Reducing the Word Length-Token Frequency Function to an Equation[1]

Michael Gradoville
University of New Mexico

Abstract

The fact that language structure is affected by usage is a cornerstone to functional linguistics. One specific idea that is generally accepted is that the words with the greatest token frequency are also the shortest (e.g. Bybee, 2002). The purpose of this paper is to outline a statistical method that may be used to perform tests on corpus data related to the word length-token frequency function. The data used to develop this method come from the spoken portion of Davies' (2005) *Corpus del español*, a 100 million word corpus of the Spanish language including sources from eight centuries. A rank-order list that includes the number of occurrences of each form was extracted from the *Corpus del español* and the 1000 most frequent forms were then tagged for length in terms of number of syllables. Using linear regression analysis, equations were created from the data presenting word length to be a function of rank in the list in one case and frequency of occurrence in the other. These equations represent an approximate average word length at any point in the rank-order list. Details for selecting data are discussed and possible future applications of this method are outlined.

0. Introduction

This article was motivated by Fenk-Oczlon (2001), which dealt with token frequency and its relationship to initial consonant obstruency in English. A question arose related to the universality of that relationship: Has anyone ever studied how often frequency was principally responsible for the reduction of an initial consonant? The resulting study evolved into an investigation regarding the statistical relationship between word length and frequency of occurrence. It is a commonly accepted fact that the most frequent words in a given language are also the shortest in length. This was first studied by Keading (1898) on German and then in the works of George Kingsley Zipf (1936). It is necessary to revisit the question of just how much frequency influences the length of words in the languages of the world.

Bybee (2002) reports on several current reductive phonological changes that serve as evidence that reductive sound changes are subject to a process of lexical diffusion, with the most frequent words in the language subject to the change before the less frequent. Given that reductive sound changes are more advanced in the most frequent words of a language, it stands to reason that the less frequent words may not have undergone as many reductive sound changes and that fact would be revealed in a statistical analysis of the phenomenon (cf. Zipf, 1936). Herein a methodology is outlined to allow a variety of statistical analyses of corpus data.

The methodology is designed in such a way that, given the proper data, it will be possible to compare these factors between different languages, particularly related languages. For example, the methodology might be applied to compare how effective reductive sound changes were in the development of the Romance languages from Latin or the Indo-Aryan languages from Sanskrit. Within the Romance languages, the Latin *factum* [‘fak.tum] yielded Portuguese *feito* [‘fei.tu], Spanish *hecho* [‘e.t🌢o], French *fait* [‘f☜(t)], and Italian *fatto* [‘fat.to]. The French example has clearly suffered more reductive sound change than the examples from the other languages, but one word does not paint a picture of the entire language. Thus, it would be useful to perform a statistical comparison between languages.

0.1. Previous Studies in Word Length

The general relationship between word length in syllables and frequency of occurrence is established in Kaeding (1898) for the German language and further elaborated by Zipf (1936). More recently, Leopold (1998) attempts to place specific word lengths within certain frequency spectra. Regarding other factors influencing word length, Nettle (1998) takes words of various frequencies from twelve West African languages to examine average word length in a language as a function of phonological inventory. Nettle’s study emphasizes the adaptive nature of language. Indeed, a casual comparison of French and Spanish would reveal that the French words seem to be shorter and that French seems to have a much larger phonemic inventory. The correlation holds in that case.

Ultimately, the synchronic relationships noted by Kaeding (1898) and Zipf (1936) are a result of the diachronic processes of lexical diffusion described by Bybee (2002). The present study contributes a methodology that permits the statistical study of this synchronic relic of diachronic processes. As such, future studies based upon the present work will contribute to what we know of the global effect of those processes, particularly their uniformity, and the relationship between languages with respect to the statistical structure of their lexicons.

0.2. Overview of Sections

This paper is divided into the following sections. Section 1 is dedicated to the process of selecting data sources. Although finding a corpus or corpora may seem a relatively easy task, one must be selective in choosing the content of the corpus. Section 1 notes some guidelines that should be followed as closely as feasible. Section 2 addresses the process of coding the data and includes a discussion of some methodological pitfalls. Section 3 deals with the analysis of the data. Two methods of relating word length are outlined that can be used in typological comparisons. Section 4 contains a discussion of the analysis that has been presented and section 5 denotes paths for further research.

1. Selecting Data Sources

The analysis of the relationship between token frequency and word length requires a rank-order frequency list of forms based on corpora of informal spoken data in the language in question. The number of occurrences should ideally be listed with each form because some of the statistical methods outlined herein call for number of occurrences rather than rank. A decision must be made regarding whether or not the transcriptions of the corpora should be orthographic so that only one form of a given word appears in the list, i.e. the Brazilian Portuguese [‘ta] and [‘se] would still be transcribed as *está* ‘copula, 3s, pres. indic.’ and *você* ‘you (fam.)’, respectively. The examples given show clear frequency effects in their abbreviation, but a decision must be made regarding whether to consider *tá* and *está* or *cê* and *você* as separate forms. In the case of the present study, forms known to be related to each other such as the relationship between *tá* and *está* were considered together as one unique form, rather than two separate forms.

Spoken data corpora are preferable because the relationship under study exists due to phonological reductions that occur in spoken language. Thus, it is of greater interest to study the forms and their corresponding token frequencies found in spoken language than in other genres. Likewise, informal spoken data is preferable because the more vernacular lexicon demonstrates the most susceptibility to the sound change that is being studied. Additionally, the corpora should preferably be of everyday conversation, rather than of interviews (e.g. sociolinguistic interviews) since interviews can affect the topics of conversation, thereby affecting the frequency counts. Everyday conversations also have the advantage that because they include two or more people, they naturally include a wider variety of speech styles in interaction. The corpora should be as large as possible and include as many different speakers as possible to approximate the variety of speech styles that exist in actual speech. The data from the *Corpus del español* are varied in terms of genre.

1.1. Considerations for a Comparative Study

For a comparative study, the corpora in each language must be tightly controlled. For example, in a comparison of Romance languages, if the French corpus contains a significant amount of conversation on political topics and the Spanish corpus does not, French could potentially appear to generally have longer words than Spanish. Thus, the results could be tainted by differences in topics in the respective corpora. It is therefore imperative to either control for topics or use very large corpora.

1.2. Dealing with Language Variation

The previous section dealt with considerations of a particular type of study that can be performed. One question that obviously arises is how to deal with phonological variation in the production of a word. To simplify the analysis into a manageable form, it may be advantageous to choose a language standard to model the phonetic representations. For the present study word length was determined based upon the number of syllables that each word has according to the standard pronunciation. In practice, speakers may truncate words, thereby reducing the number of syllables, or add epenthetic vowels, increasing the number of syllables. When there is more than one possible form, using a preestablished standard as a basis for assigning word length eliminates the need to determine how often each word was pronounced with a given number of syllables versus another number of syllables, information that may not even be available due to the lack of a sound file or a transcription method that would reflect such differences. For the corpus to be representative it needs to be reasonably large and the size of the corpus may prohibit the time-consuming task of a tight transcription of the corpus. This subject will be addressed in greater detail in section 1.4.

The author chose to base the analysis on standard forms in order to avoid additional complexities in an already complex analysis. The representation of the form is thus only as arbitrary as the language standard. In the case of the present corpus, this study is limited to the transcription methods that were used, which are not phonetic. Due to the nature of large corpora such as these, sources include a variety of corpora that employed different transcription standards, some of which account purely for the orthography and others of which orthographically denote alternate pronunciations as in the example of *está* and *tá* mentioned previously.

1.3. Ideals v. Practicalities

Ideally, the principal investigator would be in charge of gathering the corpus of each language being studied because the investigator would then have control over the above factors that can influence the results. Due to geographical and methodological constraints, this may or may not be possible. This is especially true in the case of a comparative study involving multiple languages where gathering the corpus data may be infeasible for one investigator or group of investigators to undertake. Another constraint is the time-consuming process of corpus transcription.

Such infeasibility should not discourage an investigator from attempting a study like the present one. The guidelines that I laid out above are ideals that an investigator should attempt to follow as closely as possible. Deviations will not entirely destroy the validity of the research, but they must be considered when the results are interpreted and reported. The corpus of Spanish used to develop this methodology, the oral data from Davies' *Corpus del español* (2005), is not ideal to perform the analysis because it violates many of the principles described above (formality, topic variation, etc.), however, the data provide a more than adequate means with which to test the statistical methods proposed by this study.

1.4. What Size Corpus is Enough? What Size List is Enough?

The size of the corpus depends upon the investigator's focus. If the investigator wants to treat homonyms as separate entries in the frequency list (e.g. Spanish *la* 'feminine article; feminine accusative pronoun'), the investigator may wish to opt for a smaller corpus or a tagged corpus. There are 202,783 tokens of *la* in the *Corpus del español*, so sifting through each token to sort out which are articles and which are pronouns would be impractical, especially considering that that corpus only allows contextual queries for forms that show 2000 or fewer tokens.

The size of list depends on the language or language family in question. Languages with larger lexicons probably merit longer frequency lists in investigations than do languages with smaller lexicons. In addition, languages that make heavy use of inflection probably also merit longer frequency lists because the same lexical items may appear in the list multiple times in their various inflected forms. The proposed number for Romance languages is 1000, the same number that is often used for English. Do note that Romance languages use more inflection than English, so a longer list might be appropriate. See Figure 1 for a plot of the number of syllables in the most frequent 1000 words in the *Corpus del español*. Interesting observations can be made for the entire 1000:

- Two-syllable words become possible after rank 20.
- Three-syllable words become possible after rank 50.
- Three-syllable words become favored over one-syllable words after rank 100.
- One-syllable words become extremely sporadic after rank 500 in favor of four-syllable words.
- Six-syllable words become possible only after rank 600.

Observation of Figure 1 reveals that one hundred words is really not an adequately sized corpus; with a rank-order list containing only one hundred words, three of the five observations above would not appear in the results. A much longer list would likely be necessary to find a sizeable number of seven-syllable words. The relationship being represented is exponential in nature and, therefore, as the rank-order list becomes larger, the economic law of diminishing returns applies.

2. Preparing and Coding the Data

After the rank-order list has been located, steps must be taken to assure the validity of the list and to remove undesirable elements. Once the final list has been defined, it may then be coded. The author used Microsoft Excel in preparing the test list from the *Corpus del español*. With such a large amount of data, Excel had some difficulties functioning correctly. It may be desirable in future studies to model the data in a relational database, such as Microsoft Access, which, although not an industrial database, is designed to handle a larger amount of data than Excel.

2.1. Preparing the Rank List(s)

In preparing the rank list, the author removed more than seventy items from the list for various reasons. Proper names were removed from the list since they are highly contextual and do not represent the speech of a generic community of speakers. A person would probably talk about *España* more if he or she is actually living in Spain, but a person from Cuba probably would talk about it considerably less. In addition, acronyms like *PRI* (Partido Revolucionario Institucional) were removed because they do not belong to generic speech communities, generally speaking, are relatively short-lived in the language, and do not suffer the same phonetic reductions that other words do. *PAN* created some difficulties since not only does it mean 'bread', it is also the acronym for another Mexican political party.

In some cases, forms might be separated and represented twice in the list, for example, *tá* and *está* 'copula'. If the decision has been made, as is the case with the test data, to count *tá* and *está* together, the two forms must be combined into one entry with the token frequency recalculated. As previously mentioned, in the case of the present study, the two representations in the list were combined into one entry. It goes without saying that a detailed record must be kept for all changes that are made to the list.

2.2. Methods of Measuring Length

There are three main methods to measure word length: by syllables, by graphemes, and by phonological segments. The second is problematic because graphemes often do not reflect the phonological structure of a word, or at the very least reflect it poorly. Of the two remaining options, the number of syllables is certainly the easiest to count, but it does not offer a detailed account of the phonological structure of words. On the other hand, while a count of the phonological segments would shed more light on the amount of phonological reduction that occurs in a given language, determining how to weight each segment can prove an interesting challenge to the investigator despite the fact that there is probably no universally accepted solution. A simple count of the number of segments does not reflect reductions that may occur to individual segments and the relative importance of different types of segments. At the same time, assigning weights to different types of segments is difficult to do empirically and the assigned weights could prove to be controversial.

3. Data Analysis

The equation fitting techniques described herein provide an effective means for determining the overall word length-token frequency relationship in a language. In addition, they allow for comparisons of the typological characteristics of two or more languages. Fitting an equation to the data requires the performance of a regression analysis. Since the relations that are being studied are exponential, it would be possible to use a nonlinear regression analysis to generate an equation. However, there is a much easier method of generating an equation. By taking the natural log of the either rank or token frequency it is possible to create a linear relation where none existed before, thereby permitting a linear regression analysis. The linear regression analysis is a much simpler technique for finding the equation than the nonlinear regression. The linear regression will generate an equation similar to the following:

$$y = mx + b$$

where x is the independent variable, y is the dependent variable, m is the slope of the line, and b is the y-intercept. There are two primary frequency-based factors from which an equation can be generated. The superiority or inferiority of the two different methods will be discussed in sections 3.3 and 3.4.

3.1. Length as a Function of Rank

Length can be expressed as a function of the rank of a form in a rank-order list. The linear regression analysis must be performed with form length as the y-value and the x-value must be the natural log of the rank. The calculations required in the regression analysis were performed in Microsoft Excel. For a simple guide on how to perform a regression analysis using Microsoft Excel, see Odom (2000).

Using the regression analysis described above on the 1000 most frequent words from the *Corpus del español*, we arrive at the following relation between rank and length:

(1) $L_{\acute{o}} = 0.42049 \ln r - 0.01097$

where $L_{\acute{o}}$ is the length of the form as measured in syllables and r is an integer representing the rank of the form in the rank-order list. See Figure 2 for a plot of the actual data with the points predicted by the equation. As you can see, the line follows the general pattern of the data fairly closely. The line is not going to predict the six-syllable words ranked in the six hundreds because the line represents the theoretical average of what the lengths of words are near a given rank. If this is so, about half of the data points should be above their respective point on the line generated by the equation and half should be below. This will be tested in section 3.4.

3.2. Length as a Function of Token Frequency

An alternative method of modeling form length is as a function of token frequency; that is, the length of the word is best predicted by the number of times that it appears in the corpus. Although on the surface this may appear to be a superficial distinction, it is significant. The reasons why one might wish to use one method or another will be discussed in section 3.4. The linear regression analysis must be performed with form length being the dependent y-value and the natural log of token frequency per million words being the independent x-value. Recording the frequency in terms of tokens per million regularizes the data to allow comparisons between two or more corpora of different sizes.

Using the regression analysis described above on the data we arrive at the following relation between token frequency and word length:

(2) $L_{ó} = -0.38824 \ln t + 4.59446$

where $L_{ó}$ is the length of the form as measured in syllables and *t* represents the number of tokens of the form per million words in corpus data. See Figure 3 for the plot of actual data with the points predicted by the equation with the actual frequencies from the data given as an argument for *t*. As you can see, this equation (2) predicts the data points very closely to what equation (1) predicts. The accuracy of each equation will be addressed in section 3.3. This line can be taken to predict the average length of words of a given token frequency.

3.3. Testing the Validity of the Equations

Figure 4 is a plot of the two equations in comparison with the actual data points. The two lines predicted by the equations are very close together with the exception of the first dozen items in the rank-order. Equation (1) predicts very low values in those first items. Aside from that, the results are approximately equivalent.

The averages of the actual data points and the points predicted by the two equations were calculated to all be 2.475. To detect residual differences, the differences between the actual data's average and each of the equations' averages was calculated to be zero. The difference between the average of the values predicted by equation (1) and the average of the values predicted by equation (2) was -3.9968×10^{-15}, which is a result so small that it may be due to truncation by the computer. In other words, the equations seem to have passed that test.

The two lines themselves have been determined to be approximately equivalent, although there are some differences that will be addressed in section 3.4. Figure 5 shows the residuals from equation (1). The *x*-axis represents the line predicted by equation (1). The points on the figure are the actual data with respect to the line. The data represent non-constant variance since the data points fan out and are more concentrated at the higher levels of frequency than the lower levels of frequency where the points are dispersed.

3.4. The Best Fit

Although the two equations that were generated in 3.1 and 3.2 were determined to be approximately equivalent for the data in terms of their calculations, one of the equations correlates better with what it is attempting to model. In doing the linear regression the investigator can calculate the *r*-value, which is a measure of how closely correlated a set of data plots are. *r* may be between –1 and 1. Values approaching 1 are considered to have a close positive correlation and values approaching –1 are considered to have a close negative correlation. Values near zero are less well correlated. The correlation between rank and syllable length has r = 0.43575. The correlation between token frequency and syllable length has r = -0.43648. The difference between the absolute values of the rank/syllable length correlation and the token frequency/syllable length correlation is -0.00073, which indicates that the token frequency correlation coefficient is slightly closer to –1 than the rank correlation coefficient is to 1. Thus, syllable length correlates slightly better to token frequency than to rank. Again, the difference is minor, but a close examination of the first twenty data points will reveal that the equation based upon token frequency more effectively predicts the structure of the data, as none of its values are less than zero.

The r-values for both relations are quite low, but this is due to the fact that human language simply is not that organized. In order for r to equal one, the actual data plots would have to follow the line exactly. Obviously, we do not have words that are 2.5 syllables in length. In addition, the closest a language could come to a perfect correlation would be if all one-syllable words came together, then all two-syllable words, all three-syllable words, and so forth. Such a convenient system is highly improbable in a natural language. The fact that the two relations' correlation coefficients are close to ±0.5, and not zero, suggests that there is a relation of some sort.

Of course, with such small differences between the two correlation coefficients, the technique used will ultimately depend upon what the investigator wishes to measure. For example, if the number of tokens is unavailable, it may only be possible to use rank. The equation involving token frequency might be useful in performing some sort of cross-linguistic comparison of the behavior of the exemplar model and internal representations of structures and their impact on the structure of words. On the other hand, if the investigator is strictly interested in the relationship with rank the rank-based equation might be more appropriate. The point is that the technique should be chosen based upon what is being studied.

4. Summary

The present paper has presented a brief history on research related to measuring frequency effects on the phonology of a language. The specific focus of this research is the relationship between word length and token frequency. One has to be very selective about the data source in order to perform a study such as the present one. Since the phonological reduction process occurs first in informal conversation where speech is largely unmonitored, it is important to use that type of corpus data. More formal contexts tend to yield a different distribution of lexical items, lexical items that may or may not have suffered the reductive sound changes that are presumed to have caused this relationship.

The process of preparing the word list and coding it was discussed along with some problematic issues that may be encountered. Naturally, there are no simple answers. The process for fitting a line to the data was elaborated. Syllable length was shown to be relatable to both rank and token frequency. While rank and frequency are obviously correlated, that correlation is not perfect in natural languages. As such, the choice of the dependent variable (e.g. token frequency v. rank) has a critical impact on the results of the study.

This methodology may be used to test certain statistical properties relating word length to token frequency or rank. For example, a means of testing the formation of these equations based upon the first one thousand words is whether the equations accurately predict the behavior of the next thousand words on the rank-order list. In general, one would expect the relationships to hold fairly closely for the next thousand words. If the relationship does not hold, a new hypothesis must be formulated. A related question is whether one of the relationships (rank-based, token frequency-based) more accurately predicts the next one thousand words on the list than the other relationship.

The equations presented here are statistical techniques that should be applied to data that conform more closely to the parameters defined in this paper, preferably in a comparative context. Such a study would allow the relationship between word length and frequency to be analyzed cross-linguistically. Of particular interest would be a comparison between languages that differ radically from a typological standpoint. For example, how would frequency effects on word length differ between the highly analytical Sino-Tibetan languages and the agglutinative Uto-Aztecan languages? How would different languages of the same language group behave?

Another possible application of the technique would be to see if rank or frequency of occurrence is more important. For example, if Language A's most frequent word occurs 50,000 times per million words and Language B's most frequent word occurs 40,000 times per million, does that difference make any difference to how reduced phonologically it is or any other type of difference?

Finally, this methodology requires the use of language standards to make an enormous project more manageable. While this methodology undoubtedly will open the doors to research that will advance our knowledge of the statistical relationship between word length and rank or frequency, techniques must be developed to incorporate a more usage-based framework that takes into account the variation that exists in human language.

Notes

[1] The author would like to thank Joan Bybee for her comments on a draft of this paper and Zibonele Valdez-Jasso for suggesting the usage of natural logs to make a linear regression analysis possible. The author assumes responsibility for remaining errors.

Works Cited

Bybee, J. (2002). Lexical diffusion in regular sound change. In D. Restle & D. Zaefferer (Eds.), *Sounds and Systems. Studies in Structure and Change. A festschrift for Theo Vennemann* (pp. 59-74). Trends in Linguistics. Studies and Monographs. Berlin: Mouton de Gruyter.

Davies, M. (2005). *Corpus del español*. Retrieved February 20, 2005, from http://www.corpusdelespanol.org

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 431-48). Amsterdam: John Benjamins.

Kaeding, F. W. (1898). *Häufigkeitswörterbuch der Deutschen Sprache*. Berlin: Mittler.

Leopold, E. (1998). Frequency spectra within word-length classes. *Journal of Quantitative Linguistics, 5*, 224-31.

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics, 5*, 240-45.

Odom, C. (2000). Excel tutorial on linear regression. Physics Laboratory. Clemson: Clemson University. Retrieved April, 15 2005, from http://phoenix.phys.clemson.edu/tutorials/excel/regression.html

Zipf, G. K. (1936). *The psycho-biology of language. An introduction to dynamic philology.* London: Routledge.

Figure 1. Plot of the number of syllables in a form as a function of rank. The data comes from Davies' *Corpus del español* (2005).

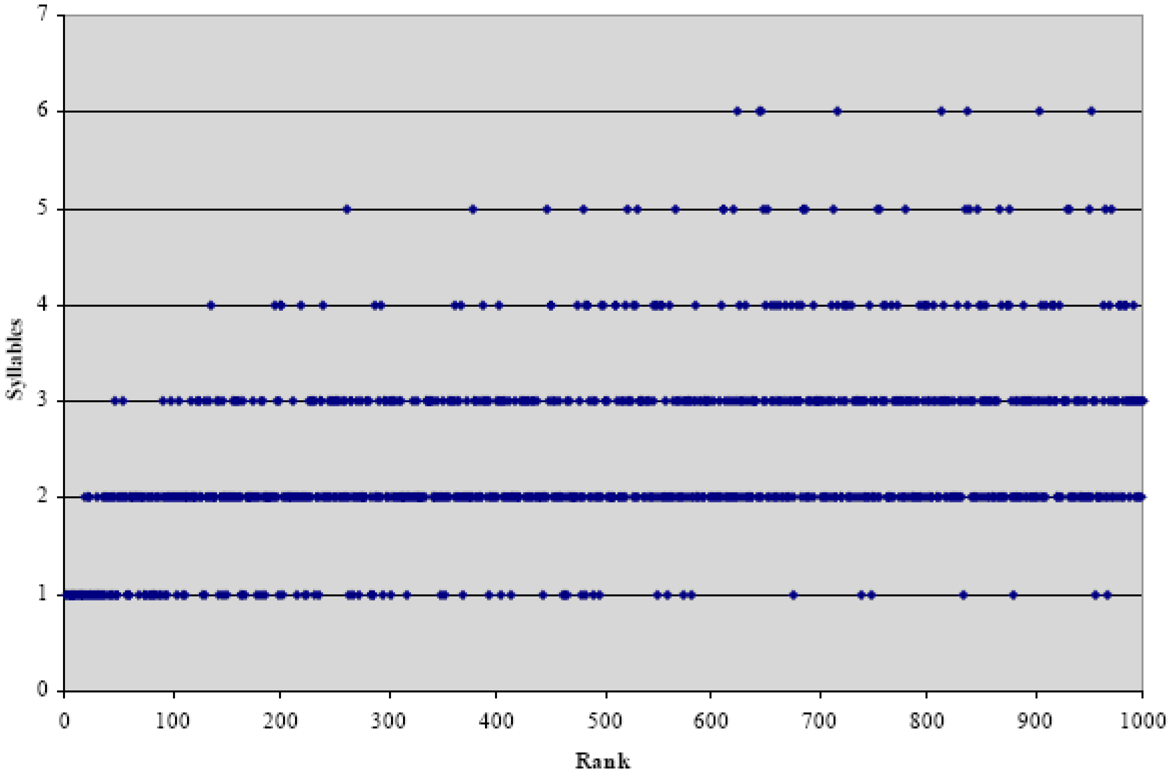
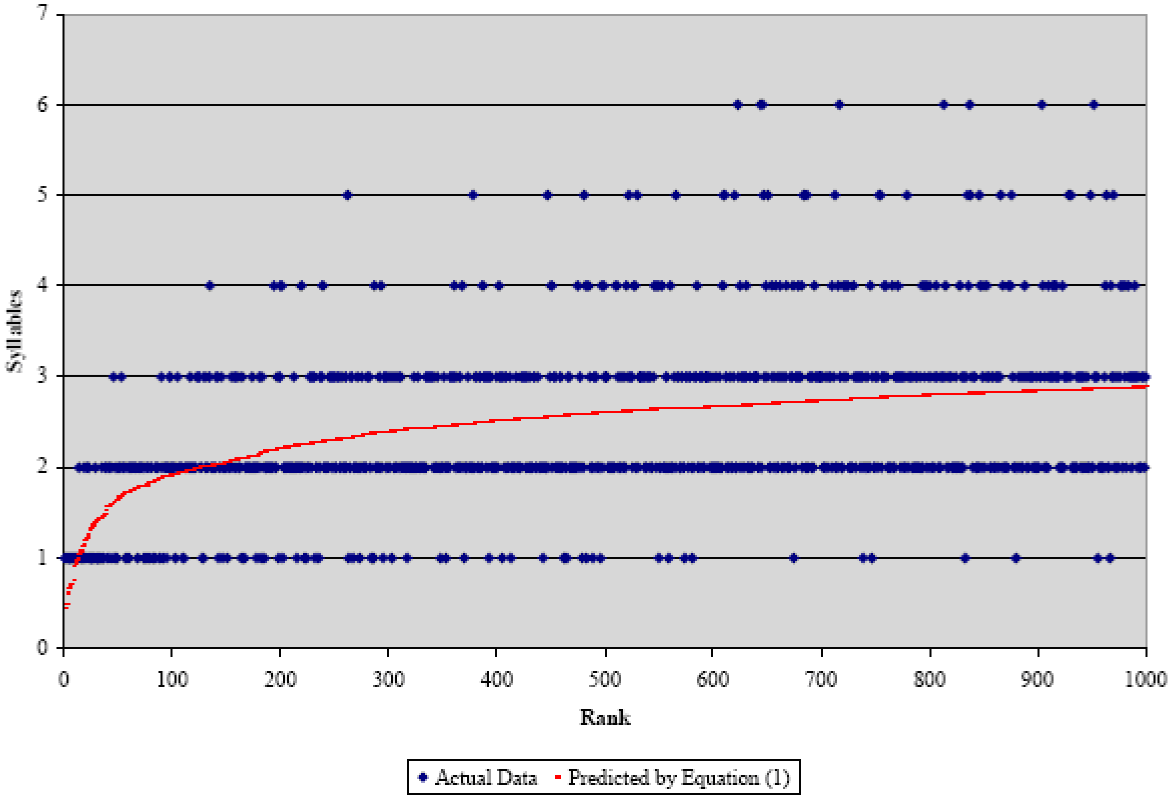


Figure 2. Plot of the actual data points against the data points predicted by equation (1).


Syllables
Rank
◆ Actual Data
- Predicted by Equation (1)