Computers and the Humanities **38:** 343–362, 2004.

Pitfalls in Corpus Research

TONI RIETVELD[1], ROELAND VAN HOUT[1] and
MIRJAM ERNESTUS[1,2]
[1]*Department of Linguistics, Radboud University Nijmegen, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands;* [2]*Max Planck Institute for Psycholinguistics*
E-mail: a.rietveld@let.kun.nl; r.v.hout@let.kun.nl; mirjam.ernstus@mpi.nl

Abstract. This paper discusses some pitfalls in corpus research and suggests solutions on the basis of examples and computer simulations. We first address reliability problems in language transcriptions, agreement between transcribers, and how disagreements can be dealt with. We then show that the frequencies of occurrence obtained from a corpus cannot always be analyzed with the traditional χ^2 test, as corpus data are often not sequentially independent and unit independent. Next, we stress the relevance of the power of statistical tests, and the sizes of statistically significant effects. Finally, we point out that a *t*-test based on log odds often provides a better alternative to a χ^2 analysis based on frequency counts.

Key words: agreement between transcribers, χ^2 analysis, corpus research, effect size, log odds, power of a test, sequential dependence, unit dependence

1. Introduction

The use of corpora has become common in language research over the last decades. In many branches of linguistics, corpora provide core data for survey research and for the development and testing of hypotheses. The origins of these corpora can be manifold: texts from the Middle Ages, series of samples from current newspapers, essays written by school pupils, letters written by emigrants to those who stayed behind, transcripts of sociolinguistic interviews or pathological speech, recordings of children's speech, or recordings used in applications of speech technology. Corpora of speech may just include transcripts, but rapid developments in storage capacity and computational power have made the direct availability of sound and video signals a reality (cf. CHILDES, http://childes.psy.cmu.edu/, TALKBANK, http://www.talkbank.org/, and the Corpus of Spoken Dutch, http://lands.Let.kun.nl/cgn/ehome.htm). Research tools have been developed to make these corpora easily accessible (e.g., the tools of the Max Planck Institute for Psycholinguistics in Nijmegen, http://www.mpi.nl/tools, such as the EUDICO linguistic annotator, which allows users to create, edit, visualize, and search annotations for video and audio data).

In spite of the rapid developments in corpus-based research, some basic problems with this type of research have not received the interest they de-

serve. Several pitfalls keep showing up, related both to the transcription and coding of corpus data, and to their analysis, particularly to the statistical analysis of frequency data. In this paper, we address some of the pitfalls.

In Section 2, we start with transcription and coding, where conflicting judgments between experts or evaluators quite often show up, partly as a result of the transcribers' expectations. The degree of conflict can be made clear by calculating agreement indices as will be exemplified. Moreover, we will show how data on which disagreement occurs ought to be dealt with in the analysis.

The statistical analysis of frequency data is the central topic of Section 3. Basically, the analysis of this type of data is fairly straightforward. The primary technique is χ^2 analysis, a technique explained in introductory textbooks on statistics. An important assumption of χ^2 analysis and equivalent statistics like Fisher's exact test and likelihood-ratio tests is the independence of observations, and precisely this assumption is problematic in corpus research. We show how two kinds of dependences may interfere in the statistical analysis, both resulting in a Type I error which is too high; that is to say that the significance of an effect is claimed too often where in fact there is no effect.

Section 4 deals with two other well-known problems in χ^2 analysis, viz. the effects of small and large samples. Small samples tend to yield few significant effects, while the 'high significance' levels obtained with large samples are often incorrectly interpreted as indicators of substantial effects. For small samples the concept of power is relevant. For large samples, we need an index which expresses the size of an effect, independently from the sample size.

In Section 5, we discuss the use of the log odds ratio as an alternative (a sometimes compelling alternative) to χ^2 analysis. Its use is still quite rare in corpus analysis (but very common in medical research), although it has outstanding statistical properties. Log odds form the basis of attractive multivariate techniques, such as logit analysis and logistic regression.

2. Transcription and Coding

2.1. Conflicting judgements

In many cases speech and language data have to be coded before they can be analyzed. Only in a small number of situations the raw data themselves are suitable for analysis. A common coding process consists of the transformation of speech fragments into discrete transcription symbols by listeners. For instance, pitch movements expressed in Hz values are coded into categorical phonological symbols, like H*L, a high pitch associated with an accented syllable and followed by a low pitch target, or phones are coded as IPA symbols. Categorical coding always results in the loss of detailed informa-

tion. Moreover, it often results in disagreement among transcribers. Some transcribers may perceive a schwa between two consonants, while others do not hear anything of the sort.

Transcribing utterances by ear is not an easy task. The transcriber must take note of all the phonetic details produced by the speaker, and decide which symbols should be used to represent the perceived sounds. It is easy to make mistakes, and the task requires great concentration. Above all, transcribing by ear is difficult because listeners normally determine what they perceive not only on the basis of the acoustic signal, but also on the basis of their expectations. While making phonetic transcriptions, transcribers should disregard all the expectations that automatically follow from their knowledge of the phonotactics of the language, the spelling of a word (Cucchiarini, 1993, p. 55), its lexical representation, its pronunciation in formal speech, and so on. Discarding these expectations is difficult, if not impossible (Vieregge, 1987, p. 9), as has been shown in a number of experiments (e.g., Kemps *et al.*, 2004).

Expectations are more prone to affect phonetic transcriptions when the speech signal is less intelligible. Casual speech is generally less intelligible than formal speech, since all kinds of contrasts tend to disappear in this register, making it difficult to distinguish [t]s from [d]s, [t]s from [s]s, and so on, and introducing uncertainty whether vowels and sonorants such as [r] or [l] are present. In casual speech we often find realizations that deviate from their canonical forms. Thus, ironically, the transcribers' expectations have a greater chance to guide perception the more the actual realizations deviate from these expectations.

The difficulty of transcribing casual speech is reflected by the high disagreement among phoneticians in their transcriptions. Ernestus (2000, p. 142) reported in her study on casual Dutch that three phoneticians judged 2136 intervocalic plosives as voiced or voiceless, and disagreed on no less than 322 plosives, that is 15% of the total. Moreover, when transcribing 274 tokens of the word *natuurlijk* (''of course'') with the unreduced form [na'ty:rlək], the three phoneticians agreed on the presence/absence of the first vowel in less than half of the cases (116 tokens). Similarly, Kuijpers and Van Donselaar (1997) reported that their three phonetically trained transcribers generally disagreed in more than 10% of cases on the presence/absence of schwa in Dutch sentences read aloud.

We have to conclude that disagreement among listeners is an inherent characteristic of human coding and transcription. Moreover, agreement between transcribers is no guarantee for valuable transcriptions. Validity is a difficult aspect, as we hardly ever know what the speaker actually realized, or wanted to realize. We must accept, as was also stated by Keating (1998), that pronunciation variability is probably necessarily confounded with transcription variability in studies with human transcribers.

2.2. Assessing agreement

The disagreement between human observers asks for an index on the basis of which we can assess the degree of (dis)agreement, and which also shows whether the agreement is based on chance or not. Note that agreement is not equivalent to reliability. Inter-observer agreement expresses the extent to which listeners agree in their judgments. Reliability expresses the extent to which the error variance is part of the total variance of the ratings. It is a measure of the covariation between the raters' judgments, and is only relevant for ratings expressed at the interval or ratio level (cf. Rietveld and Van Hout, 1993). It is possible to have a low index of inter-observer agreement, and, at the same time, a high index of reliability. This is, for instance, the case in the ratings (1) and (2) of two observers (A and B) on five speech samples. The ratings covary to a large extent (high reliability), while the two observers use different parts of the scale (low agreement).

$$\mathrm{A} : 1, 3, 2, 6, 3 \tag{1}$$

$$\mathrm{B} : 3, 5, 4, 8, 5 \tag{2}$$

Categorical judgments, like +/− voiced, constitute nominal scales. For this type of scales the concept of reliability (covariation) does not make sense. We are left with indices which only express the degree of agreement between observers. We illustrate the use of some of these indices on the basis of the artificial data set in Table I, consisting of the voiced and voiceless scores of two transcribers.

A frequently reported index for agreement is the percentage of agreements between the judges. In Table I, there are $10 + 7 = 17$ disagreements, and $20 + 25 = 45$ agreements between A and B. Thus we obtain a percentage of agreement of $45/(45 + 17) = 71\%$. The percentage of agreement as an index is problematic in two regards (see Rietveld and Van Hout, 1993; Cucchiarini, 1996, p. 137; Carletta, 1996):

– percentage agreement is based on the assumption that agreement between transcription symbols is all-or-none;

Table I. The absolute numbers of plosives scored as voiced and voiceless by transcribers A and B

	Transcriber A		
	Voiced	Voiceless	Total
Transcriber B			
Voiced	20	10	30
Voiceless	7	25	32
Total	27	35	62

– percentage agreement does not enable us to distinguish between agreement due to chance and genuine agreement.

The second problem is the most important drawback of this index: its sensitivity to chance agreement, which depends on the number of alternatives available. The coefficient κ (Cohen, 1960) adjusts for chance agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

P_e is the expected proportion of agreement solely on the basis of chance (cf. Rietveld and van Hout, 1993, p. 219). P_o is the observed proportion of agreement. On the basis of κ and its standard error, a z score can be computed by means of most statistical packages, which indicates whether the achieved agreement is due to chance. For our example κ is 0.449; the associated standard error is 0.113, and $z = 3.973$, which is significant at the 0.01 level. Clearly the agreement between the two transcribers is not only due to chance.[1]

How relevant is a significant κ? Agresti (2002, p. 435) observes: "It is rarely plausible that agreement is no better than expected by chance." He concludes that it is far more relevant to estimate the strength of agreement, by the magnitude of κ (taking into account its confidence interval).

The coefficient κ was developed to assess the agreement between two observers. Fleiss (1971) presented an extension which can be used to assess the agreement between more than two observers. Using more than two observers in a transcription task is not common when a large corpus is processed. However, in clinical applications, with smaller corpora, the use of relatively large panels of transcribers is not uncommon at all (cf. Vieregge and Maassen, 1999). Fleiss' κ is provided by dedicated software (e.g., AGREE, provided by THE SCIENCE PLUS GROUP).

A different situation arises if observers are not asked to detect specific phenomena, like voiced or voiceless segments, but to give a complete description of a speech fragment in terms of transcription symbols. In Table II we give an example in which both transcribers use the same number of symbols; thus we do not have to deal with an alignment problem (cf. Kruskal, 1983; Cucchiarini, 1996).

The transcribers agree in three instances, and disagree in two. The coefficient κ can be used to deal with this kind of data. However, this κ does not take into account that some symbols express more dissimilar speech sounds than others.[2] The difference between symbols like [i] and [ɪ] is intuitively smaller than that between [i] and [a]. Obviously we need a metric to express these differences. One metric is provided by feature counts. Thus, for Dutch the difference between [i] and [I] is a difference of 1 (difference on the tongue height feature "mid"), while the difference between [i] and [a] is 2 (differences

Table II. The transcription of a Dutch sentence fragment *dat is* "that is" by two transcribers, A and B

	Segments				
Transcriber A	d	a	t	ɪ	s
Transcriber B	d	a	d	i	s

on the tongue features "high" and "back"). Weighted κ is an extension of κ in which weights can be assigned to disagreements. For the calculation and mathematical details of weighted κ we refer to Wickens (1989, p. 241).

2.3. Dealing with disagreements

If the number of disagreements is relatively high, we have to decide what to do next. In some studies (e.g., Van de Velde and Van Hout, 2001), the stretches of speech on which the transcribers disagree are replayed, and the judges decide whether they are willing to agree on the same transcription. This method may not *a priori* yield a more valid transcription, as the judges, when listening for the second time, know each other's transcriptions and can be influenced by them, so that the transcription which is eventually accepted may not be the best one, but the one obtained from the most "confident" transcriber. Disagreements may encompass both mistakes on the part of one or more raters and systematic differences between raters. Moreover, it is important to realize that a high degree of agreement does not prove the validity of the ratings involved.

A better method of dealing with disagreements may be to discard the problematic stretches of speech. The researcher, however, should be aware that the removal of problematic utterances can affect the conclusions that are drawn from the data. The number of transcriptions that remains may be too small to warrant firm conclusions. Moreover, a complete category might be removed from the data set, for instance all or most realizations in one condition, or the realizations by one speaker, which diminishes the scope of the investigation. Moreover, statistical analyses may yield significant differences that would not have been found if all stretches of speech had been transcribed unanimously. This may be the case if one condition differently affects the probabilities that the data points are removed from the data set. Since we do not know the "real" classifications, it is difficult to ascertain when we are dealing with such a situation. A possible, but not completely reliable, solution is to discard tokens only if the numbers of agreements and disagreements have the same distribution over the conditions. We can test

Table III. Hypothetical data: the absolute numbers of plosives unanimously classified as voiced or voiceless in conditions A and B

	Condition A	Condition B
Voiced classification	60	52
Voiceless classification	45	71

this assumption of equal distributions by applying χ^2 as a goodness-of-fit test between the distribution of agreements and the distribution of disagreements over the conditions.

We would like to clarify this with the following hypothetical example. Imagine that Table III presents the number of plosives in conditions A and B, which were unanimously transcribed as voiced or voiceless. The difference between conditions A and B is statistically significant $\chi^2 = 5.009$, $df = 1$, $p = 0.025$ without continuity correction; all reported χ^2 values are without continuity correction, see Fienberg, 1980, p. 22). Imagine that Table IV presents the ''real data'', i.e., all plosives, including those that were not transcribed unanimously. The difference between conditions A and B is not significant in this data set ($\chi^2 = 2.720$, $df = 1$, $p = 0.099$). The difference is significant in the transcription data (Table III), whereas it is not in the ''real'' data (Table IV). Apparently, Condition A leads to more disagreements for voiceless realizations than for voiced ones. The factor ''Condition'' appears to affect the transcription of the plosives, but not necessarily their realization as voiced or voiceless. Such an explanation is suggested by the different distributions of agreements and disagreements over the two conditions.

We should calculate a χ^2 on the numbers of agreements and disagreements in the conditions concerned. Only if the assumption of equal distributions is met, we have an argument (but not more than that) to restrict the analysis to the occurrences where the observers agreed in their judgments. In our example, the numbers of agreements in Conditions A and B are 105 and 123, respectively, and the numbers of disagreements are 15 and 3. The corresponding $\chi^2 = 9.280$, $df = 1$, $p = 0.002$. The assumption of equal distributions of agreements and disagreements over the conditions is not warranted.

Table IV. Hypothetical data: the absolute numbers of plosives actually realized as voiced and voiceless in conditions A and B

	Condition A	Condition B
Voiced realization	65	55
Voiceless realization	55	71

If more than two observers are involved, alternative procedures are available. A simple alternative is to make the majority of observers decide, leaving the decision to the researcher (or better: to chance) if votes tie. The outcome then remains a binomial variable. A second alternative, which only makes sense if more than three observers are involved, is to take the relative number of one of the two outcomes as dependent variable. For instance, with four observers, the outcome then varies between 0/4, 1/4, 2/4, 3/4, and 4/4. This procedure transforms the dependent variable from the nominal level to a continuous one. A discussion of the merits and demerits of such a transformation lies beyond the goals of this contribution.

3. Frequency Data and Dependences

3.1. χ^2 Analysis

The data obtained in corpus research is very often of the nominal level: counts of observations (frequencies) in different categories. Examples of this are the contingency Tables III and IV. The default statistical test for frequency data is the χ^2 statistic. This non-parametric statistical test is widely used in sociology, sociolinguistics, and linguistic corpus research.

More than 50 years ago Lewis and Burke (1949) published an article called "The use and misuse of the χ^2 Test". This article was followed by a series of articles defending and criticizing current (at that time) practice (see also Delucchi, 1983). We can still benefit from this debate, which warns against the unthoughtful use of χ^2. The use of χ^2 tests (or equivalent tests like Fisher's exact test or likelihood-ratio tests) is based on the assumption that the data or observations are independent.

This assumption is often neglected in practice though. For instance, researchers normally take more than one occurrence for every speaker or writer into account. The rationale for this approach of repeated sampling is that language and speech are highly varying phenomena, and that the variable of interest may induce variability both among and within speakers and writers. The speaker or writer level normally does not appear in the analysis, and the data obtained from the different speakers or writers are pooled. Two types of dependence may occur in the resulting data set:

- sequential dependences, by which an observation can be predicted by the outcomes of preceding observations,
- unit dependences, which are the consequence of pooling the data from the units used in the data collection.

We discuss sequential and unit dependences (e.g., speakers) and methods to avoid them, in Subsections 3.2 and 3.3.

3.2. Sequential dependences

Observations are sequentially dependent if the category of one token affects the probability that the next token is of a certain category. This is often the case in speech and text, although dependence rapidly falls off with distance (Dunning, 1993, p. 64). To give an example, assume that speakers tend to realize consonant clusters in a specific way depending on how they realized the preceding cluster, for instance, because speakers may try to maintain their mode of speaking. The observations – presence of all or absence of some consonants in a cluster – are then sequentially dependent. Having observed observation i, we can then predict observation $i + 1$ to some extent.

Another well documented example of a phenomenon with sequential dependence is the use of pitch accents, such as H*L (high-low), L*H (low-high), H*LH (high-low-high). Although we do not yet know the transitional probabilities of the different pitch accents, it is clear that they are not equal, as the use of one specific pitch accent seems to bring about the realization of another specific pitch accent (cf. Dainora, 2002).

The sequential dependence can sometimes be characterized by windows within which the dependence is possible, and outside which the dependence can be assumed not to exist. One way to determine the size of the window for binary data is the ONE SAMPLE RUNS TEST (cf. Siegel and Castellan, 1988), which is available in most statistical software packages. This test enables us to detect lack of randomness in the sequence of binary data. The test evaluates the number and length of sequences with the same observations. Let us assume that a researcher wants to know whether H*L accents are more often used than L*H accents. The accents are coded as 1 and 0 respectively. The fictitious data (57 observations, 16 0s and 41 1s) is as follows:

11101110001111110001111001111111001100111111101100111111111

The runs test yields a z value of -2.333, $p = 0.020$, which means that we have to reject the hypothesis of independence between occurrences of the two types of pitch accents. The dependence window is larger than 1. We then try a window of $k = 2$, taking the first observation in each window as the observation to be analyzed (thus always skipping the second observation). The resulting z value of 0.914 is not significant. When the second observation in each window is taken, the resulting z value is 0.986; again this is not significant. Because both tests in a time or sequence window with the length 2, produce a non-significant outcome, assuming a time window of 2 appears to be appropriate. This suggests that the type of pitch accent only depends on the directly preceding pitch accent.

Altham (1979) suggested the following adjustment of χ^2 by the length of the time window:

$$\chi^2_{\text{adapted}} = \frac{\chi^2}{(2K-1)} \tag{4}$$

in which K is the length of the time window, which must be chosen in such a way that dependences between observations are absent (cf. Wickens, 1989, p. 29). This adjustment does not yield an estimate of the true χ^2 statistic, but a lower bound estimate and consequently a conservative p value. In our fictitious example of pitch accents the adjusted χ^2 is $\frac{1}{3}\ \chi^2$. When we test the hypothesis of an equal number of 0s and ls in in a χ^2 one-sample test in our accent example (observed 16 0s, 41 ls; expected 28.5 0s, 28.5 ls), we get a significant χ^2 value of 5.802 ($df = 1$). Assuming a time window of 2, Altham's adjustment gives a value of $5.802/3 = 1.934$, which is not significant at the 5% level.

3.3. Unit dependence

Unit dependences may occur when the units used in the collection do not match those used in the analysis of the data. In corpus research the units of data collection normally are the speakers, or writers, or texts. Most often these units do not return in the analysis of the observed frequencies, and the data of the different units are pooled. The researcher wants to tackle a research question directly, which often boils down to the comparison of groups of speakers (e.g., low educational level versus high educational level), or the comparison of speech conditions (e.g., face-to-face interactions versus formal addresses). Disregarding the actual units of sampling in processing frequencies, however, may cause serious problems, since the observations from the same speaker, writer, or text may be more similar than those from different ones. Disregard of the unit or level of sampling can imply the violation of the assumption of independence.

We would like to clarify this with a hypothetical example. There are two speakers, A and B, who realized specific phonetic sequences in two contexts. In these sequences they could apply a specific assimilation rule. Speaker A applies assimilation with a higher relative frequency in Context 2 than in Context 1 (see Table V). For speaker B, who provided fewer observations, it is the other way round (see Table VI). Contingency Table VII contains the data pooled over speakers A and B.

It seems straightforward to pool over the two participants, and to analyze the resulting contingency table with a χ^2 test. The χ^2 for the pooled data is 4.572, $df = 1$, $p = 0.033$, which is a significant result at the 0.05 level. Thus, on the basis of the pooled data we might think that context affects the occurrence of assimilation. However, this conclusion is not correct, as the pooling procedure is not allowed for three related reasons (cf. Wickens, 1993, p. 192):

Table V. Hypothetical data (frequencies of occurrence): the occurrence of assimilation as a function of context for Speaker A

	Context 1	Context 2	Total
+Assimilation	19	51	70
−Assimilation	59	40	99
Total	78	91	169

Table VI. Hypothetical data (frequencies of occurrence): the occurrence of assimilation as a function of context for Speaker B

	Context 1	Context 2	Total
+Assimilation	22	8	30
−Assimilation	6	12	18
Total	28	20	48

Table VII. Hypothetical data (frequencies of occurrences): the occurrence of assimilation as a function of context; Data pooled over speakers A and B.

	Context 1	Context 2	Total
+Assimilation	41	59	100
−Assimilation	65	52	117
Total	106	111	217

– Each participant is the source of a large number of observations. They cannot be treated as one source of independent observations;
– The association between the two variables – in our case context and assimilation – may vary between participants. Especially if participants do not realize the same number of occurrences, specific participants may dominate the overall results and hide the association realized by the other participants.
– We miss the possible interaction between participant and context in the data.

We illustrate the serious effects of disregarding the speaker or text level with some computer simulations. In each simulation we had 200 observations, half of which came from speakers with a low educational background, the other half from speakers with a high educational background. The basic sampling units were the speakers. Each speaker was assigned a random value using the normal distribution, with a mean of 0 and a standard deviation of 1. To each speaker value generated in this way we applied a logistic function.

The resulting value (p) represents the probability that a certain linguistic phenomenon is present in an observation of that speaker. Since all random values come from the same normal distribution, the speakers in the two groups do not differ in their use of the linguistic phenomenon. We generated observations for each speaker, by means of a binomial distribution with p being equal to the probability of the occurrence of the linguistic phenomenon for that speaker.

In the first series of 5000 simulations, we generated a single observation for each speaker. These simulations represent the situation in which each speaker is the source of one occurrence only, and the occurrences are completely independent. The binomial distribution resulting from these simulations shows that the χ^2 test yields a significant difference between the two groups of speakers in 5% of cases, if the significance level is 0.05. The significance level achieved perfectly matches the a priori type I error level.

In the next series of simulations (5000 per series), we increased the number of occurrences per speaker, while keeping the total number of observations per group constant. Thus, the number of speakers decreased, which implied an increasing violation of the independence assumption. Figure 1 shows the probability of a type I error as a function of the number of observations per speaker in the data set, with the significance level set at 5%. It clearly shows the dramatic effects of the number of observations per speaker or subject. If

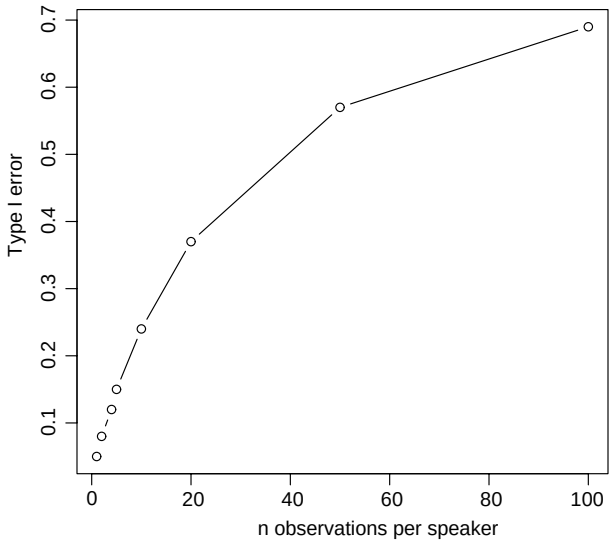

Figure 1. Probability of a type 1 error as a function of the number of observations per subject.

the data set contains 10 observations for every speaker, the probability on a type I error is 24%. In case every group is represented by a single speaker, which means that 100 occurrences per speaker are sampled, the type I error is 70%. Therefore, the researcher claims a difference between two groups in 70% of the cases, whereas in fact there is no difference at all.

Another important factor affecting the type I error is the standard deviation of the distribution which generates the mean values for the speakers. The normal distribution with a mean of 0 and a standard deviation of 1 produces a logistic distribution with a mean value of 0.500 and a standard deviation of 0.209. The standard deviation can be changed stepwise to see what happens to the type I error. We have done this for the situation in which each speaker is represented by ten observations. The results are shown in Figure 2. The standard deviation runs from a value of 0.001 to 10. A standard deviation of 10 in a normal distribution corresponds to a value of 0.459 in a binomial distribution. Increasing the standard deviation strongly increases the type I error.

Note that such an effect is only found when more occurrences per speaker are analyzed as independent occurrences.

The simulations clearly show that the observations of different speakers or writers should not be pooled, as pooling results in an unacceptably high Type I error. Two solutions are possible: Researchers restrict themselves to one observation for every speaker or writer, or they apply different statistical approaches, such as the ones that we discuss in Section 5.

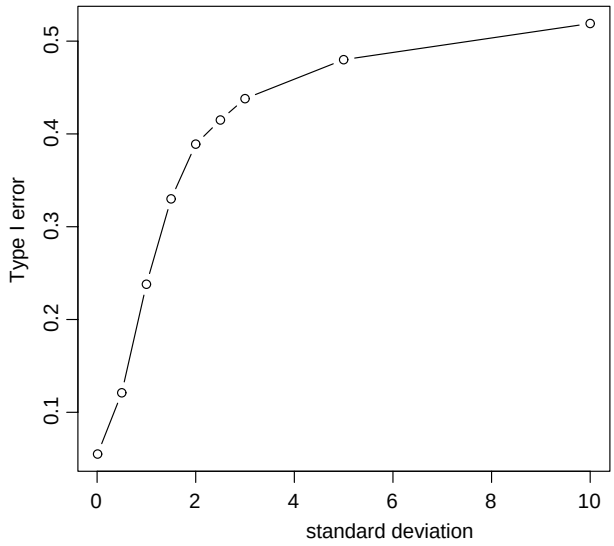

Figure 2. The probability of a type 1 error as a function of the value of the standard deviation.

4. χ^2 and Sample Size

4.1. Small *n*

Speech corpora are very often explored to answer specific questions, such as the occurrence of assimilation in specific contexts (for instance, in non-accented bisyllabic content words with specific consonant clusters) or some morphosyntactic phenomena (*they* vs. *them* in subject position). The phenomena of interest to the researcher may seldom occur in a corpus, and as a consequence the research is based on a very small number of occurrences only. In these cases, the power of the statistical test is relevant. The power of a test is the probability that it detects an effect which is present in the population(s) under investigation. It is dependent on four factors:

- the effect size to be detected (for instance a small effect of 10% points versus an effect of 30% points);
- the adopted significance level;
- the variation in the populations at issue;
- the sample size.

Our illustration of the effects of effect size and sample size on power is based on hypothetical data obtained by sampling two subpopulations and recording the occurrences of a phenomenon, such as assimilation. According to the H_0, the relative frequencies in both subpopulations are 0.50. In the first example the effect to be detected is 0.10: In population 1 the relative frequencies of occurrence of + and – assimilation are 0.50, whereas in population 2 they are 0.60 and 0.40, respectively. The effect size of the second example is 0.20: In population 1 the relative frequencies are 0.50, whereas in population 2 they are 0.70 and 0.30, respectively.

In Figure 3, we show the effect of sample size – ranging from 20 to 100 in each sample of the two subpopulations – on the power for 2×2 contingency tables (the effects were calculated with the package SamplePower of SPSS). The α level was set at 0.05, one-tailed. The figure illustrates the importance of obtaining relatively large samples. In order to detect an effect size of 20% (50% vs. 70%), a sample size of about 70 is needed to achieve a reasonable power of 80%. If the effect size is 10%, a much larger sample size is necessary. Small samples can only reveal large effects, and non-results may not be very informative.

4.2. Large n

Research of written language is generally based on very large corpora, like records of books, magazines and daily papers. Sizes of 1,000,000 tokens are no exception. Statistical tests are powerful enough to reveal even very small effects in such populations. Here researchers face another problem. The significance levels of χ^2 tests cannot be considered as indices of the

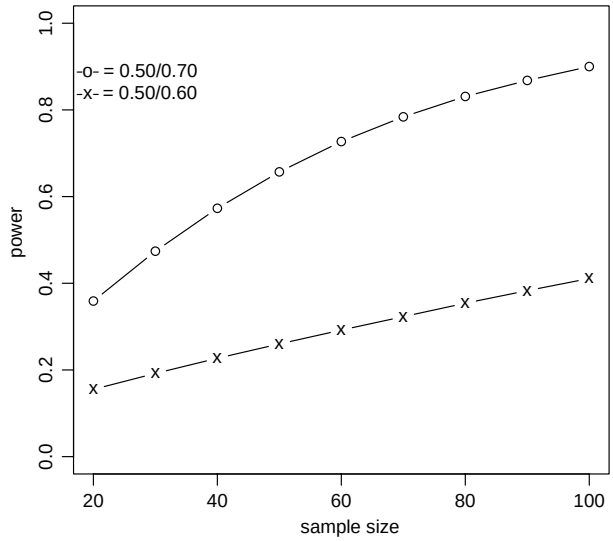

Figure 3. The power of 2 × 2 contingency tables as a function of sample size, for effects of 10% and 20%, respectively.

magnitudes of the detected effects (Kilgarriff, 2001, p. 102). Our example is based on the hypothetical data in Table VIII. In Text type 1, syntactic construction 1 occurs in 49.925% of cases, in Text type 2 the percentage is 50.090, a difference of 0.165%. The associated χ^2 is 5.436, $df = 1$, $p = 0.020$, significant at the 5% level.

A large number of indices is available (see for instance Reynolds, 1977; Liebetrau, 1983; Wickens, 1989) which aim at expressing the strength of association between the variables of contingency tables. The ideal index is:

– easy to interpret, because the possible values it can take range from 0 to 1, with 0 indicating absence of association and 1 complete association,
– independent of marginal distributions, which allows the researcher to compare effects obtained in different contingency tables,
– independent of sample size; the value of χ^2 is an extreme example of dependence on sample size,

Table VIII. Hypothetical data: the occurrence of a specific syntactical construction as a function of text type

	Text type 1	Text type 2	Total
Construction 1	500000	501800	1001800
Construction 2	501500	500000	1001500
	1001500	1001800	2003300

– has a known standard error and sampling distribution, which makes it possible to test absence of association. Obviously, this requirement is superfluous if χ^2 is calculated, and a significant association is established.

We do not know any index which fulfills all requirements mentioned above, and refer, therefore, to Liebetrau (1983) for a good overview of all pros and cons of available measures of association. For illustrative purposes we calculated a well-known measure of association, Goodman-Kruskal's λ for the data given in Table VIII. This coefficient expresses the relative decrease in the probability of an error in guessing the response (here: the occurrence of a specific syntactic construction) when the condition (here: text type) is known. The measure λ is sensitive to heterogeneity of marginal distributions, but, fortunately, the marginal distributions in our example are homogeneous. The asymmetrical version of λ calculated for our data is very low: 0.001; this means that despite the significant χ^2, knowledge of text type does hardly decrease the probability of an error in guessing the occurrence of a syntactic construction.

5. Log Odds Ratios

In the preceding sections, we showed that the analysis of frequency data on the basis of the χ^2 statistic is not always warranted. Especially the violation of the assumption of independent observations is a serious problem (cf. Section 3). Alternatives to the χ^2 statistic have been developed which lend themselves to analyses in the context of well-known techniques, like logit analyses, logistic regression, and analyses of variance (Rietveld and Van Hout, p. 1993). An important concept in this context is the log odds ratio. We illustrate its use on the basis of Tables V–VII in Section 3.3, which represent a frequently occurring situation in which several participants (speakers) are recorded, and their responses are analyzed. For convenience's sake we reproduce a general form of these 2×2 tables in Table IX.

As argued above, we need an index which expresses the extent to which the conditions (here contexts) determine the distributions of the answers (responses) over the categories for every speaker. An important requirement is that the index is independent of N, the number of observed data per unit (here: speaker). The odds ratio (also called the cross-product ratio) is such an index. For a 2×2 table, with conditional probabilities $p_{i|j}$, it is

$$\alpha = \frac{(p_{1|1})/(p_{2|1})}{(p_{1|2})/(p_{2|2})} = \frac{(p_{1|1})/(p_{2|2})}{(p_{1|2})/(p_{2|1})} \tag{5}$$

This ratio has a simple interpretation: $p_{1|1}/p_{2|1}$ is the odds ("likelihood") of observing a phenomenon of the type labeled in the first row (here: + assimilation) in the condition labeled in the first column (here

Table IX. Conditional probabilities $p_{i|j}$ in a 2×2 table

	Context 1	Context 2	Total
+Assimilation	$p_{1\|1}$	$p_{1\|2}$	$p_{1.}$
−Assimilation 1	$p_{2\|1}$	$p_{2\|2}$	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	

Context 1). The first probability (assimilation in Context 1) is compared with the second probability (no assimilation in the same context) on the basis of a ratio. For speaker A (Table V), it is $19/78 = 0.244$ versus $59/78 = 0.756$. Analogously $p_{1|2}/p_{2|2}$ is the odds of observing assimilation in context 2. The odds ratio α gives the relative value of these odds, and it is consequently a ratio of ratios. The odds ratio ranges from 0 to $+\infty$. Its value is 1 if the odds are independent of the columns (here: contexts). Note that p can be changed into n in Equation (6) without changing the value of α. In most cases the natural logarithm of α on the basis of counts n_{ij} is used (see Equation 6). The index then gets the value 0 ($= \ln 1$), if there is no association between conditions and responses. The log odds ratio has a range from $-\infty$ to $+\infty$.

$$y = \ln \frac{n_{11} n_{22}}{n_{12} n_{21}} \tag{6}$$

The odds ratio has two nice properties: (a) invariance to marginal distributions and (b) invariance under interchanges of rows and columns. When one of the frequencies is zero, often a constant of 0.5 is added to each frequency value. As a matter of fact there are number of adaptations of the log odds available to cope with the "zero problem". Gart and Zweifel (1967) showed that adding 0.5 is the preferable transformation as long as $n \times p > 1$, which will be very often the case in the applications under discussion here; for more details, see Agresti (2002, p. 397), but also Wickens (1993).

Testing the effect of condition on the responses now involves the following steps:

- Compute the log odds ratio y_j for the table of each participant j;
- Test the hypothesis that $\mu_y = 0$ with a t-test. The standard error used for this t_{k-1} test is the standard deviation of the k log odds, estimated from the sample, divided by $\sqrt{k}$.

The log odds ratio of the first subtable of our example (Table V) is:

$$y = \ln \frac{19.5 \times 40.5}{51.5 \times 59.5} = \ln \frac{789.75}{3064.25} = -1.356 \tag{7}$$

The second log odds ratio (Table VI) is 1.627. The mean of the two is 0.136 and the standard error is $s/\sqrt{k} = 2.109/1.414 = 1.492$; thus we obtain $t_1 = 0.136/1.492 = 0.091$, which is not significant at any reasonable signifi-

cance level. Apparently there is an interaction in this example between speakers and context, which affects the probability of the occurrence of assimilation; this can also be a very relevant finding.

Of course, the use of *t* tests in this context has possible drawbacks. The first one is the relatively small power of the test when just a small number of speakers is involved, as the degrees of freedom is equal to the number of subtables minus 1. We think, however, that this reduced power is fully compensated by the realistic *p* values we obtain, compared with the situation in which the assumptions of χ^2 are not fulfilled. A second possible drawback consists in worries about the use of *t* tests when the normality of the population from which the samples are drawn is not warranted. However, as early as in 1960 Boneau showed on the basis of simulation studies that the *t* test is quite robust against this possible violation. Wickens' (1993) simulation studies showed that both power and type I errors of the proposed test remain quasi unaffected by the presence of asymmetry in the marginal distributions of the contingency tables.

6. Conclusion

The aim of this paper was to make researchers more aware of possible pitfalls associated with the analysis of corpus data. We started with a demonstration of problems and biases connected with the transcription of spoken speech data. One fundamental problem is disagreement between observers. We presented some indices of between-observer agreement, and we suggested a number of steps to follow in case disagreements between observers occur. An important conclusion was that the mere deletion of data which observers disagree on is not a self-evident solution at all. It is only acceptable if the disagreements are uniformly distributed over the research conditions. We demonstrated how this distribution can be tested.

As the analysis of corpus data very often involves the analysis of frequencies of occurrence, we extensively discussed a crucial assumption on which χ^2 and equivalent statistics are based, viz. the independence of observations. This assumption is often not fulfilled in real data sets. A simulation experiment showed the dramatic consequences of not meeting the assumption of non-dependences.

We made a distinction between sequential dependence and unit dependence. The first type can be dealt with by applying the window method suggested by Altham (1979). This approach takes into account the assumed or tested number of dependent observations in a sequence. Unit dependence concerns the matching of the level of sampling and the level of analysis. The two levels have to be the same. Our computer simulations showed that

neglecting this assumption, which is fairly common in corpus research, has dramatic negative consequences for the statistical validity of the results.

We discussed the notion of power in the context of small data sets, and, in relation to large data sets, we mentioned the usefulness of indices which express the size of an effect independently of the sample size. Finally, we pointed out the important role of the log odds ratio in frequency analysis. They often provide a good alternative to χ^2 analysis. We explained how a t test can be used to test whether log odds differ from zero.

Acknowledgements

We thank the reviewers for many valuable suggestions for improvement of the manuscript.

Notes

[1] According to Liebetrau (1983), κ divided by its standard error is approximately a normal variable "for n sufficiently large", n being the number of objects to be judged. We suggest to follow the recommendations of Siegel and Castellan (1988) made for the use of χ^2, as the square root of the latter corresponds with the z value associated with κ, (a) when $n \leq 20$, use an exact test, (b) when n is between 20 and 40, κ divided by its standard error can be used, as long as all expected values are 5 or more. Another suggestion, given in Wickens (1989, p. 240) for κ, corresponds with these recommendations: the sample size for an $a \times a$ Table should be at least $16a^2$. For our example this amounts to $16 \times 4 = 64$, 2 more than the actual sample size. Fortunately, exact tests for κ are available in statistical packages. For Table I, $\kappa = 0.449$, the "approximate significance", obtained with SPSS, is 0.000, whereas the "exact significance" is 0.001.

[2] Moreover, as shown by Schouten (1985), the interpretation of κ is seriously hindered by unequal marginal totals.

References

Agresti A. (2002) *Categorical Data Analysis*. Wiley and Sons, Hoboken, NJ.

Altham P. (1979) Detecting Relationships between Categorical Variables Observed over Time: A Problem of Deflating a Chi-Quared Statistic. *Applied Statistics*, 28, pp. 115–125.

Boneau C. (1960) The Effects of Violations of Assumptions Underlying the t Test. *Psychological Bulletin*, 57, pp. 49–64.

Carletta J. (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22, pp. 250–254.

Cohen J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, pp. 37–46.

Cucchiarini C. (1993) *Phonetic Transcription: A Methodological and Emperical Study*. Unpublished PhD-thesis, University of Nijmegen.

Cucchiarini C. (1996) Assessing Transcription Agreement: Methodological Aspects. *Clinical Linguistics & Phonetics*, 10, pp. 131–155.

Dainora D. (2002) Does Intonational Meaning Come From Tones or Tunes? Evidence Against a Compositional Approach. In *Proceedings of the 1st International Conference on Speech Prosody*, pp. 235–238.

Delucchi K. (1983) The Use and Misuse of Chi-Square: Lewis and Burke Revisited. *Psychological Bulletin*, 94, pp. 166–176.

Dunning T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19, pp. 61–74.

Ernestus M. (2000) *Voice Assimilation and Segment Reduction in Casual Dutch, a Corpus-Based Study of the Phonology – Phonetics Interface*. LOT Utrecht.

Fienberg S.E. (1980) *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, MA.

Fleiss J. (1971) Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76, pp. 378–382.

Gart J., Zweifel J. (1967) On the Bias of Various Estimators of the Logit and its Variance with Application to Quantal Bioassay. *Biometrika*, 54, pp. 275–281.

Keating P. (1998), Word-Level Phonetic Variation in Large Speech Corpora. In Alexiadou, H., Fuhrhop U., Kleinhenz U., Law P. (eds.), *Papers of the Conference "The Word as a Phonetic Unit"* [*ZAS Papers in Linguistics* 11]. Zentrum für Allgemeine Sprachwissenschaft, Sprachtypologie und Universalienforschung, Berlin, pp. 35–50.

Kemps R., Ernestus M., Schreuder R., Baayen R. H. (2004) Processing Reduced Word Forms: The Suffix Restoration Effect. *Brain and Language*, 90, pp. 117–127.

Kilgarriff, A. (2001) Comparing Corpora. *International Journal of Corpus Linguistics*, 6, pp. 97–133.

Kruskal J. (1983) An Overview of Sequence Comparison. In Kruskal J., Sankoff D. (eds.), *Time Wraps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.

Kuijpers C., Van Donselaar M. (1997) The Influence of Rhythmic Context on Schwa Epenthesis and Schwa Deletion in Dutch. *Language and Speech*, 41, pp. 87–108.

Lewis D., Burke C. J. (1949) The Use and Misuse of the Chi-Square Test. *Psychological Bulletin*, 46, pp. 433–489.

Liebetrau A. (1983) *Measures of Association*. Sage Publications, London.

Reynolds H. (1977) *The Analysis of Cross-Classifications*. The Free Press, New York.

Rietveld T., Van Hout R. (1993) *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin.

Schouten H. (1985) *Statistical Measurement of Interobserver Agreement*. University of Rotterdam, Rotterdam.

Siegel S., Castellan N.J. (1988) *Nonparametric Statistics*. McGraw Hill, New York.

Van de Velde H., Van Hout R. (2001) The Devoicing of Fricatives in a Reading Task. In Van der Wouden, T. Broekhuis H. (eds.), *Linguistics in the Netherlands 2001*. John Benjamins, Amsterdam, pp. 219–229.

Vieregge W. (1987) Basic Aspects of Phonetic Segmental Transcription. In Almeida H., Braun A. (eds.): *Probleme der Phonetischen Transkription*. Franz Steiner Verslag Wiesbaden GMBH, Stuttgart.

Vieregge W., Maassen, B. (1999) IPA Transcriptions of Consonants and Vowels Spoken by Dyspractic Children: Agreement and Validity. In Maassen B., Groenen P. (eds.), *Pathologies of Speech and Language*. Whurr, London, pp. 275–284.

Wickens T. (1989) *Multiway Contingency Tables Analysis for the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wickens T. (1993) Analysis of Contingency Tables With Between-Subjects Variability. *Psychological Bulletin*, 113, pp. 191–204.