Running head: BIGRAM MEASURES AND ANAGRAMS

Type and Token Bigram Frequencies for Two through Nine Letter Words and the Prediction of Anagram Difficulty.

David Knight and Steven J. Muncer

University of Durham

Author Note

David Knight, Department of Psychology, University of Durham; Steven Muncer, Department of Psychology, University of Durham.

We thank Prof. Laura Novick of Vanderbilt University and Prof. Jim Sherman of Indiana University for generously allowing us to reanalyze their data.

Correspondence concerning this article should be addressed to Steven Muncer, Applied Psychology, University of Durham, Thornaby-on-Tees, TS17 6BH, United Kingdom. E-mail: s.j.muncer@durham.ac.uk.

Abstract

Recent research on anagram solution has produced two original findings. First it has shown that a new bigram frequency measure called Top Rank, which is based on a comparison of summed bigram frequencies, is an important predictor of anagram difficulty. Second, it has suggested that the measures from a type count are better than token measures at predicting anagram difficulty. Testing these hypotheses has been difficult because the computation of the bigram statistics is difficult. We present a program that calculates bigram measures for 2 to 9 letter words. We then show how the program can be used to compare the contribution of Top Rank and other bigram frequency measures derived from both a token and a type count. Contrary to previous research, we report that type measures are not better at predicting anagram solution times, and that Top Rank is not the best predictor of anagram difficulty. Lastly we use this program to show that type bigram frequencies are not as good as token bigram frequencies at predicting word identification reaction time.

Type and token bigram frequencies for two through nine letter words and the prediction of anagram difficulty.

In this article we describe a computer program that calculates bigram frequencies (two-letter sequences) for two through nine letter words and letter strings derived from a token count and a type count. This program is unique in calculating measures from both type frequencies and token frequencies from the same word corpus and thus, allows for a comparison of the importance of the different frequencies, which are defined below. The program calculates all of the major bigram frequency measures that have been proposed as being important in predicting anagram solution and word identification tasks. All of these bigram frequency measures are based on positional frequency counts of a bigram in a particular position in a word. The summed bigram frequency (SBF), for example, is the aggregate frequency of each bigram in each position. Examples of the calculation of all of the important bigram measures will be given below.

In order to demonstrate the use of the programme we re-examined research on five-letter anagrams by Novick and Sherman (2004, 2008) that challenged previous explanations of anagram solution in two major ways. First they suggested that bigram frequency measures calculated as a type measure of frequency are more important than those calculated as a token measure of frequency. Most previous anagram research has used token bigram frequencies provided by Mayzner and Tresselt (1965). They provided bigram counts derived from 100 samples of 200 words taken from a variety of newspapers, magazines, and both fiction and nonfiction books. They counted each instance of every bigram in 3 to 7 letter words in this corpus in each possible position. So for example, every time the word *because* appeared the initial bigram *be* received a count of 1 in the first position, and so on. If there were 20 appearances (tokens) of *because* in the word samples, then *be* in the first position received a frequency of 20 from this word alone. A type measure of bigram frequency indicates the number of different words that contain the targeted

bigram, rather than the number of tokens with the targeted bigram. So the *be* in the word *because* has a type count of 1. Solso, Topper, and Macey (1973) made a similar distinction between what they called bigram frequency and bigram "versatility". They used the example of the bigram *of* to demonstrate that there will be differences between bigram frequency (token) and versatility (type) measures. The bigram *of* has a relatively high token frequency in the English language but this is largely based on the frequency of the word *of*.

Novick and Sherman (2004) suggested that a type frequency would be a better predictor of anagram difficulty because it was not confounded with word frequency. Novick and Sherman (2004) also pointed out that the Mayzner and Tresselt (1965) tables were derived from only a small subset of the words in the English language. For example, there were only 856 different five-letter words. Accordingly they produced a set of type frequencies for five-letter English words based on 2,550 different words, which were used in their program.

Novick and Sherman (2004) also proposed that a new bigram frequency measure called Top Rank, which was calculated from the type frequencies used in their computer program, would be a better predictor of anagram solution. They compared Top Rank against the SBF measure derived from Mayzner and Tresselt's (1965) token frequencies and found that the Top Rank measure was a better predictor of anagram solution time. Novick and Sherman (2004) concluded "that type-based bigram frequency is a better predictor of the difficulty of anagram solution than is token-based frequency" (p.397).

The relative importance of type and token measures is a controversial debate in word processing (Hofmann, Stenneken, Conrad & Jacobs, 2007). There is, however, no doubt that its resolution will be very important for current and future models of word recognition in general (Conrad, Carreiras & Jacobs, 2008). We believe that Novick and Sherman's (2004) conclusion that type frequencies were more important than token measures is misleading, as their study confounded two variables. They

have shown that a type frequency (Top Rank) was more important than a token frequency (SBF), but did not make the appropriate comparison between a type and token frequency measure of both SBF and Top Rank. They have also ignored several other bigram frequency variables that have been considered more important in anagram solution than SBF, and which may also be more important than their Top Rank measure.

Previous anagram research has suggested a multitude of variables that influence the difficulty of anagrams. Some of these are related to features of the solution word such as; word imagery, concreteness of the word, familiarity, objective frequency, age of acquisition, meaningfulness, number of vowels, starting letter and some bigram frequency measures. Others are related to the composition of the anagram such as the similarity of the word and anagram, and the pronounceability of the anagram (Gilhooly & Johnson, 1978). Later research has suggested that the bigram frequency measures are the most important features of the solution word in predicting anagram difficulty (Gilhooly & Johnson, 1978; Mendelsohn, 1976; Mendelsohn & O'Brien, 1974).

The three main frequency measures that were proposed for use in the prediction of anagram difficulty were the summed bigram frequency (SBF; Mayzner & Tresselt, 1959), bigram rank (BR; Mendelsohn & O'Brien, 1974) and greater-than-zero (GTZero; Mendelsohn, 1976). Novick and Sherman (2004) only included SBF and their new Top Rank measure in their analysis. The starting frequencies for all of these measures were usually provided by the Mayzner and Tresselt (1965) tables that give frequencies for each bigram, in each position for words between 3 and 7 letters long. As mentioned earlier, SBF is the aggregated frequency of each bigram in each position. For example, for the word *light* we would calculate the frequency of *li* in the first two positions of a five-letter word (*li* = 36), the frequency of *ig* in positions two and three (*ig* = 101), the frequency of *gh* in positions three and four (*gh* = 94) and

the frequency of *ht* in positions four and five (*ht* = 95). These frequencies are then summed to produce a token SBF for *light* of 326.

Top Rank is based on a comparison of the summed bigram frequency (SBF), which is the sum of the frequencies of a word's successive bigrams. A word is top ranked if it has the highest SBF of the 120 possible combinations of bigrams for a five-letter word. Novick and Sherman's (2004, 2008) Top Rank was based on a comparison of the SBF of the type frequencies of each of the words used in their study, but it is possible to produce Top Rank from token frequencies as well.

Bigram Rank (Mendelsohn & O'Brien, 1974) can also be calculated from the Mayzner and Tresselt (1965) tables, although it is not easily calculated by hand as it requires the completion of a bigram frequency matrix for all of the bigrams (all combinations of two letter sequences) in a word in all of the possible positions. There are 80 cells in the matrix as there are four bigram positions in a five-letter word (see above) and there are 20 possible different combinations of bigrams. Bigram Rank (BR) is the total number of cell entries in the bigram frequency matrix that have higher frequencies than the four correct cell frequencies. For example, the bigram frequency matrix for the word *beach* shows that in the correct position *be* has a frequency of 41, *ea* in the correct position a frequency of 134, *ac* of 51 and *ch* of 126. In the matrix for *beach* there are 5 incorrect bigrams that have higher values than *be*, 1 incorrect bigram has a higher frequency than *ea*, 5 exceed *ac* and 1 exceeds the value of *ch*. Thus the BR for *beach* is 12. The incorrect bigram with the highest frequency, which is higher than both *ea* and *ch,* is *he* in the second and third position (263). Higher numbers for BR thus indicate more competition for the locations of the letters and therefore greater difficulty of solution.

GTZero (Mendelsohn, 1976) is also calculated from the bigram frequency matrix. GTZero is the total number of bigrams in a word with a frequency greater than zero in the bigram frequency matrix. For example, for the anagram *igthl* (*Light*), *hg, ht, hl, gt, tg, tl, lh, lg, lt* would all have a frequency of 0 in the first two positions. The measure

is a development from Ronning's (1965) "rule-out" theory of anagram solution, which proposes that anagrams with a low number of bigram combinations that can be "ruled-out" of consideration will be harder to solve. The more non-zero entries there are, the greater the possible competing solutions, which makes the anagram harder to solve (Mendelsohn, 1976). Again for any five-letter word, there are 80 possible cells in the bigram frequency matrix, as each bigram can appear in four positions. That is, in the word *light*, the bigram *li* can appear in competing words in the first and second position, the second and third position, the third and fourth position or the fourth and fifth position, and there are 20 possible bigram combinations in a five-letter word. The GTZero for the word *light* is 33.

Novick and Sherman (2004) produced a computer program that calculates SBF and their new bigram measure, Top Rank, for five-letter words. The program, however, does not calculate GTZero or BR. They did not include these bigram frequency measures in their analysis, but only included SBF which fewer researchers have argued predicts anagram difficulty. Furthermore, they did not include a Top Rank measure from a token frequency count. Our re-analysis will correct these omissions by examining the relative importance of all bigram frequency measures.

It is possible to calculate a type and token count for all of the bigram frequency measures apart from GTZero. This is because GTZero is calculated by counting the nonzero bigram frequencies in the bigram frequency matrix, without considering their size as do other measures. So if a cell has a frequency above 1 it will contribute to the GTZero score in the same way from either a bigram frequency matrix based on type or token counts. For GTZero, the distinction between type-and token frequencies is irrelevant conceptually. It is, however, not possible to produce both type and token frequencies even for the other variables from either the program described by Novick and Sherman (2004), or the tables described in Mayzner and Tresselt (1965) of token frequency, as neither provide an alternative. Furthermore, as these two studies are based on different word corpora, any differences in

predictive ability might be based on differences in the size and quality of the word corpora. We have, therefore, used the bigram frequencies provided by Solso and Juel (1980) to derive a set of type and token statistics from the Kucera and Francis (1967) corpus. Solso and Juel (1980) refer to the token frequencies as "positional frequencies", which is how many times a bigram appears in a specific position per one million words. They called type bigram frequencies "versatilities", which is in how many different words a bigram appears in a specific position per one million words.

Novick and Sherman (2004) noted that one of the limitations of the Solso and Juel (1980) norms was that only a printed version of them existed and that calculating these bigram measures from frequency tables is both a laborious and potentially error-prone process. For example, Seidenberg (1987, 1989) argued that the evidence for the syllable as a functional unit during reading could be explained by the fact that bigram frequencies at the boundary of two syllables are less than intrasyllabic bigram frequencies. Readers could, therefore, be sensitive to this bigram trough in which bigram frequencies at a syllable boundary are lower than the preceding and following bigram, rather than to syllables per se. Rapp (1992) conducted an experiment to test the "bigram trough" hypothesis that required the calculation of bigram frequencies at, before and after a syllable boundary. In Rapp's (1992) experiment, the bigram frequencies were calculated incorrectly for 4 out of 117 words. Errors are even more likely in the calculation of more complicated statistics like BR and GTZero that require the completion of a frequency matrix. For example, in Gilhooly and Johnson's (1978) study of the impact of 12 variables on anagram solution, their token bigram frequency measure called GTZero was incorrect in 38 of 80 cases.

Our computer program calculates all relevant bigram frequency statistics that have been used in anagram research in both type and token form. We used this program to calculate all relevant bigram measures for the data reported by Novick

and Sherman (2004) and others. This enabled a thorough comparison of both the importance of the type and token distinction, and an evaluation of the different bigram frequency measures in predicting anagram difficulty. This comparison was not confounded by size and quality of word corpora as both counts were based on the same corpus. Novick and Sherman (2004) noted that the calculation of bigram and other sublexical frequencies has been important "across a variety of fields over at least the past 40 years" (p.397). The importance of sublexical frequency measures in general language processing is also recognised (Aichert & Zielgler, 2005: Hoffman et al., 2007). Therefore, we also examined the relationship between the various bigram measures and time taken on a lexical decision task, in order to demonstrate the wider applications of the program.

The Program

There are 577 different bigrams in the Solso and Juel (1980) tables with a frequency count for both word tokens and for word types in each bigram position for words between 2 and 9 letters long. Each bigram can appear in numerous positions dependent on word length. So for a two-letter word there is only one bigram position, first and second, but for a five-letter word a bigram can appear in four positions and so on. Our computer program called "Bigram calculator for Solso and Juel" computes all of the major bigram frequency statistics for any letter string with a length of between 2 and 9 letters using the Solso and Juel (1980) tables. These include the simplest statistics such as the bigram frequency in each position, which could be used to calculate bigram troughs, and also SBF. It also calculates the more complicated statistics such as BR, GTZero, and the Top Rank measure suggested by Novick and Sherman (2004). In order to calculate Top Rank it has to compute the SBF of the 120 possible orders of the five letters in the letter string and then rank them. Our program refers to this ranking as the Likelihood Rank (LR), as from a bigram frequency perspective it is a measure of the likelihood of that combination of

letters. The program also gives the lowest ranked combination of letters and the SBF for this lowest combination.

Our MS Windows based program consists of six files (compressed as a single zip file) that can be downloaded from http://spider.dur.ac.uk/bical. The "Table.csv" file consists of position-sensitive token and type bigram norms taken from Solso and Juel (1980) for 2 to 9 letter words. The "BiCal.exe" file is the main executable program file and provides the bigram frequency for each position for the correct solution, the SBF, GTZero, Bigram Rank, Likelihood Rank and Top Rank based on these norms. The final 4 files make up the help system that gives directions on running the program and can be accessed either from within the program from the "Help" menu, or by launching the "BiCal.hlp" file.

In order to run the program a word is entered into the Input box and the 'Go' button is clicked. The token bigram frequency totals for each bigram, their sum (SBF), BR, GTZero and LR appear on the left, and the type frequencies on the right. Words with an LR of 1 would be given a Top Rank score of 1 and the rest 0. For the word *light*, the token SBF is 8361, the BR is 12 and the LR is 8, whereas the type SBF is 89, BR is 51 and the LR is 8. The GTZero from both the type and token frequencies is 49 and will always be identical, as explained earlier. GTZero, however, is likely to be affected by the size of the corpus from which it is calculated; the more words that are included the greater the likelihood that one of them will have a bigram in each position. For example, for the word 'light' the GTZero from Solso and Juel is 49, calculated from Novick and Sherman's (2004) type frequencies it is 45, whereas it is 33 when calculated from Mayzner and Tresselt's (1965) smaller token corpus. It is possible, of course, that if a corpus is very large it will include many words that are unfamiliar to most people and this may mean that it is less effective at predicting anagram solution.

The program can take several seconds to calculate results for longer words. This is due to the number of calculations necessary to work out the Likelihood Rank, as

the number of possible permutations of the letters in the word increases exponentially with the word length. For instance, a 9 letter word consists of 362,880 different possible letter combinations and these contain a total of 2,903,040 bigrams. Each of these has to be looked up in the Solso and Juel (1980) statistics table in order to calculate the SBF for that permutation of letters, and then the results are ranked.

The program can also perform batch processing on a list of words. The list of words to be processed should be saved as a standard text file with each word on a separate line. This list of words can then be loaded into the program by going to the "File" menu and selecting the "Load Word List" option. The program will scan the list and will automatically reject any words that are not between 2 and 9 characters in length (and will notify the user of this) – although it will continue processing any subsequent words. After this, the program prompts for the name of an output results file (again, a standard text file). The program calculates the various results/scores for each of the valid words in the list and then saves these into the results file. This file can then be viewed and analysed accordingly.

In the present study we have used frequency statistics from this program to re-examine Novick and Sherman's (2004) data in a regression that includes all relevant bigram statistics calculated in both type and token forms as independent variables, and solution time as a dependent variable. We have also compared the type and token measures as predictors of lexical decision time using data from Balota et al. (2002). This follows the suggestion of Hofmann et al. (2007) who argued that a useful contribution to the type/token controversy "would be to conduct a regression analysis with type and token measures as predictors, to find out which measure is most predictive" (Hofmann et al., 2007, p. 623). Lastly, we have examined evidence for the "bigram trough hypothesis" as an explanation of word identification times.

Results

Re-analysis of Novick and Sherman (2004)

Novick and Sherman (2004) provided solution time and accuracy of solution scores for 108 five-letter anagrams, which they generously made available to us. As the correlation between the dependent measures is extremely high, $r(106) = -.97$, $p < .0005$, only the reaction time measure will be reported in detail.

The correlations between reaction time to solve the anagram (maximum time allowed of 30 sec) and all of the bigram frequency measures are shown in Table 1.

Insert Table 1 here

It is apparent that neither the token nor the type SBF are significantly correlated with reaction time; nor are they significantly different to each other, $t(105) = 0.92$, $p > .05$. Although type Top Rank is significantly correlated with reaction time, it does not have a significantly higher correlation with reaction time, $t(105) = 1.11$, $p > .05$, than token Top Rank. In fact, there are no significant differences between any of the correlations between the various matched token and type measures and reaction time. Furthermore, there were also no significant differences between the matched token and type measures of SBF, $t(105) = .013$, $p > .05$ and Top Rank, $t(105) = 1.19$ $p > .05$ with reaction time, when these are calculated using the Novick and Sherman frequency count and the Mayzner and Tresselt (1965) count.

We conducted a stepwise regression analysis for reaction time with both the Top Rank variable and Bigram Rank (BR) calculated from the type and token counts and GTZero as independent variables. We used the Top Rank variable rather than Likelihood Rank (LR), which has a higher correlation with the dependent measure, because it was found to be an important predictor in Novick and Sherman's (2004) earlier analysis of the data. In addition, the type and token versions of the Top Rank measure are more distinct than are those versions of the LR measure, $r(106) = .35$ vs. $R(106) = .48$, respectively. The most important predictive variable was GTZero,

R = .53, $F(1,106) = 42.31$, $p < .001$ with a β = .53, $t(106) = 6.51$, $p < .001$. The second and only other variable to be entered was BR calculated from the token norms, R = .58, $F(2,105) = 26.33$ with a β =.24, $t(105) = 2.77$, $p < .01$. GTZero has a significantly higher correlation with reaction time than type Top Rank, $t(105) = 3.15$ $p < .01$. Thus it is clear that GTZero is a better predictor of anagram solution times than type Top Rank.

Novick and Sherman (2004) pointed out that the Kucera and Francis (1967) word frequencies were based on a subset of English words and would, therefore, be less inclusive than their frequencies based on dictionary definitions. In particular they argued that a number of ordinary words were omitted that would contribute to people's knowledge of bigram frequencies. We, therefore, conducted a second stepwise regression in which we included Novick and Sherman's (2004) estimates of Top Rank, but this did not change the regression equation. Furthermore, GTZero calculated from Solso and Juel (1980) has a significantly higher correlation with anagram solution reaction time than Novick and Sherman's (2004) Top rank measure, $t(105) = 2.53$, $p < .01$, as does GTZero calculated from Mayzner and Tresselt (1965), $t(105) = 2.29$, $p < .05$.

Results also showed that the matched type and token measures were highly correlated when taken from the same corpus of words. In particular, BR, which is one of the more important variables for anagram prediction, was correlated at $r(106) = .80$, $p < .001$ between the token and type measures of Solso and Juel (1980). BR was also highly correlated across corpora. For example, Novick and Sherman's BR was correlated with Solso and Juel's (1980) type BR at $r(106) = .77$, $p < .001$ and with their token BR at $r(106) = .78$, $p < .001$. Furthermore, there were no significant differences between any of the matched type and token correlations with solution time when they were taken from the Solso and Juel (1980) corpus.

It is interesting to note that the GTZero measure based on frequencies from Mayzner and Tresselt (1965) had a similar correlation, $r(106) = .52$, $p < .001$, with

solution time as the GTZero based on Solso and Juel (1980). The Mayzner and Tresselt (1965) GTZero was significantly lower (M = 42.06, SD = 7.60) than the Solso and Juel (1980) GTZero (M = 60.08, SD = 6.89), $t(107) = 44.10$, $p < .0005$; $r(106) = .83$, $p < .001$, but the correlation with solution time was similar. The larger corpus of words does not seem to improve the quality of prediction of GTZero in this case. It is argued that GTZero is important as it provides a measure of which bigram combinations and positions can be 'ruled out' as possible solutions. Hence the higher the GTZero, the harder the anagram is to solve. It is worth noting here, however, that the smaller corpus may reflect a lay person's knowledge of words adequately, as a larger corpus may include more unfamiliar words that will suggest that some unlikely bigram positions are permissible.

Our reanalysis of Novick and Sherman (2004) suggested that the distinction between type and token frequencies is of little importance in predicting time for anagram solution. Generalizing the results of language experiments is particularly problematic, however, as a significant result tells us only that the result is likely to generalize to a new set of participants and not necessarily to a new set of stimuli (Clark, 1973; Coleman, 1964). It was important to demonstrate that these results apply to other anagrams, other indices of anagram difficulty as well as other participants

Re-analysis of other anagram studies

In order to investigate whether similar results would be obtained with different anagrams and different participants, we re-analysed the results from Gilhooly and Johnson's (1978) study that looked at ease of solution of 80 five-letter anagrams. This study used the number of participants successfully solving the anagram as the dependent measure. In their original analysis, Gilhooly and Johnson (1978) found that starting letter, anagram solution similarity to the target word, pronounceability of the anagram and two token bigram frequency measures were most important in determining anagram difficulty. The more important of these two bigram measures

was GTZero, which was calculated by hand from a bigram frequency matrix from Mayzner and Tresselt (1965).

Our reanalysis using all relevant measures derived from Solso and Juel (1980) showed that on this occasion GTZero had the highest correlation with anagram solution score, $r(78) = -.46$, $p < .005$). Again there were no significant differences between the correlation of any of the matched type and token measures from the same corpus and solution score.

The importance of GTZero and relative unimportance of the type-token distinction can also be seen in other anagram studies. For example in the Mayzner and Tresselt (1966) study that looked at solution times for 42 anagrams in six conditions, the correlation between the average time taken to solve an anagram across conditions and GTZero calculated from our program is $r(40) = .35$, $p < .05$. Similarly for Ronning's (1965) study, GTZero has a correlation of $r(18) = .61$, $p < .02$, with solution time. In every case the correlation of GTZero and the anagram difficulty measure was higher than the Top Rank measure whether calculated from Novick and Sherman (2004) or from Solso and Juel (1980). Furthermore both the token- and type- BR measures were also highly correlated with solution score, and there was never a significant difference between the sizes of the correlation with the dependent measure. In fact, there were no significant differences between any of the matched type and token measures and the dependent variable.

GTZero derived either from Mayzner and Tresselt (1965) or from our program was a good predictor of anagram solution, and there were no significant differences between them. Furthermore, the GTZero measure calculated from Novick and Sherman's (2004) frequencies is an equally good predictor. There is, therefore, clear evidence that the distinction between type and token counts has little importance for anagram solution. It is important to note, however, that anagram problems may have limited relevance to word processing skills. In particular, it is unlikely that GTZero will be a useful variable in predicting performance on other psycholinguistic tasks, as it

seems particularly suited to the demands of anagram solution. The next section explores this issue by looking at the relationship of GTZero with a lexical decision task.

Bigram measures and lexical decision making

We investigated the relationship of GTZero with reaction time in a lexical decision task for the words employed by Novick and Sherman (2004) and Gilhooly and Johnson (1978). The English Lexicon Project (Balota et al., 2002) provided lexical decision reaction times for 173 words used in the above studies after we removed any duplicates and words for which times were not available. In their lexical decision task (Balota et al., 2002) participants were presented with a string of letters (either a word or nonword) and asked to press one button if the string was a word and another button if it was a nonword. As expected we found no relationship between GTZero and mean lexical decision reaction time, $r(171) = .03$, $p = .72$.

Given the nature of the lexical decision task, one might expect the LR measure to be a reasonable predictor of lexical decision reaction time. LR indicates the relative frequency of the combination of bigrams in a word compared to other combinations. Frequent combinations should, therefore, be most like English words and therefore quicker to identify. In this case the correlation between token LR with reaction time, $r(171) = .29$, $p < .0005$, was not significantly higher than the type LR correlation with reaction time, $r(171) = .17$, $p < .05$; $t(170) = 1.75$, $p < .10$. Surprisingly the highest correlation with lexical decision reaction time is between the token BR measure, $r(171) = .31$, $p < .0005$, which is significantly higher than the correlation between the type BR measure and reaction time, $r(171) = .17$, $p < .05$; $t(170) = 3.14$, $p < .01$. It is also worth noting that this pattern of results would be the same if Solso and Juel's token measure was compared to a type measure derived from Novick and Sherman (2004).

Overall the correlations between the token measures and lexical decision reaction time were never significantly lower than those of type measures and sometimes they were significantly higher.

It should also be noted that the correlation between bigram frequencies taken from a token and type count will be fairly high for five-letter words, which makes it unlikely that there would be significant differences between their correlation with other variables. The correlation between token and type measures from Solso and Juel (1980) regardless of word length is $r(4614) = .41$, $p < .0001$. The correlation between total token and total type bigram frequency for five-letter words is $r(575) = .75$, $p < .001$, and it was significantly higher for six-letter words ($r(575) = .95$; $z = 13.74$, $p < .001$). Smaller words, of course, have smaller correlations as there are bigger discrepancies in their frequencies. For example, the correlation was only $r(577) = .59$, $p < .0001$, for four-letter words and was significantly lower for three-letter words $r(577) = .17$, $p < .001$; $z = 8.57$, $p < .001$. Therefore, the distinction between type and token counts might well be more important for words of less than five letters.

The Bigram Trough Hypothesis

The lexical decision data and our program can also be used to examine the bigram trough hypothesis, which argues that syllable effects are caused by differences in bigram frequency. The number of syllables in a word has been shown to be positively correlated with reaction time in lexical decision tasks, even after important covariates such as word length have been controlled (Yap & Balota, 2009). There was a significant correlation between number of syllables and reaction time, $r(171) = .18$, $p = .02$, in the present data. If Seidenberg (1987) is correct the syllable effect should be caused by the presence of bigram troughs in multisyllabic words. It is surprising, however, that there are quite a large number of one syllable words (57/114) that have a bigram trough as defined by Rapp (1992). For example, in the word *blush* there is a trough between the first and third bigrams; the first bigram *bl*

has a frequency of 687, the second bigram *lu* only has a frequency of 205 and the third bigram *us* has a frequency of 919. There are also quite a few multisyllablic words that do not have a bigram trough (17/59). For example, in the word *basic;* the first bigram *ba* has a frequency of 812, the second bigram *as* has a frequency of 939, the third bigram *si* has a frequency of 714 and the fourth bigram *ic* has a frequency of 595. It is true that multisyllabic words are more likely to have a trough, χ^2 (1, N = 173) = 7.02, p = .01, but the large number of one syllable words with a trough makes it unlikely that it can be an explanation of the syllable number effect.

So far, research examining the bigram trough hypothesis as a cause of the various syllable effects in reading has focused on multisyllabic stimuli (Conrad, Carreiras, Tamm & Jacobs, 2009; Rapp, 1992). It is clear, however, that if such effects are caused by relative bigram frequencies, they should also occur in one syllable words with troughs. In this case, there was no significant difference in reaction times between one syllable words with and without troughs, t(112) = 1.12, p = .26. There was also no significant difference in reaction times between two syllable words with and without troughs, t(57) = .60, p =.55. The latter result offers some support to the recent finding by Conrad et al (2009) that syllable frequency effects in bisyllabic Spanish words were unaffected by the presence or absence of a bigram trough.

Discussion

We have presented a program that will calculate GTZero, Bigram Rank, Likelihood Rank and SBF from both type- and token-based systems simultaneously using the Solso and Juel (1980) frequencies. This program will perform these calculations on any word or letter string between two and nine letters long. This program extends previous research that provided a program for the calculation of some of these bigram statistics from a type-based system for five-letter words (Novick & Sherman, 2004).

Our reanalysis of previous research using this program has shown that GTZero, and Bigram Rank calculated from either a type or token count, are the most important bigram frequency variables in determining the difficulty of anagrams. We have also shown that there is no evidence that type-based systems will be superior in predicting anagram difficulty.

We have demonstrated that this program will be useful not just in anagram research but also in areas of visual word recognition. In particular, we have shown that the distinction between type and token bigram frequencies has little importance when predicting word identification reaction times. In fact, in our analysis, which uses all of the matched token and type frequencies, the token measure of Bigram Rank was the best predictor of word identification reaction times, and was significantly better than the type measure of Bigram Rank derived either from Novick and Sherman (2004) or Solso and Juel (1980). This is particularly important because although anagram solution may be regarded as an unusual task, lexical decision has been described as a "*defacto* gold standard in visual word recognition research" (Yap & Balota, 2009, p. 502).

Lastly we used the program to investigate the bigram trough explanation of the syllable number effect. Previous research in English on the bigram trough hypothesis has used frequencies derived by hand from the Solso and Juel (1980) tables, which as we noted is a time-consuming process. In this research only multisyllabic words have been investigated even though monosyllabic words also have troughs as we have shown. It was in fact surprising that so many monosyllabic words in our sample had a bigram trough, and this makes it unlikely that syllable effects are caused by troughs. Clearly if it is the presence or absence of a bigram trough that causes syllable effects then one syllable words with troughs should have longer lexical decision reaction times than those without troughs. We found, however that there were no significant differences in lexical decision time between words with and without troughs for either mono- or polysyllabic words. It would be important to

demonstrate this again with other variables considered, such as consonant-vowel structure. Our program makes it easy to include variables such as a bigram trough and its relative position in any future studies of syllabic or other effects in word processing. We believe that our program should be useful for all research that looks at the impact of sublexical features on visual word recognition and reading processes (Aichert & Ziegler, 2005).

References

Aichert, I. & Ziegler, W. (2005). Is there a need to control for sublexical frequencies? *Brain and Language, 95,* 170-171.

Balota, D. A., Cortese, M.J., Hutchison, K.A., Loftis, B., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). *The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English Words and Nonwords*, Washington University, http://elexicon.wustl.edu/

Clark, H.H. (1973). The language as fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behaviour*, 12, 335-339.

Coleman, E.B. (1964). Generalizing to a language population. *Psychological Reports*, 14, 219-226.

Conrad, M., Carreiras, M., & Jacobs, A. M. (2008). Contrasting effects of token and type syllable frequency in lexical decision. *Language and Cognitive Processes, 23*, 296-326.

Conrad, M., Carreiras, M., Tamm, S., & Jacobs, A.M. (2009). Syllables and bigrams: Orthographic redundancy and syllabic units affect visual word recognition at different processing levels. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), 461-479.

Gilhooly, K., & Johnson, C.E. (1978). Effects of solution word attributes on anagram difficulty: A regression analysis. *Quarterly Journal of Experimental Psychology, 30*, 57-70.

Hofmann, M.J., Stenneken, P., Conrad, M, & Jacobs, A.M. (2007). Sublexical frequency measures of orthographic and phonological units in German. Behavior Research Methods, Instruments, & Computers, *36*, 397-401.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Mayzner, M.S. & Tresselt, M.E. (1959). Anagram solution times: A function of transition probabilities. *Journal of Psychology*, *47*, 117-125.

Mayzner, M. S., & Tresselt, M. E. (1965). Tables of single-letter and bigram frequency counts for various word length and letter position combinations. *Psychonomic Monograph Supplement, 1*, 13-31.

Mayzner, M. S., & Tresselt, M. E. (1966). Anagram solution times: a function of multiple- solution anagrams. *Journal of Experimental Psychology, 71*, 66-73.

Mendelsohn, G. A. (1976). An hypothesis approach to the solution of anagrams. *Memory & Cognition, 4*, 637-642.

Mendelsohn, G. A. & O'Brien, A.T. (1974). The solution of anagrams: A reexamination of the effects of letter transition probabilities, letter moves, and word frequency on anagram difficulty. *Memory and Cognition, 3,* 566-574.

Novick, L. R., & Sherman, S. J. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods , Instruments and Computers, 36(3),* 397-401.

Novick, L. R., & Sherman, S. J. (2008). The effects of superficial and structural information on on-line problem solving for good versus poor anagram solvers. *Quarterly Journal of Experimental Psychology*, *61*, 1098-1120.

Rapp, B. (1992). The nature of sub-lexical orthographic organization: The bigram trough hypothesis examined. *Journal of Memory and Language*, *25*, 461-475.

Ronning, R.R. (1965). Anagram solution times: A function of the "ruleout" factor. *Journal of Experimental Psychology, 69,* 35-39.

Seidenberg, M.S. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy? In M. Coltheart (Ed.), *Attention and Performance XII: The psychology of reading* (pp. 245-263). Hillsdale, NJ: Erlbaum.

Seidenberg, M.S. (1989). Reading complex words. In G.Carlson & M.Tannenhaus (Eds.), Linguistic structure in language processing (pp. 53-105). Dordrecht, the Netherlands: Kluver Academic.

Solso, R.L., & Juel, C.L. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods & Instrumentation*, *12(3)*, 297-343.

Solso, R. L., Topper, G. E., & Macey, W. H. (1973). Anagram solution as a function of bigram versatility. *Journal of Experimental Psychology, 100*, 259-162.

Yap, M.J., & Balota, D.A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60(4), 502-509.

Table 1

Intercorrelation between Solson and Juel (1980) bigram measures and anagram solution time for Novick and Sherman (2004).

Variable	1	2	3	4	5	6	7	8	9	10
					Words (n=108)					
1 GT0	--	-.13	-.19	.43**	.42**	.28*	.24*	-.25*	-.39**	.53**
2 Token SBF		--	.41**	-.60**	-.43**	-.49**	-.28**	.46**	.35**	-.06
3 Type SBF			--	-.36**	-.52**	-.35**	-.60**	.22	.28*	.04
4 Token Bigram Rank				--	.80**	.74**	.61**	-.52**	-.47**	.43**
5 Type Bigram Rank					--	.52**	.85**	-.43**	-.58**	.36**
6 Token Likelihood Rank						--	.48**	-.40**	-.32**	.26*
7 Type Likelihood Rank							--	-.24*	-.40**	.26*
8 Token Top Rank								--	.35**	-.13
9 Type Top Rank									--	-.25*
10 Reaction time										--

*p <.01, **p <.001