Extracting the Lowest-Frequency Words: Pitfalls and Possibilities

Marc Weeber*
University of Groningen

Rein Vos†
University of Groningen, University of Maastricht

R. Harald Baayen‡
Max Planck Institute for Psycholinguistics

In a medical information extraction system, we use common word association techniques to extract side-effect-related terms. Many of these terms have a frequency of less than five. Standard word-association-based applications disregard the lowest-frequency words, and hence disregard useful information. We therefore devised an extraction system for the full word frequency range. This system computes the significance of association by the log-likelihood ratio and Fisher's exact test. The output of the system shows a recurrent, corpus-independent pattern in both recall and the number of significant words. We will explain these patterns by the statistical behavior of the lowest-frequency words. We used Dutch verb-particle combinations as a second and independent collocation extraction application to illustrate the generality of the observed phenomena. We will conclude that a) word-association-based extraction systems can be enhanced by also considering the lowest-frequency words, b) significance levels should not be fixed but adjusted for the optimal window size, c) hapax legomena, words occurring only once, should be disregarded a priori in the statistical analysis, and d) the distribution of the targets to extract should be considered in combination with the extraction method.

1. Introduction

The research reported here arose from an attempt to determine the conditions under which optimal recall and precision are obtained for the extraction of terms related to side effects of drugs in medical abstracts. We used the standard technique of defining a window around a seed term, *side-effect* in our case, and selected as potentially relevant terms those words that appeared more often in these windows than expected under chance conditions.

Our original question concerned the extent to which recall and precision are influenced by the size of the window. It turns out, however, that a preliminary question needs to be answered first, namely, how to gauge the significance of the large effect of the lowest-frequency words on recall, precision, and the number of words extracted as potentially relevant terms.

* Groningen University Institute for Drug Exploration, Department of Social Pharmacy and Pharmacoepidemiology, Ant. Deusinglaan 1, 9713 AV Groningen, The Netherlands. E-mail: marc@farm.rug.nl
† Faculty of Health Sciences, Department of Health Ethics and Philosophy, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: rein.vos@zw.unimaas.nl
‡ Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen. E-mail: baayen@mpi.nl

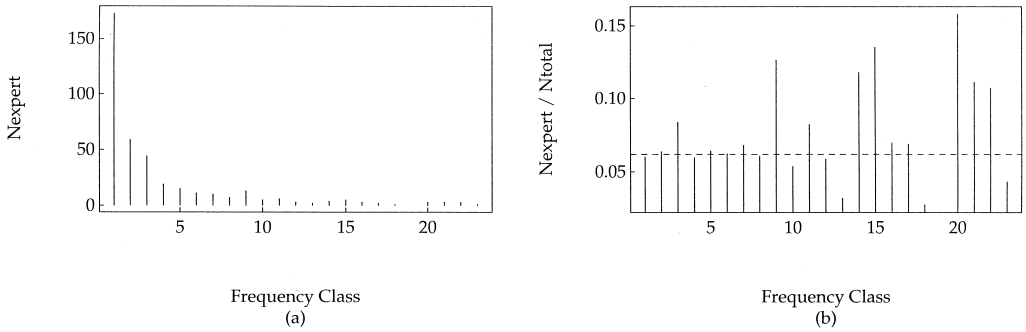

Figure 1
Frequency distribution of medical expert word types. Panel (a) shows the number of side-effect-related word types as judged by a medical expert (Nexpert) as a function of the first 23 frequency classes. Panel (b) shows the proportion of expert types/total corpus types (Ntotal) for the first 23 frequency classes. The horizontal dashed line indicates the mean proportion of 0.0619.

It is common practice in information retrieval to discard the lowest-frequency words a priori as nonsignificant (Rijsbergen 1979). In Smadja's collocation algorithm Xtract, the lowest-frequency words are effectively discarded as well (Smadja 1993). Church and Hanks (1990) use mutual information to identify collocations, a method they claim is reasonably effective for words with a frequency of not less than five.

A frequency threshold of five seems quite low. Unfortunately, even this lower frequency threshold of five is too high for the extraction of side-effect-related terms from our medical abstracts. To see this, consider the left panel of Figure 1, which plots the number of side-effect-related words in our corpus of abstracts as judged by a medical expert, as a function of word-frequency class. The side-effect-related words with a frequency of less than five account for 295 of a total of 432 expert words (68.3%). The right panel of Figure 1 shows that the first 23 word-frequency classes are characterized by, on average, the same proportion of side-effect-related words. The a priori assumption of Rijsbergen (1979) that the lowest-frequency words are nonsignificant is not warranted for our data, and, we suspect not for many other data sets as well.

The recent literature has seen some discussion of the appropriate statistical methods for analyzing the contingency tables that contain the counts of how a word is distributed inside and outside the windows around a seed term. Dunning (1993) has called attention to the log-likelihood ratio, G^2, as appropriate for the analysis of such contingency tables, especially when such contingency tables concern very low frequency words. Pedersen (1996) and Pedersen, Kayaalp, and Bruce (1996) follow up Dunning's suggestion that Fisher's exact test might be even more appropriate for such contingency tables.

We have therefore investigated for the full range of word frequencies whether there is an optimal window size with respect to recall and the number of significant words extracted using both the log-likelihood ratio and Fisher's exact test. In Section 2, we will show that indeed there seems to be an optimal window size for both statistical tests. However, a recurrent pattern of local optima calls this conclusion into question. Upon closer inspection, this recurrent pattern appears at fixed ratios of the number of words inside the window to the number of words outside the window (complement).

In Section 3, we will relate the recurrent patterns of local optima at fixed window-complement ratios (henceforth W/C-ratios) to the distributions of the lowest-frequency words over window and complement. We will call attention to the critical effect of the choice of W/C-ratios on the significance of the lowest-frequency words.

As the improvement in the extraction of side-effect terms from medical abstracts, as gauged by the F-measure, which combines recall and precision (Rijsbergen 1979), is small, we also applied the same approach to the extraction of Dutch verb-particle combinations from a newspaper corpus. In Section 4, we report substantially better results for this more lexical extraction task, which is subject to the same statistical behavior of the lowest-frequency words.

In the last section, we will discuss the consequences of our findings for the optimization of word-based extraction systems and collocation research with respect to the lowest-frequency words.

2. An Optimal Window Size for Medical Abstracts?

The MEDLINE bibliographic database contains a large number of abstracts of scientific journal papers discussing medical and drug-related research. Typically, abstracts discussing medical drugs mention the side effects of these drugs briefly. Information on side effects is potentially relevant for finding new applications for existing drugs (Rikken and Vos 1995). We are therefore interested in any terms related to the side effects of drugs.

Before proceeding, it may be useful to clarify the way in which the present research differs from standard research on collocations. In the latter kind of research, there is no a priori knowledge of which combinations of words are true collocations. Moreover, the most salient collocations generally are found at the top of a list ranked according to measures for surprise or association, such as G^2 or mutual information (Manning and Schütze 1999). The large numbers of word combinations with significant but low values for these measures are often of less interest. Low-frequency words are predominant among these kinds of collocations. In our research, we likewise find many low-frequency terms for side effects with low ranks in medical abstracts. The relatively well-known side effects that are mentioned frequently can be captured by examining the top ranks in the lists of extracted words. At the same time, the rarely mentioned side-effect terms are no less important, and in post marketing surveillance the extraction of such side-effect terms may be crucial for the acceptance or rejection of new medicines.

Is reliable automatic extraction of both low- and high-frequency side-effect terms from MEDLINE abstracts feasible? To answer this question, we explored the efficacy of a standard collocation-based term extraction method that extracts those words that appear more frequently in the immediate neighborhood of a given seed term than might be expected under chance conditions.

We compiled two corpora on the side effects of the cardiovascular drugs captopril and enalapril from MEDLINE abstracts. The first corpus contains all abstracts mentioning captopril and the word *side*. The second corpus contains all abstracts mentioning captopril and at least one of the compounds *side-effect*, *side effect*, *side-effects*, and *side effects*. Thus, the second corpus is a subset of the first. The first corpus is comprised of 118,675 tokens and 7,678 types; the second corpus 103,603 tokens and 6,582 types. A medical expert marked 432 of the latter word types as side-effect-related terms. The left panel of Figure 1 summarizes the head of the frequency distribution of these terms in the larger corpus. Note that most side-effect-related terms have a frequency lower

Table 1
General 2×2 contingency table. A = frequency of the target in the window corpus, B = frequency of the target in the complement corpus, W = total number of words in the window, C = total number of words in the complement. Corpus size $N = W + C$.

	window	complement	
frequency of target	A	B	$A + B$
sum frequency of other words	$W - A$	$C - B$	$W + C - A - B$
	W	C	$W + C$

than five. What we need, then, is an extraction method that is sensitive enough to select such very low frequency terms.

In the collocation-based method studied here, the neighborhood of a given seed term is defined in terms of a window around the seed term. We constructed windows around all seed terms in the corpus, leading to a window corpus and a complement corpus. The window corpus contains all words that appear within a given window size of the seed term. For instance, with a window size of 10, any word appearing from five words before the seed to five words after the seed as well as the seed itself is included in the window corpus. The word tokens not in the window corpus comprise the complement corpus. Any type in the window corpus is a potential side-effect-related term. For any such target type, we tabulate its distribution in window and complement corpora in a contingency table like Table 1.

Given W and C, we need to know whether the frequency of the target in the window corpus, A, is high enough to warrant extraction. Typically, given the marginal distribution of the contingency table, a target is extracted for which $\frac{A}{W-A} > \frac{B}{C-B}$, and for which the tabulated distribution is nonhomogeneous according to tests such as G^2 and Fisher's exact test for a given α-level.

In this approach, the window size is a crucial variable. At small window sizes, many potentially relevant terms fail to appear in the window corpus. However, at large window sizes, many irrelevant words are found in the window corpus and may be extracted spuriously.

To see to what extent window size may affect the results of the extraction procedure, consider the solid lines in panels (a) and (b) of Figure 2. The left panel shows the results for recall when we use the log-likelihood ratio, G^2, the right panel the results for Fisher's exact test. We define recall as the proportion of the number of side-effect words extracted and the total number of side-effect words available in the window.

For both statistical tests, recall seems to be optimal at window size 2. However, at this window size, the number of words extracted is very small. This can be seen in panels (c) and (d). Considered jointly, panels (a) and (c) suggest an optimal window size of 24 for our larger corpus (corpus 1), as recall is still high, and the number of significant words is maximal. When Fisher's exact test is used instead of G^2, panels (b) and (d) suggest 42 as the optimal size.

The dashed lines in panels (a) to (d) show the corresponding results for our smaller corpus (corpus 2). Unsurprisingly, the general pattern for this subcorpus is quite similar, although the drops in recall and the number of significant words, Nsig, occur at somewhat smaller window sizes.

Interestingly, we can synchronize the curves for both corpora by plotting recall and the number of significant items, Nsig, against the window-complement ratio (W/C). This is shown in panels (e) and (f). These panels suggest not an optimal window size

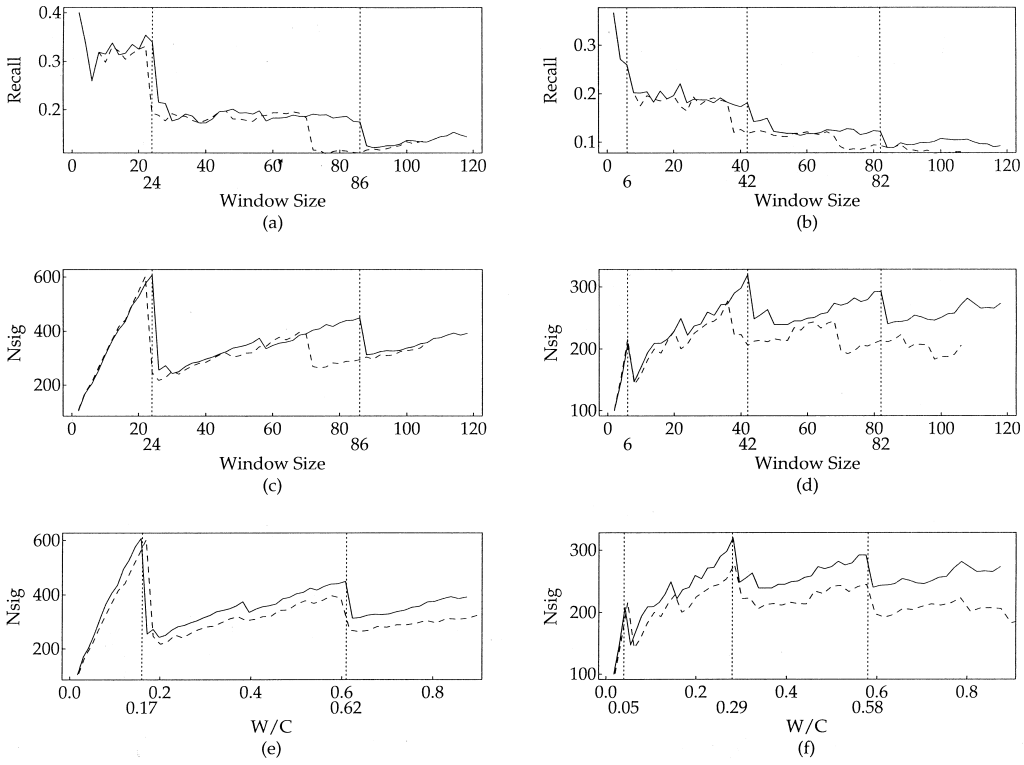

Figure 2
Results of the word extraction procedure ($\alpha = 0.05$). Solid line = corpus 1, dashed line = corpus 2. Panel (a) shows the log-likelihood, G^2, recall results as a function of the window size. Panel (b) shows recall values for Fisher's exact test. Panel (c) shows the total number of significant words (Nsig) as a function of the window size for G^2. Panel (d) shows the same as (c) but for Fisher's exact test. Panel (e), G^2, and (f), Fisher's exact test, also show the total number of significant words, but as a function of the W/C-ratio; the ratio of the number of words in the window corpus to the number of words in the complement corpus.

but an optimal W/C-ratio (0.17 for G^2 and 0.29 for Fisher's exact test). Although we now seem to have shown that recall and Nsig depend on the choice of window size, the sudden drops in recall and Nsig and the reoccurrence of such drops at various W/C-ratios is a source of worry, not only for G^2 results, but also for the results based on Fisher's exact test. A further source of worry is the fact that the two tests diverge considerably with respect to the optimal W/C-ratio.

3. Contingency Tables and the Lowest-Frequency Words

Before we can have any confidence in the optimality of a given W/C-ratio, we should understand why the saw-tooth-shaped patterns of Nsig arise. Both the log-likelihood ratio (G^2) and Fisher's exact test compute the significance of contingency tables similar to Table 1. So why is it that the left panels in Figure 2 differ from the right panels? G^2 has a χ^2-distribution as $N \rightarrow \infty$. This convergence is not guaranteed for low expected frequencies and sparse tables, which renders use of G^2 problematic for our lowest-frequency words in that it may suggest words to be more remarkable than they

Table 2
Contingency tables for hapax legomena, dis legomena, and tris legomena. W = number of words in window corpus; C = number of words in complement corpus. Total corpus size: $N = W + C$.

(a):	1	0	(b):	2	0	(c):	1	1
	$W-1$	C		$W-2$	C		$W-1$	$C-1$
(d):	3	0	(e):	2	1	(f):	1	2
	$W-3$	C		$W-2$	$C-1$		$W-1$	$C-2$

really are. Fisher's exact test, on the other hand, does not use an approximation to a probability distribution but computes the exact hypergeometric distribution given the marginal totals of the contingency table. While Fisher's exact test is suitable for the analysis of sparse tables, it is inherently conservative because it regards the marginal totals not as stochastic variables but as fixed boundary conditions. Consequently, this test is likely to reject words that are in fact remarkably distributed in the contingency table. The difference in behavior of the two tests is clearly visible in panels (c) and (d) of Figure 2: the number of significant words (Nsig) according to G^2 is roughly twice as large as that according to Fisher's exact test.

When a hapax legomenon[1], a word with frequency 1, occurs in the window corpus, we use contingency table (a) as shown in Table 2. For dis legomena, words with a frequency of 2, that appear at least once in the window corpus, we obtain the two contingency tables (b) and (c). The interesting contingency tables for tris legomena are tables (d) to (f). These six tables are relevant for 63.8% of the side-effect-related terms as judged by our medical expert.

How do changes in the W/C-ratio affect G^2 and Fisher's exact test, when applied to contingency tables (a) to (f)? In other words, how does the choice of the window size affect whether a low-frequency word is judged to be a significant term, for fixed A and B (e.g., $A = 1$ and $B = 0$ for a hapax legomenon)?

First, consider contingency tables with $B = 0$, for instance tables (a), (b), and (d). For small A, ($A \ll W, C$), it is easily seen (see the appendix) that the critical W/C-ratio based on the log-likelihood ratio is:

$$\frac{W}{C} = \frac{1}{\sqrt[A]{e^{X/2}} - 1}, \tag{1}$$

with X the χ^2 value corresponding to a given α-level with 1 degree of freedom. For $A = 1$ and $\alpha = 0.05$, $X = 3.84$, the critical W/C-ratio equals 0.1718. This is exactly the W/C-ratio in panel (e) in Figure 2 at which the first and largest drop in the number of significant words occurs. Up to this ratio, any hapax legomenon appearing in the window corpus is judged to be a significant term. For $W/C > 0.1718$, no hapax legomenon will be extracted.

Fisher's exact test is far more conservative. For this test, the critical W/C-ratio is

1 The term hapax legomenon (literally 'read once') goes back to classical studies and was originally used to refer to the words used once only in the works of a given author, e.g., Homer. By analogy, dis legomenon and tris legomenon have come into use to refer to words occurring only twice or three times.

Table 3
Critical W/C-ratios where sparse and skewed contingency tables lose significance. Equations 1 and 2 provide the ratios for the $B = 0$ cases. The other ratios are obtained by simulations.

	distribution	G^2		Fisher	
	A–B	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
hapax legomena	1 – 0	0.1718	0.0375	0.0526	0.0101
dis legomena	1 – 1	0.0400	0.0092	0.0260	0.0050
	2 – 0	0.6204	0.2348	0.2880	0.1111
tris legomena	1 – 2	0.0232	0.0053	0.0172	0.0033
	2 – 1	0.1917	0.0824	0.1565	0.0626
	3 – 0	1.1155	0.4938	0.5833	0.2746

(see the appendix for details):

$$\frac{W}{C} = \frac{\sqrt[A]{P}}{1 - \sqrt[A]{P}}, \tag{2}$$

where P is the α-level. For $A = 1$ and $P = 0.05$, the critical W/C-ratio for a hapax legomenon equals 0.0526. In panel (f) of Figure 2, we observe the first drop in the number of significant words at precisely this W/C-ratio. For very small W/C-ratios, *any* hapax legomenon in the window corpus is also judged to be significant according to Fisher's exact test. Compared to G^2, Fisher's exact test rejects hapax legomena as significant at much smaller W/C-ratios. Note that when $W/C = 0.05/0.95 = 0.0526$, i.e., when the window corpus is exactly 1/20 of the total corpus, the probability that a hapax legomenon appears in the window corpus equals 0.05. Our conclusion is that, with the W/C-ratio as the only determinant of significance, the windowing method is not powerful enough to distinguish between relevant and irrelevant hapax legomena. In other words, hapax legomena should be removed from consideration a priori.

For dis legomena that appear exclusively in the window corpus, the critical ratios are 0.6204 for G^2, corresponding to the second major drop in panel (e) of Figure 2, and 0.2880 for Fisher's exact test, corresponding to the severe drop following the maximum of Nsig in panel (f). The third major drop in this panel corresponds to the critical W/C-ratio for tris legomena occurring three times in the window corpus.

For contingency tables with $B > 0$; $A > B$; $A, B \ll W, C$, critical W/C-ratios are not easy to capture analytically. We therefore carried out a simulation study for $W + C =$ 100,000. For fixed A and B and a given α-level, we calculated the critical W/C-ratio by iterative approximation. Results are summarized in Table 3.

When we highlight these critical ratios in Figure 2 by means of vertical dashed lines, we obtain Figure 3. Panels (a) to (d) correspond to the curves for corpus 2 in the first four panels of Figure 2. For the log-likelihood ratio, we observe that both the major and minor drops in recall and the number of significant words (Nsig) occur at the W/C-ratios where different distributions of the lowest-frequency words lose significance. For Fisher's exact test, we observe exactly the same pattern. Panels (e) and (f) show the number of significant words for a pseudorandomized version of corpus 2 where we used the same tokens but randomized the order of their appearance. Although the number of significant words is lower, the saw-tooth-shaped pattern with the sudden drops at fixed ratios reemerges.

We conclude that W and C are the prime determinants of both recall and the number of significant words. At first sight, Fisher's test is clearly preferable to the

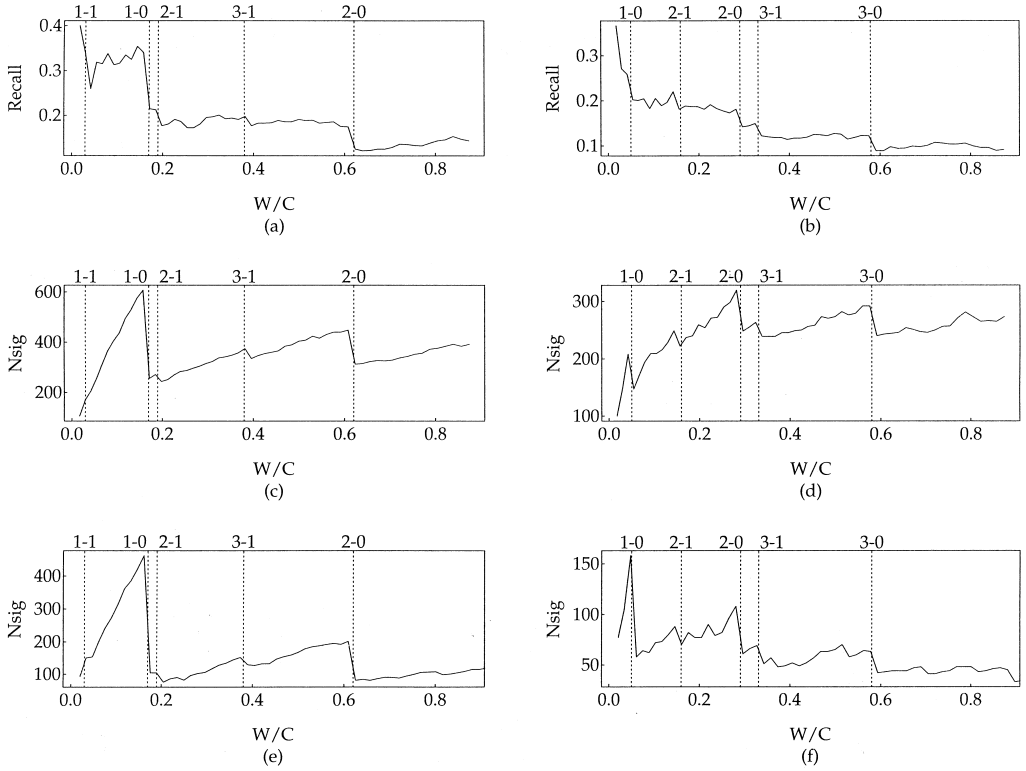

Figure 3
Results of word extraction procedure ($\alpha = 0.05$) with A–B distributions. Panels (a), log-likelihood ratio, G^2, and (b), Fisher's exact test, show the recall results of the extraction procedure for corpus 2. Panels (c) and (d) show the total number of significant words (Nsig), again for G^2 and Fisher's exact test, respectively (see also Figure 2). Panels (e) and (f) show the results for a randomized corpus for G^2 and Fisher's exact test. The numbers above the panels indicate the A–B distribution of the contingency tables in Table 2.

log-likelihood ratio because the extreme saw-tooth-shaped pattern is substantially reduced. However, the use of Fisher's exact test does not eliminate the effect of the choice of window and complement size on the number of significant words and recall. At specific W/C-ratios, nonnegligible numbers of words with the lowest frequency of occurrence suddenly lose significance. Moreover, in our discussion thus far, we have not taken extraction precision into account nor the trade-off between precision and recall. For the assessment of overall extraction results, we turn to the F-measure (Rijsbergen 1979), a measure that assigns equal weights to precision (P) and recall (R):

$$F = \frac{2PR}{P + R}. \tag{3}$$

Figure 4 plots precision, recall, and F as a function of the W/C-ratio. The common trade-off between recall and precision is clearly present for the smaller window sizes, with the F-measure providing a kind of average.

Thus far, we have applied a common collocation extraction technique to a semantic association task. Actual extraction performance is low: F is maximally 0.17. To gauge

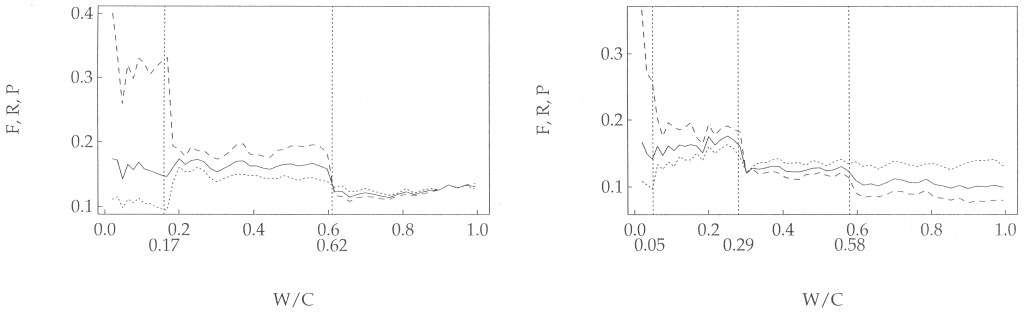

Figure 4
F, recall, and precision as a function of the W/C-ratio. Recall (R, dashed line), F (solid line), and precision (P, dotted line) using G^2 (left panel) and Fisher's exact test (right panel) for our second corpus plotted as a function of the W/C-ratio.

whether better results can be obtained with the present techniques, we examined the extraction of Dutch verb-particle combinations.

4. Extracting Verb-Particle Combinations

In English, the particle of verb-particle combinations always follows the verb, as in *she rang him up*. In Dutch, the particle can occur either before or after the verb. When it occurs before the verb, it is separated from the verb by *te* ('to') and/or one or more auxiliary verbs. Extracting such particle-verb combinations is relatively straightforward. However, when the particle follows the verb, it may be separated from the verb by many constituents of arbitrary complexity: *Hij zegt de belangrijke afspraak met de programmeur voor vanmiddag af* ('he says the important meeting with the programmer for this afternoon off'; i.e., he cancels the meeting). How well does our present approach lend itself to the extraction of verb-particle combinations with the particle *af* ('off') when the particle follows the verb?

We investigated this question by studying verb-particle combinations with *af* from a Dutch newspaper corpus of about 4.5 million word tokens. We extracted by hand all sentences from the corpus that contain *af* (3,802 sentences, 97,903 tokens) and singled out those sentences in which *af* belongs to a verb-particle combination in which the verb occurs to the left of the particle (2,202 sentences with 42,825 tokens). The targets to extract from the 2,202 sentences are 436 different verb inflections, of which 276 have a frequency of less than five. Just as the judgments of a medical expert were used in the preceding extraction task to provide a frame of reference for the evaluation of precision and recall, the present lexical extraction task has as its frame of reference the 2,202 sentences that we judged to contain a verb followed at some point to the right by a particle. How many of the 436 different verb inflections can we extract with our windowing technique, and what is the trade-off between recall and precision?

To answer this question, we defined windows to the left of the seed term *af* in the range of positions $[-12, -1]$. We calculated the W/C-ratio for each window size. For each word in all windows, we calculated its significance according to G^2 and Fisher's exact test. Using the 436 target verb inflections as a frame of reference, we computed precision, recall, and F. Panel (a) of Figure 5 plots F as a function of the W/C-ratio. F reaches a maximum F of 0.31 at $W/C = 0.59$ for G^2 (the solid line in the figure) and a maximum of 0.27 at $W/C = 0.50$ for Fisher's exact test (the dashed line). These

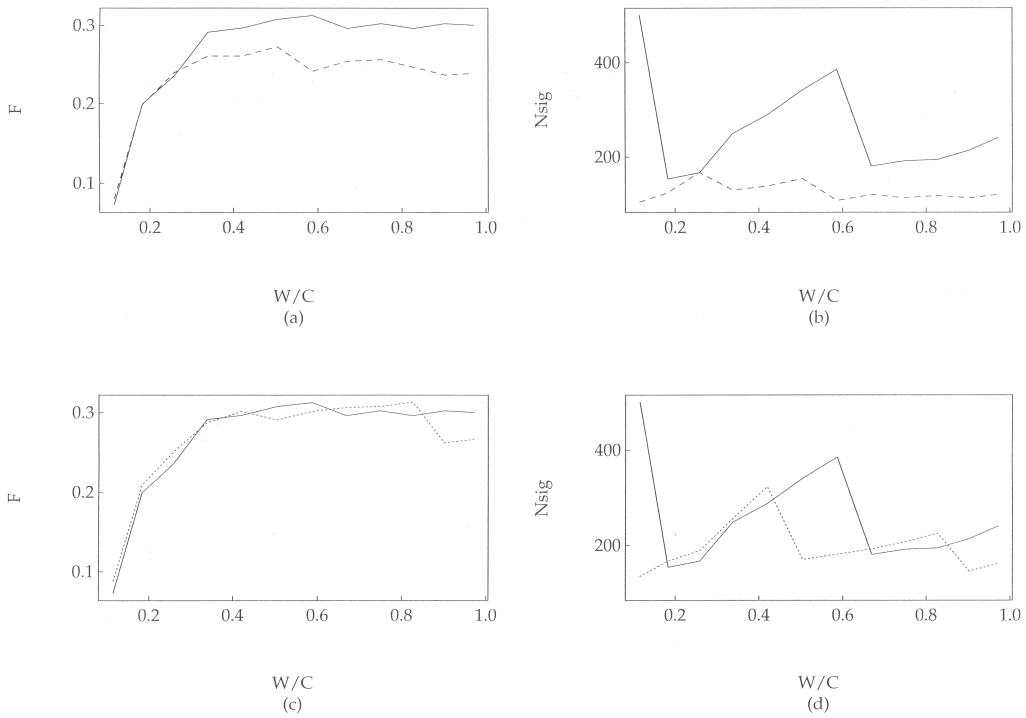

Figure 5
Extraction results for the *af* corpus. Panel (a) shows F for G^2 (solid line) and Fisher's exact test (dashed line) as a function of the W/C-ratio. Panel (b) displays the number of significant words (Nsig) according to both tests. Panel (c) shows F for G^2 at $\alpha = 0.05$ (solid line) and Fisher's exact test at $\alpha = 0.1$ (dotted line). Panel (d) shows Nsig for G^2 at $\alpha = 0.05$ and for Fisher's exact test at $\alpha = 0.1$.

results compare favorably with the maximum F of 0.17 obtained for the extraction of side-effect terms from medical abstracts.

Panel (b) of Figure 5 shows the by-this-time familiar saw-tooth-shaped pattern of the number of significant word types as function of the W/C-ratio. We observe again that Fisher's exact test is more conservative, and in the extraction task, less successful, than G^2. However, by opting for a more liberal α-level we can compensate for the conservatism of Fisher's exact test and obtain an F profile that is indistinguishable from that of G^2 as shown in panel (c) for $\alpha = 0.1$. Panel (d) returns to the number of significant terms (Nsig) when Fisher's exact test is used with $\alpha = 0.1$. Note that the optimal W/C-ratio according to F for G^2 (0.59) still leads to a higher Nsig than the optimal W/C-ratio (0.83) for Fisher's exact test with $\alpha = 0.1$. However, in the case of Fisher's exact test, the precision is much higher than when G^2 is used. These results suggest that the choice of G^2 or Fisher's exact test should be guided by the desired trade-off between precision and recall.

5. Discussion

The question that originally motivated the present research concerned the determination of the optimal window size for the extraction of side-effect-related words. Most

words that are judged by a medical expert to be related to side effects have frequencies of use that are so low that they fall below the frequency thresholds generally used in standard information extraction techniques. Is it nevertheless possible to single out such low-frequency terms through optimal window size estimation, especially since the log-likelihood ratio and Fisher's exact test have recently been advanced as suitable techniques even for the analysis of the lowest-frequency ranges?

Manipulation of the window size revealed a saw-tooth-shaped pattern in the number of significant words (Nsig) that depends not on the window size itself but on the W/C-ratio. This saw-tooth-shaped pattern arises most prominently when the log-likelihood ratio is used to extract significant words, but it is also clearly visible when Fisher's exact test is used. This pattern is due to the way in which these tests evaluate surprise as a function of the window size for the lowest-frequency words. We argue that hapax legomena should be disregarded a priori, while for low-frequency words with frequency greater than 1, only the most extreme distributions over window and complement are reliable in that we are confident that these terms are really related to the seed. For dis and tris legomena, for instance, all occurrences should in effect be concentrated in the window. Only then are we confident that there is truly a relationship between the seed and the target.

With these restrictions, the optimum W/C-ratio for our side-effect data is just smaller than 0.2880, using Fisher's exact test, which amounts to an optimal window size of 36. Of the 295 terms with a frequency of 4 or less that a medical expert judged to be side-effect-related terms, we capture 14, which amounts to 4.8%. When we exclude the hapax legomena as impossible to extract reliably a priori, we capture $14/122 = 11.5\%$. Although the gain in number of significant low-frequency items is small, the success for the low-frequency items is still reasonable when compared to the corresponding success rate of $26/137 = 19.0\%$ for the items with a frequency of 5 or more. These results suggest that the windowing technique is far from optimal for the extraction of side-effect terms from medical abstracts, irrespective of the frequencies of these terms.

The windowing technique applied to the extraction of Dutch verb-particle combinations led to more encouraging results. Choosing 0.4625 as the optimal W/C-ratio for the *af* data, which amounts to accepting dis legomena with a 2–0 distribution, and using $\alpha = 0.1$ with Fisher's exact test, we obtain an optimal window size of 5. With this window, we extract 42 of the 139 lowest-frequency words in the 2 to 4 range, i.e., 30.2%. This compares favorably to the success rate of $60/170 = 35.2\%$ for verbs with a frequency greater than 4. When we use G^2 instead of Fisher's exact test to obtain improved recall at the cost of lesser precision, we extract $58/139 = 41.7\%$ of the lowest-frequency words in the 2 to 4 range and $64/170 = 37.6\%$ of the higher-frequency words (optimum W/C-ratio 0.6204, corresponding window size of 7). For this more lexical extraction task, extraction success rates are comparable for the lower-frequency and the higher-frequency words. Neglecting the extraction of the lower-frequency words a priori would have led to the loss of nearly half of the words currently extracted.

The difference in the results between the two extraction tasks, side effects in medical abstracts and verb-*af* combinations in a newspaper corpus, is due to the difference in the distributions of the targets around the seed terms. Concentrating on the lowest-frequency word tokens, the left panel of Figure 6 shows their distribution for the side-effect corpus. The right panel shows the corresponding distribution for the *af* corpus. The side-effect terms reveal a wide scatter around the seed at position 0. By contrast, verbs predominantly cluster close to the left of *af*. Apparently, the distance between the verb and the particle is more constrained than the distance between side-effect terms and the seed term. The optimal window size of 7 (position -7) for G^2

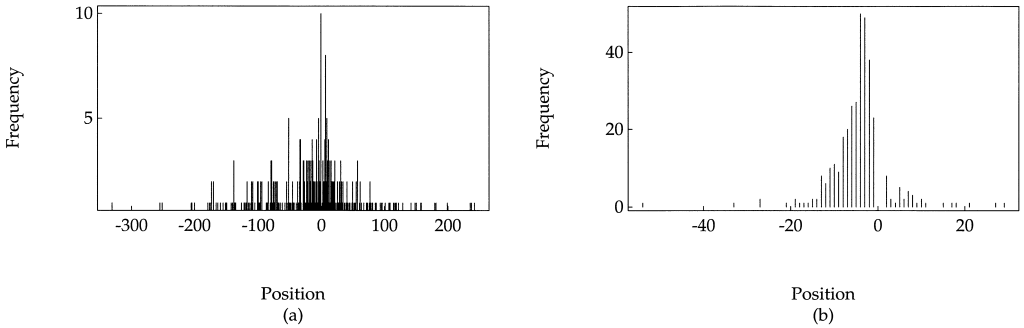

Figure 6
Frequency distribution of words occurring two to four times. Panel (a) shows for the side effect corpus how the expert words with a frequency of 2, 3, and 4 are distributed around the seed term. Panel (b) shows this distribution for the *af* corpus.

obtained above ties in with the distribution of the lowest-frequency words: 68% of all lowest-frequency tokens are in this window. For the side-effect corpus, only 31% of all low-frequency tokens are in the optimal window of 36 for Fisher's exact test. This suggests that the optimal window size must be ascertained on the basis of the distribution of targets around the seed, on the one hand, and by optimizing the statistics, on the other hand.

As an illustration of how the statistics can be optimized, we return to the *af* data. When we look at the distribution of the lowest-frequency words in Figure 6, an optimal window size of 8 to the left suggests itself. This translates into a W/C-ratio of 0.6689. Given that we want to retain dis legomena with a 2–0 distribution, we proceed to compute the corresponding significance levels for both G^2 and Fisher's exact test by Equations 1 and 2. The critical χ^2 value for G^2 equals 3.65, the critical P for Fisher's exact test is 0.161. The extraction results for both tests as measured by F are 0.31 and 0.33, respectively. This procedure allows us to extract $64/139 = 46.0\%$ of the low-frequency words and $66/170 = 38.8\%$ of the high-frequency words using G^2, and $64/139 = 46.0\%$ and $79/170 = 46.7\%$, respectively, using Fisher's exact test. Note that this technique is optimal for the extraction of the lowest-frequency words, leading to identical performance for G^2 and Fisher's exact test for these words. For the higher-frequency words, Fisher's exact test leads to a slightly better recall with the same precision scores (0.31 for both tests).

While we have observed reasonable results with both G^2 and Fisher's exact test, we have not yet discussed how these results compare to the results that can be obtained with a technique commonly used in corpus linguistics based on the mutual information (MI) measure (Church and Hanks 1990):

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}. \tag{4}$$

In (4), y is the seed term and x a potential target word. A high MI score for a given target word suggests an association between this target and the seed term. Or perhaps more precisely, a low MI score suggests a dissociation between target and seed word (Manning and Schütze 1999). To compute recall, precision, and F, we require a cut-off value. As there is no theoretically motivated cut-off value, we vary it systematically. Panel (a) of Figure 7 plots the results for the *af* corpus. The x-axis represents the MI

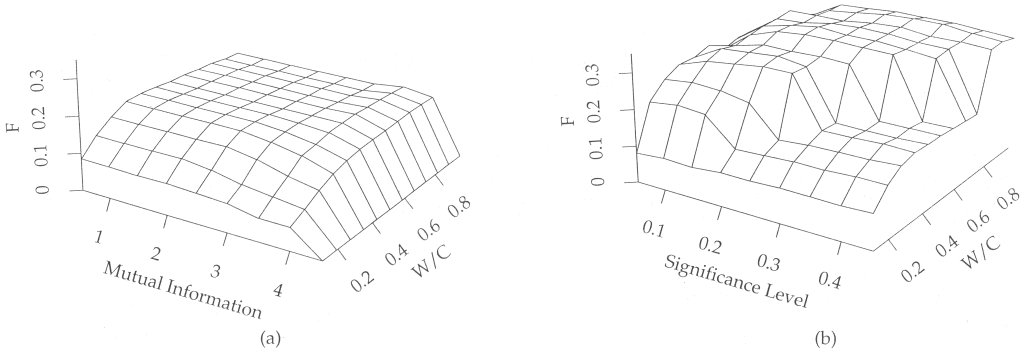

Figure 7
Extraction results (F) for the *af* corpus for mutual information and Fisher's exact test. Panel (a) shows the F score as a function of both W/C-ratio and mutual information cut-off value. Panel (b) shows F as a function of W/C-ratio and the significance level α used with Fisher's exact test.

cut-off value, the y-axis the W/C-ratio, and the z-axis the F value. Note that F is rather indifferent to variation in window size and MI cut-off value. It varies between 0 (at the right-hand edge) to 0.17, with most values around 0.15 (the plateau in the figure). Interestingly, the highest possible MI cut-off point equals 4.27: the right-hand edge of the plateau. In fact, 4.27 is the maximum MI score for this corpus size (42,825) and the frequency of the seed term *af* (2,206), irrespective of the frequency of the target word, reached when all occurrences of the target word are concentrated in the window (see the appendix for details). Consequently, any hapax legomenon appearing in the window will automatically be assigned the maximum value of MI, along with target words with the most extreme W/C distributions (Window–Complement: 2–0, 3–0, 4–0, etc.). This has the unfortunate consequence that, with regard to their MI score, truly remarkably distributed target words become indistinguishable from the statistically unremarkable hapax legomena.

Panel (b) of Figure 7 displays the corresponding results when we use Fisher's exact test rather than MI. Instead of varying the MI cut-off value, we vary the significance level α. Note that the resulting F scores tend to be roughly twice as high as those obtained with MI-based extraction. As there are a number of very similar local maxima, the choice of window size and significance level should be based on the desired trade-off between precision and recall given the general distribution of the target words around the seed term.[2] We conclude that, at least for the present word extraction task, Fisher's exact test compares favorably to mutual information (as does G^2).

All the analyses presented thus far are conditional analyses, in the sense that we compiled new corpora from the database of abstracts and from the newspaper corpus containing only relevant abstracts (containing the drug names captopril and enalapril as well as the term *side-effect*) and relevant sentences (containing the particle *af* and its verb to its left), respectively. The size of the complement was always determined with respect to these new conditional corpora, and not with respect to all MEDLINE

2 Note that we manipulate the α-levels in the same way as the MI cut-off values. In the present technique, the α-level is a parameter that we vary to optimize extraction results for a training data set. Our use of α should be carefully distinguished from the function of preset α-levels when testing the significance of observed differences in experimentally obtained data.

Table 4
General and specific 2 × 2 contingency tables for low-frequency words. Table (a) provides the general notation of the counts in a 2 × 2 contingency table. In table (b), A = frequency of rare words $(1, 2, 3, \ldots)$, W = number of words in window, C = number of words in complement. Corpus size $N = W + C$.

(a):

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n_{++}

(b):

A	0	A
$W - A$	C	$W + C - A$
W	C	$W + C$

abstracts or to the complete newspaper corpus. This raises the question of whether better results might have been obtained if the complete data sets had been used. In principle, more data might imply more power. At the same time, more data also entails the risk of more noise. At least for our *af* data, enlarging the complement leads to worse performance. When we allow any sentence that contains *af* in our analyses, F decreases from 0.31 to 0.23 for G^2. When we base the analyses on the complete newspaper corpus, F reduces further to 0.19. The reason for this decrease in performance is probably due to the W/C-ratio being very low for all practical window sizes, i.e., at the very left part of the saw-tooth-shaped pattern characterizing Nsig as a function of W/C. Consequently, any low-frequency word is singled out as a significant item whenever it occurs at least once in the window. Given the Zipfian structure of word-frequency distributions, a great many spurious low-frequency words are extracted.

As mentioned in the introduction, the received wisdom is that the windowing method is unreliable for events with a frequency of less than 5. By means of an analysis of the behavior of statistical tests for 2 × 2 contingency tables with sparse data, a method for optimizing the use of these tests has been developed. We hope that this technique will prove to be useful for domains in which the extraction of low-probability events is crucial.

Appendix

Log-Likelihood Ratio

For the general contingency table, table (a) in Table 4, the log-likelihood ratio is defined by (Agresti 1990):

$$G^2 = 2 \sum_i \sum_j n_{ij} \ln(n_{ij}/\hat{m}_{ij}),$$

where $\hat{m}_{ij} = n_{i+} n_{+j} / n_{++}$. When we use the specific contingency table for hapax legomena, table (b) in Table 4, we obtain for a specific G^2 of X the formula:

$$\begin{aligned} X/2 &= A \ln \frac{W + C}{W} + (W - A) \ln \frac{(W - A)(W + C)}{W(W + C - A)} + C \ln \frac{W + C}{W + C - A}, \\ &= \ln (W - A)^{W-A} - \ln W^W + \ln (W + C)^{W+C} - \ln (W + C - A)^{W+C-A}, \\ &= \ln \frac{(W - A)^{W-A} (W + C)^{W+C}}{W^W (W + C - A)^{W+C-A}}, \\ e^{X/2} &= \frac{(W - A)^{W-A} (W + C)^{W+C}}{W^W (W + C - A)^{W+C-A}}. \end{aligned}$$

We rewrite the latter equation to:

$$\begin{aligned}\frac{e^{X/2}W^W}{(W-A)^{W-A}} &= \frac{(W+C)^{W+C}}{(W+C-A)^{W+C-A}},\\ \frac{e^{X/2}W^W(W-A)^A}{(W-A)^W} &= \frac{(W+C)^W(W+C)^C(W+C-A)^A}{(W+C-A)^W(W+C-A)^C}.\end{aligned}$$

Because $W \gg A$ and therefore $W + C \gg A$, we rewrite the formula above as follows:

$$\begin{aligned}\frac{e^{X/2}W^WW^A}{W^W} &= \frac{(W+C)^W(W+C)^C(W+C)^A}{(W+C)^W(W+C)^C},\\ e^{X/2}W^A &= (W+C)^A,\\ \sqrt[A]{e^{X/2}}W &= W+C.\end{aligned}$$

So that the ratio is:

$$\frac{W}{C} = \frac{1}{\sqrt[A]{e^{X/2}} - 1}.$$

When $N > 10{,}000$, the error of this equation is smaller than 0.001.

Fisher's Exact Test
With Fisher's exact test, the observed marginal totals are used to compute the hypergeometric distribution, which is defined for the general 2×2 table, table (a) of Table 4, as (Agresti 1990):

$$\frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix}\begin{pmatrix} n_{2+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} n_{++} \\ n_{+1} \end{pmatrix}}.$$

The probability of every possible table with given marginal totals adds to 1. We use Fisher's exact test that sums the hypergeometric probabilities of all tables with probabilities less than or equal to the observed table. With $B = 0$, table (b) in Table 4 is the only table we are interested in so that the probability P for this contingency table is:

$$\begin{aligned}P &= \frac{\begin{pmatrix} A \\ A \end{pmatrix}\begin{pmatrix} W+C-A \\ W-A \end{pmatrix}}{\begin{pmatrix} W+C \\ W \end{pmatrix}},\\ &= \frac{\frac{(W+C-A)!}{(W-A)!C!}}{\frac{(W+C)!}{W!C!}},\\ &= \frac{W!(W+C-A)!}{(W-A)!(W+C)!},\\ &= \frac{W(W-1)\ldots(W-A+1)(W-A)!\quad(W+C-A)!}{(W-A)!\quad(W+C)(W+C-1)\ldots(W+C-A+1)(W+C-A)!}.\end{aligned}$$

Because $A = 1, 2, 3, \ldots$, $W \gg A$ and therefore $W + C \gg A$, we allow ourselves to formulate $W! = W^A(W \quad A)!$ and $(W + C)! = (W + C)^A(W + C - A)!$. We therefore rewrite Fisher's exact test as follows:

$$\begin{aligned} P &= \frac{W^A W!(W+C)!}{(W+C)^A W!(W+C)!}, \\ &= \frac{W^A}{(W+C)^A}, \\ \sqrt[A]{P} &= \frac{W}{W+C}. \end{aligned}$$

The W/C-ratio is then:

$$\frac{W}{C} = \frac{\sqrt[A]{P}}{1 - \sqrt[A]{P}}.$$

When $N > 20{,}000$, the error of this equation is smaller than 0.001.

Practical Issues Using Fisher's Exact Test. We used a network algorithm to compute Fisher's exact test (Mehta and Patel 1986; Clarkson, Fan, and Joe 1993). This algorithm is computationally intensive, but since many words have the same table, only a few tables have to be computed and their results can be cached. It takes an average of 50 seconds to compute one window size in a 100,000 word corpus on a Pentium 133MHz, 48MB Linux machine.

Source code for the algorithm can be found at: `http://www.acm.org/pubs/citations/journals/toms/1986-12-2/p154-mehta/`

Mutual Information

Given the definition of Mutual Information (Church and Hanks 1990),

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)},$$

we consider the distribution of a window word according to the contingency table (a) in Table 4. $P(x)$ is the relative frequency of the target word, $P(y)$ is the relative frequency of the seed term, and $P(x, y)$ is the frequency of the target word in the window. In terms of the contingency table, we have:

$$I(x, y) = \log_2 \frac{\frac{n_{11}}{n_{++}}}{\frac{n_{1+}}{n_{++}} \cdot \frac{S}{n_{++}}},$$

where S is the frequency of the seed. Substituting $n_{11} = n_{1+} - n_{12}$, we find that

$$\begin{aligned} I(x, y) &= \log_2 \frac{\frac{n_{1+} - n_{12}}{n_{++}}}{\frac{n_{1+}}{n_{++}} \cdot \frac{S}{n_{++}}}, \\ &= \log_2 \frac{\frac{1}{n_{++}}}{\frac{n_{1+}}{n_{++}(n_{1+} - n_{12})} \cdot \frac{S}{n_{++}}}, \end{aligned}$$

$$= \log_2(n_{++}) - \log_2(S) - \log_2(n_{1+}) + \log_2(n_{1+} - n_{12}).$$

For a given corpus and extraction task, corpus size (n_{++}) and the frequency of the seed term S are fixed, so that we can write

$$I(x,y) = \mathbf{C} - \log_2(n_{1+}) + \log_2(n_{1+} - n_{12}).$$

As $n_{12} < n_{1+}$, $I(x,y)$ reaches its maximum value ($\mathbf{C}$) when $n_{12} = 0$, i.e., when all instances of the target word are in the window, irrespective of the frequency of the target.

Acknowledgments

We are indebted to three anonymous reviewers whose criticisms have led to substantial improvements. This study was financially supported by the Dutch National Research Council NWO (PIONIER grant to the third author).

References

Agresti, Alan. 1990. *Categorical Data Analysis*. John Wiley & Sons, New York.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Clarkson, Douglas B., Yuan-An Fan, and Harry Joe. 1993. A remark on algorithm 643: FEXACT: An algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *ACM Transactions on Mathematical Software*, 19(4):484–488.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Manning, Christopher D. and Hinrich Schütze, 1999. *Foundations of Statistical Natural Language Processing*, chapter 5, Collocations. The MIT Press, Cambridge, MA.

Mehta, Cyrus R. and Nitin R. Patel. 1986. Algorithm 643. FEXACT: A fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software*, 12(2):154–161.

Pedersen, Ted. 1996. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, pages 188–200, Austin, TX.

Pedersen, Ted, Mehmet Kayaalp, and Rebecca Bruce. 1996. Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, pages 455–460.

Rijsbergen, Cornelis J. van. 1979. *Information Retrieval*. Second edition. Butterworths, London.

Rikken, Floor and Rein Vos. 1995. How adverse drug reactions can play a role in innovative drug research. *Pharmacy World and Science*, 17(6):195–200.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.