure of correlation that is uninfluenced by frequency, which is computed according to following formula and thus potentially avoiding this apparent weakness of chi-square test and their *p*-values:

$$V = \sqrt{\frac{x^2}{n \cdot (\min [k, m] - 1}}$$; where k and m refer to the number of rows and columns of the 2×2 table respectively (cf. Agresti 2002:112);

(iv) *d*, a standardized measure of effect size proposed by Cohen, which is computed according to the following formula:

$$d = 2 \cdot \sqrt{\frac{x^2}{n - x^2}}$$; (cf. Rosenthal, Rosnow, and Rubin 2000:15)

(v) *d**, a standardized measure of effect size which is computed according to the following formula:

$$d^* = \sqrt{\frac{3}{\pi}} \cdot (\ln(a) + \ln(d) - \ln(b) - \ln(c))$$; where a, b, c, d as usually refer to the cells of a 2×2 table (cf. Hasselblad and Hedges 1995).

Finally, I inspected these values resulting from the word-vs.-word comparisons in each pair; the desired outcome is that the replication should yield only a modest proportion of significant and/or relevant results because of the exhaustive permutation of subcorpora from the same cor-

pus. Since the total number of these comparisons across all 45 frequency list pairs is very high (1,208,819, to be precise), in what follows I discuss the results of these steps rather summarily; cf. Table 1 and Table 2 in the appendix for more specific results from all number of word-vs.-word comparisons per file pair as well as Figure 1 in the appendix for histograms of variables as well as scatterplots of the variable combinations most relevant for the discussion below.

3.2. Results

There are several important results for our present purposes. First, Kilgarriff is supported in the sense that the number of significant results is much higher than one would expect on the basis of chance alone, which would result in a potentially very high number of false hits: While the median chi-square value is 1.4 (interquartile range = 2.245) and thus relatively small, the mean is much larger (arithm. mean = 6.133 $\pm$ 0.094; 95% CI), and out of all individual comparisons that were performed, 247,223 ($\approx$ 20.45%) reached the standard 5% level of significance (cf. the long left tail of points representing small *p*'s in panel 1 of Figure 1), which shows that − in obvious accordance with Kilgarriff − null hypothesis testing by means of chi-square tests does not appear to be a truly fruitful strategy for the word-frequency comparison of corpora.

Second, however, these undesired results change considerably once we do something that most corpus linguists seem to do rather rarely − and I explicitly include my own previous work here − namely apply a correction for multiple *post-hoc* testing. One frequently used procedure is the Bonferroni approximation to the exact formula by Duncan, but since it has proven to be a very conservative correction I use Holm's sequentially rejective approach here (cf. Wright 1992 for an overview). The surprising result is that, once Holm's post-hoc correction is applied to the *p*-values derived from the chi-square tests, the p_{corr}-values are dramatically raised and the average rate of significant word-vs.-word comparisons across the 45 file comparisons is reduced to 3.7% ($\pm$ 0.52%; 95% CI), which is rather close to the 5% level we should have reached if our data were from a really random distribution; cf. panels 3 and 5 of Figure 1 for the distributions of p_{corr} as well as p and p_{corr}. While these results do of course not contradict the general methodological point that with a large enough sample we will get a significant result, the results do contradict Kilgarriff's statement that "arbitrary associations between word frequencies and corpora are systematically non-random": Once the prescriptively correct statistical procedure is adopted, we get the desired result that the random baseline of false hits is not exceeded even with chi-

square tests on word frequencies, a tendency which has so far been overlooked in much discussion of at least the statistical word frequency analyses of the present kind.[6]

The third set of findings relevant at present emerges from inspecting the values of Cramer's *V* as well as the effect sizes *d* and *d** (cf. panels 7, 8, 9, and 10 for the relationships between the *p*-values and *d* and *d**). As it turns out, the overall median of all the Cramer's *V* values is rather small: 0.0015 (interquartile range = 0.0011; arithm. mean = 0.0021 ± 4.43E-05; 95% CI). More precisely, 95% of the Cramer's *V* values are equal to or smaller than 0.0054, and 99% are equal to or smaller than 0.011. This small value clearly underscores that, even though the chi-square values and the corresponding *p*-values may indicate significant differences in many word-vs.-word comparisons, the vast majority of these are practically of a rather limited importance and could thus be omitted from consideration.

Similar results can be gleaned from *d*. The median resulting from all 1,208,819 comparisons is 0.003 (interquartile range = 0.0022; arithm. mean = 0.0043 ± 8.87E-06; 95% CI). More specifically, 95% are equal to or smaller than 0.0108, and 99% are equal to or smaller than 0.0226. Since, as a (much-debated) rule of thumb at least, one assumes that an effect size of $d = 0.2$ corresponds to a weak effect (cf. Cohen 1969:23), the present effect sizes again show that by far most of the observed differences are practically rather irrelevant (cf. the left peak in panel 2 of Figure 1). In fact, the mean *d* observed in the comparison of the subcorpora are close to random *d*'s: I did a small simulation in which I generated 2,000 random 2×2 tables with $n = 100{,}000$ and 2,000 random 2×2 tables with $n = 10{,}000$ and computed the average *d* across all these tables. The median *d* is 0.007 (interquartile range = 0.011; arithm. mean = 0.0104 ± 0.0003; 95% CI), which underscores the fact that the word-vs.-word comparisons did in fact mostly result in the desired practically irrelevant differences predicted by the null hypothesis given our random sampling.

The results are different for *d**, however, even though *d* and *d** are positively correlated to some degree (cf. panel 6 of Figure 1). Looking at the absolute values − since we are only concerned with the strength of the effects and not their directions − the overall median is 0.272 (interquartile range = 0.462; arithm. mean = 0.43 ± 0.0008; 95% CI); 95% are equal to or smaller than 1.313 and 99% are equal to or smaller than 1.932; cf. panel 4 of Figure 1 for an overview. This is not only higher than one would assume random *d**'s to be intuitively − pointing to undesired findings given the random sampling − it is also higher than

the median d^* obtained from the same simulation as conducted for d above: median abs(d^*) of 0.008 (arithm. mean = 0.011 ± 0.0004; 95% CI).

Finally, a few words concerning the corpus file sampling. As shown in Table (i), the corpus files come from different media, domains, and genres. Interestingly, the present data set allows us to at least approach the issue whether the whole enterprise is worthwhile anyway: For example, we have data from five 'periodical' files and from five 'miscellaneous-published' files. If we look at the mean values (chi-square, Cramer's V, p_{corr}, d, and d^*), we can now determine

- how similar the files from different media/domains/genres are to each other;
- which measure makes this effect most visisble.

And, we can of course do the same for domains and genres; the expected finding is always that, if such a word-vs.-word comparison is a sensible thing to begin with, files from the same medium/domain/genre should be more similar to each other than to other levels of the same factor (i. e., exhibit larger p^*s and lower effect sizes). I have to restrict myself to referring the reader to the results given for the medium (periodicals vs. miscellaneous published) in Figure 2 in the appendix; suffice it here to mention that, e. g., chi-square and p_{corr} fail the test (since 'miscellaneous-published' files are on average least similar to each other) while the other measures yield the 'right' results.[7]

4. Conclusion

Interestingly, the results convey a rather mixed and confusing picture. On the one hand, it is obvious that null-hypothesis significance testing has a variety of well- and long-known shortcomings as illustrated for decades in other discplines, and the present paper by Kilgarriff (as well as sections 2 to 2.2 of Kilgarriff 2001) illustrates one such shortcoming by demonstrating that he obtained many significant results where he shouldn't have.

On the other hand, it has hopefully become clear that the story is not quite that simple. Scholars in other disciplines, who noticed many problems with null-hypothesis significance testing, have long ago proposed several ways to overcome these and other problems and analyze the data more fruitfully. As a result of utilizing some of the methodological changes as discussed in other disciplines, the present small-scale case study has shown that corpus linguists appear to have more than one out of several theoretically conflicting ways to arrive at the desired results:

(i) using a null-hypothesis testing paradigm produces the correct quantitative results but incorrect qualitative results: once the prescriptively correct technique of multiple corrections for *post-hoc* testing is applied, even the less-than-optimal chi-square test used on random subcorpora yields the desired proportions of significant results, namely ones that conform rather well to the expected proportion of false hits. Qualitatively, on the other hand, the significance tests may be suboptimal for several reasons (dependence on *n*, sensitivity to small *n*) and yielded the wrong results for the comparison of media.
(ii) using a proposed alternative to null-hypothesis testing produces good quantitative results for Cramer's V and d and not so good results for d* and at the same time qualitatively good results across all measures: once effect sizes are computed, the effect sizes *d* and Cramer's *V* yield the desired small effect sizes (and, thus, practical irrelevance) than one would expect to result from the random sampling and that may in fact even be fairly close to those expected from chance. On the other hand, the effect size *d** yields results which appear to indicate that the corpus files are more different than one would expect and hope for in this simulation but all effect sizes produced the desired results when applied to the comparison of the two media and do away with many null-hypothesis testing problems.

On the basis of these results, it is very difficult to decide what the right quantitative apporach in such word-frequency studies may be. What is not so difficult is to notice that much more exploration in these areas is necessary and that the results show how much we may benefit from taking up methodological proposals from other disciplines to refine, and add to, our methods (cf. for several other examples, Gries, to appear). However, the question that remains is: Do the points of critique and the proposals in section 2 as well as the present findings also mean that we as corpus linguists should more or less abandon null-hypothesis significance testing? Does all this mean we should now turn to effect sizes − if so, which? −, Bayesian statistics, exploratory data analysis, confidence intervals, Trellis graphs, likelihood measures, information criteria AIC/BIC etc. instead? Or is this not an either-or question anyway and we should *always* do both? Comments and proposals are welcome …

Received August 2005
Revisions received August 2005
Final acceptance August 2005

Max Planck Institute
for Evolutionary Anthropology,
Leipzig

Notes

* I thank Dagmar Divjak, Anatol Stefanowitsch, Stefanie Wulff, and particularly Stefan Evert as well as Daniel Stahl for their input and feedback; the usual disclaimers apply.

1. I would like to stress from the outset that these proposals are neither my own nor particularly new (cf. the references mentioned in the following section). It seems, however, as if these proposals, which are much more lively discussed in the disciplines of psychology or statistics, as well as others have not yet made their way into corpus linguistic methodology.
2. A quasi intermediate position is taken by Dixon (1998), who illustrates how *p*-values from several traditional null-hypothesis tests (*z*-tests, *t*-tests, sign tests) are in an approximate linear relationship with likelihood ratios that allow us to compare different hypotheses with each other and, thus, further our understanding in a way traditional null-hypothesis significance testing often cannot.
3. This last point is discussed particular lucidly in Maindonald and Braun (2003); one of the standard references is Tukey (1977).
4. Cf. Table (i) for the files that were analyzed.

Table (i). *The corpus files from the BNC World Edition comparisons inlcuded in the simulation plus some of their characteristics (from David Lee's BNC Index)*

File	Medium	Domain	Genre
HJ0	miscellaneous published	social sciences	non-academic: social sciences
HJ1			
HHV		world affairs	Hansard extracts
HHW			
HHX			
K97	periodical		broadsheet: miscellaneous
K5D			broadsheet:
K5M			reportage
HH3			non-academic: politics/law/education
CRM		natural science	non/nat/sc

5. In order to replicate Kilgarriff's experiment properly, I also include low-frequency words (his frequency list includes at least frequency ranks up to 81,920 items) and also use the chi-square test for 2×2 tables, although it is only reliable if $n \geq 8$ and $p_{\text{rarer alternative}} > 0.2$ (cf. Camilli and Hopkins 1979), which will be violated for some comparisons of infrequent words. I cannot investigate here to which degree this choice of test alone has biased Kilgarriff's data in a from the point of the null hypothesis undesirable way.
6. This is not to imply that post hoc corrections are without problems since, for example, the loss of power resulting from post hoc correction could also undermine an empirical design considerably; this issue is still hotly debated in the pertinent literature.

7. Unfortunately, considerations of space do not allow me to discuss the multitude of additional results obtained from the simulation. Gries (2005) discusses the corpus part-specific results concerning the media/domains/genres in more detail on the basis of a resampling approach and I am currently investigating the words identified as important by the Holm-corrected *p*-values to those considered important by the effect sizes to hopefully arrive at a more refined picture. Also, I must leave aside first meta-analytic results concerning the identification of reasonable sizes of effect sizes to be expected in corpus linguistic studies (cf. Gries and Stahl, in prep.).

Appendix

Table 1. *Results of the corpus comparisons concerning p-values. First cell from top: number of word-vs.-word comparisons per pair of corpus files, second cell: relative frequency of significant word-vs.-word comparisons, third cell: relative frequency of significant (at $p < 0.05$) word-vs.-word comparisons after Holm's correction for multiple post-hoc testing*

	HH3	HHV	HHW	HHX	HJ0	HJ1	K5D	K5M	K97
CRM	34,678 23.40% 4.38%	28,775 31.28% 8.28%	27,913 27.95% 6.46%	29,457 30.88% 7.45%	26,262 23.79% 4.84%	26,351 23.63% 4.5%	33,613 27.97% 5.52%	32,806 28.15% 5.50%	38,153 25.43% 5.42%
HH3		27,606 22.01% 3.60%	26,829 17.35% 2.42%	28,085 19.97% 2.95%	26,971 17.67% 3.38%	27,000 18.24% 3.31%	31,423 14.86% 1.74%	30,687 15.34% 1.84%	35,601 15.90% 1.82%
HHV			19,477 14.31% 1.83%	20,684 13.62% 1.54%	20,391 25.90% 5.40%	20,559 25.44% 5.67%	25,559 22.35% 3.57%	24,741 21.38% 3.64%	30,217 25.82% 3.98%
HHW				20,011 10.21% 0.66%	19,540 21.36% 4.79%	19,699 20.44% 4.78%	25,054 18.91% 2.60%	24,204 18.44% 2.66%	29,664 18.72% 3.07%
HHX					21,093 24.98% 4.99%	21,212 23.74% 4.87%	26,164 20.94% 2.95%	25,325 20.68% 3.08%	30,761 20.91% 3.57%
HJ0						17,244 9.74% 0.99%	25,550 20.16% 4.05%	24,666 20.34% 4.19%	30,222 18.90% 4.08%
HJ1							25,666 19.68% 3.88%	24,752 19.63% 4.00%	30,340 18.60% 3.94%
K5D								27,826 8.07% 0.23%	33,333 12.83% 1.22%
K5M									32,655 13.23% 1.27%

Table 2. *Results of the corpus comparisons concerning effect sizes. First cell from top: mean Cramer's* V*; second cell: mean* d*; third cell: mean of absolute values of* d*

	HH3	HHV	HHW	HHX	HJ0	HJ1	K5D	K5M	K97
CRM	2.43E-03 0.0049 0.4362	2.59E-03 0.0052 0.6290	2.63E-03 0.0053 0.4923	2.53E-03 0.0051 0.5926	2.60E-03 0.0052 0.4886	2.61E-03 0.0052 0.4927	2.50E-03 0.005 0.4661	2.54E-03 0.0051 0.4632	2.39E-03 0.0048 0.4983
HH3		2.07E-03 0.0041 0.5106	2.03E-03 0.0041 0.3683	1.98E-03 0.004 0.4706	2.36E-03 0.0047 0.3878	2.33E-03 0.0047 0.3857	1.89E-03 0.0038 0.3340	1.93E-03 0.0039 0.3230	1.83E-03 0.0037 0.3761
HHV			1.53E-03 0.0031 0.3581	1.40E-03 0.0028 0.2809	2.33E-03 0.0047 0.5525	2.33E-03 0.0047 0.5582	1.94E-03 0.0039 0.4599	1.97E-03 0.0039 0.4737	1.97E-03 0.0039 0.4532
HHW				1.36E-03 0.0027 0.3140	2.52E-03 0.005 0.4569	2.50E-03 0.005 0.4578	1.96E-03 0.0039 0.3341	2.00E-03 0.004 0.3509	1.95E-03 0.0039 0.3758
HHX					2.29E-03 0.0046 0.5179	2.27E-03 0.0045 0.5212	1.88E-03 0.0038 0.4285	1.91E-03 0.0038 0.4398	1.93E-03 0.0039 0.4226
HJ0						1.89E-03 0.0038 0.2354	2.40E-03 0.0048 0.4432	2.46E-03 0.0049 0.4381	2.25E-03 0.0045 0.4698
HJ1							2.37E-03 0.0047 0.4431	2.43E-03 0.0049 0.4354	2.23E-03 0.0045 0.4709
K5D								1.48E-03 0.003 0.2492	1.64E-03 0.0033 0.3201
K5M									1.68E-03 0.0034 0.3365

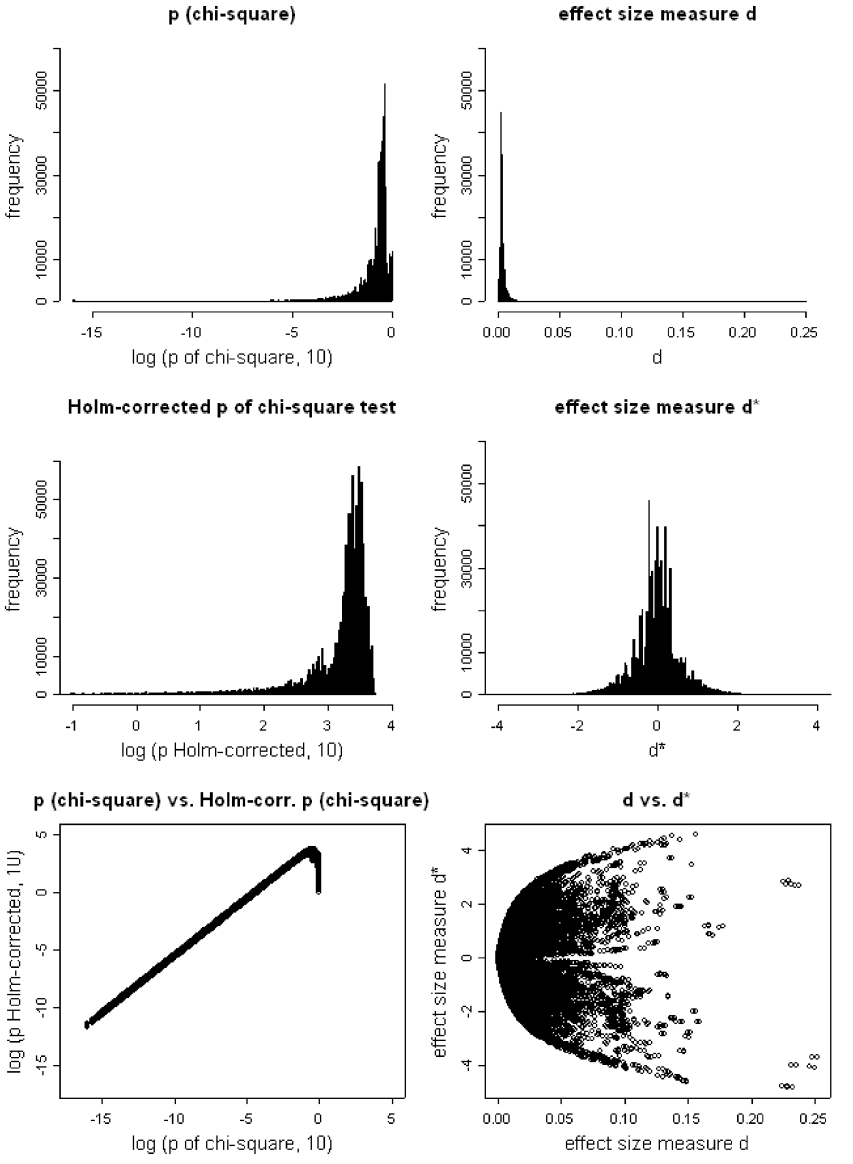


Figure 1. *Histograms and scatterplots for the most relevant (pairs of) variables*

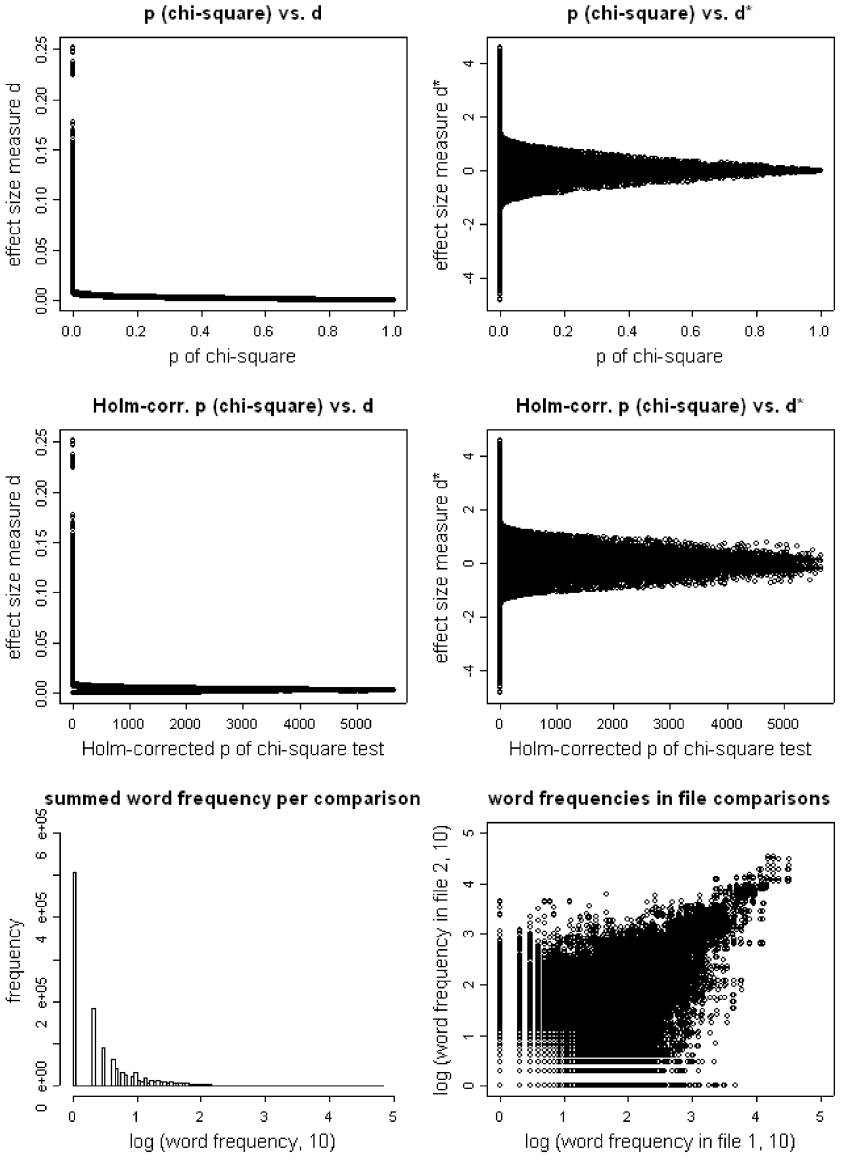
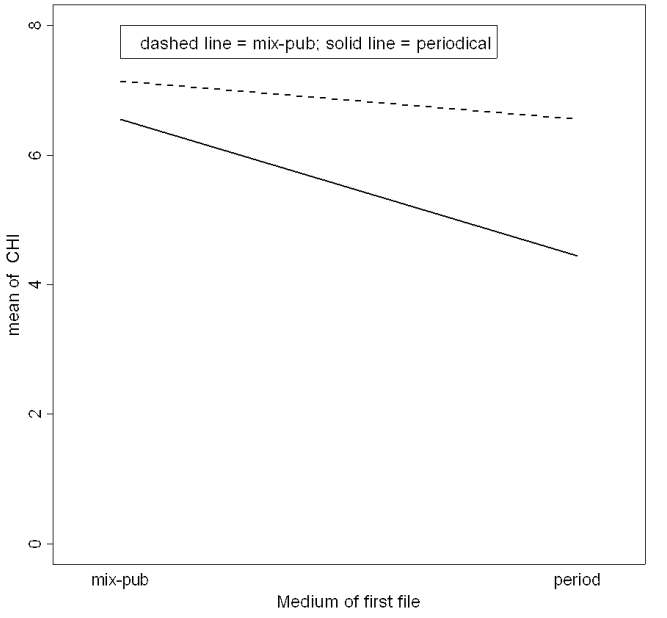


Figure 1. *(continued)*



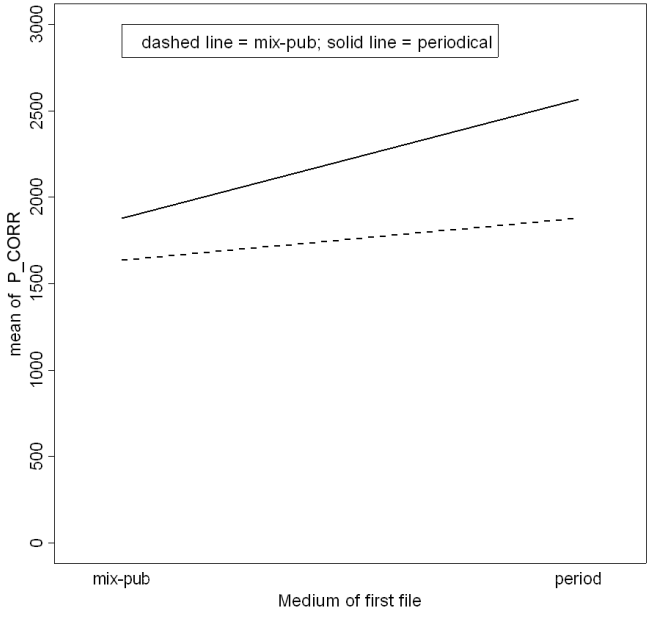


Figure 2. *Interaction plots for the media with chi-square,* $p_{corr,}$ *Cramer's* V, d, *and* d*

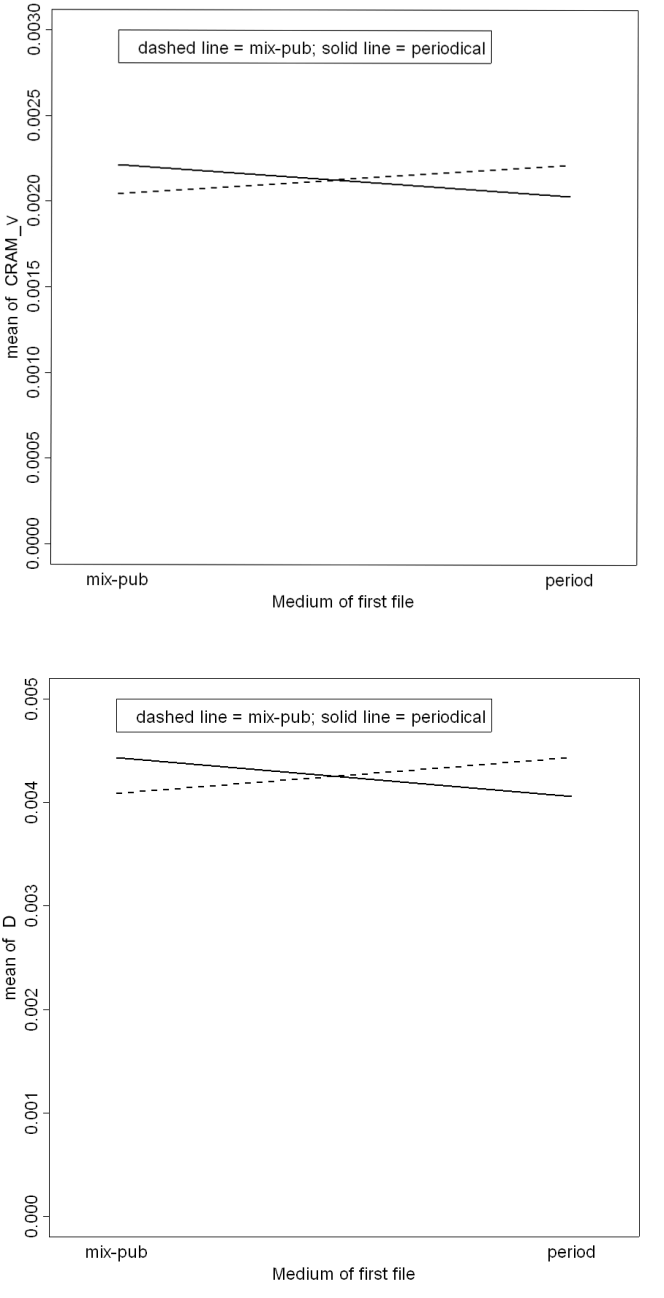


Figure 2. *(continued)*

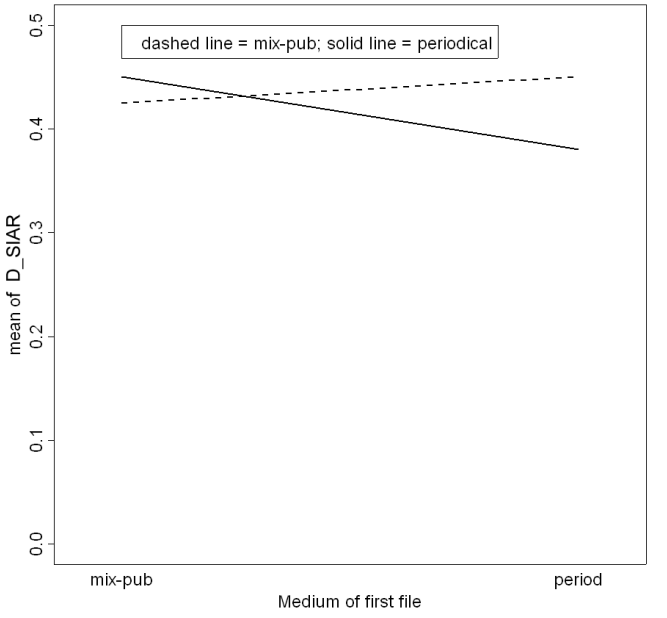


Figure 2. *(continued)*

References

Agresti, Alan
2002 *Categorical data analysis*. 2nd edition. Hoboken, NJ: John Wiley.
Camilli, G. and K. D. Hopkins
1979 Testing for association in 2 × 2 contingency tables with very small sample sizes. *Psychological Bulletin* 86(5), 1011−1014.
Cohen, Jacob
1969 *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
1994 The earth is round ($p < .05$). *American Psychologist* 49(12), 997−1003.
Denis, Daniel J.
2003 Alternatives to null hypothesis significance testing. *Theory and Science* 4.1.
Dixon, Peter
1998 Why scientists value *p* values. *Psychonomic Bulletin and Review* 5(3), 390−396.
Gries, Stefan Th.
2005 Resampling corpora. Paper presented at the workshop "Corpus statistics: objectives, methods, problems", University of Leipzig.
in press Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*.
Gries, Stefan Th. and Daniel Stahl
in prep. Effect sizes in corpus linguistics.

Hasselblad, Victor and Larry V. Hedges
1995 Meta-analysis of screening and diagnostic tests. *Psychological Bulletin* 17(1), 167−178.
Kilgarriff, Adam
2001 Comparing corpora. *International Journal of Corpus Linguistics* 6(1), 1−37.
2005 Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1.2.
Loftus, Geoffrey R.
1991 On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology* 36(2), 102−105.
1996 Psychology will be a much better science when we change the way we analyze data. *Current Directions on Psychological Science* 5(6), 161−171.
Maindonald, John and John Braun
2003 *Data analysis and graphics using R: an example-based approach*. Cambridge: Cambridge University Press.
Rosenthal, Robert, Ralph L. Rosnow, and Donald R. Rubin
2000 *Contrasts and effect sizes in behavioral research*. Cambridge: Cambridge University Press.
Tukey, John W.
1977 *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Wright, S. Paul
1992 Adjusted *P*-values for simultaneous inference. *Biometrics* 48(4), 1005−1003.