A Quantitative Analysis of Word-Definition in a Machine-Readable Dictionary

Robert W.P. Luk and Venus M.K. Chan
Department of Chinese, Translation and Linguistics, City University of Hong Kong

rwpluk@ctspc05.cityu.edu.hk

Abstract

This paper investigates some of the distributional properties of word definitions in a machine-readable dictionary which was obtained from the Oxford Text Archive. Three types of distributions were examined: (1) frequency-ranked distribution of words in the definitions, (2) the length distribution of word definitions and (3) the frequency distribution of the number of unique tags of an entry. In addition, the coverage characteristics of headwords over word definitions are also explored. A rough-and-ready comparison of distributional properties between tokens and their morphologically decomposed ones are made. Our result shows that morphological dcomposition does not change the length distribution of the word definitions nor the ranked-frequency distribution of words, significantly. However, it increases the coverage of word definitions dramatically compared with no decompositions. Furthermore, the frequency distribution of the number of unique tags per entry is approximately linear when the data is suitablely scaled (i.e. linear or logarithmic).

1 Introduction

Many (English) dictionaries (e.g. Collins [1], OALD [2], Longman [3]) have been made available online but they have rarely been used as data in quantitative analysis. Preliminary analysis [4] such as compiling word usage frequency in word definitions, thereby compiling conceptual primitives has been carried out. A notable exception is the mathematical modeling of word length distribution [5]. In this paper, we focus on a quantitative analysis of word definitions. In particular, we examined:

1. the ranked-frequency distribution of words in the definitions;
2. the length distribution of word definitions;
3. the frequency distribution of the number of unique tags in word definitions;
4. the coverage statistics of words in dictionary definitions.

These quantitative descriptions facilitate a more objective comparison between different dictionaries than merely the number of entries in a dictionary. Ultimately, they can serve as part of a more comprehensive evaluation metrics for dictionaries. For example, word length distribution may indicate whether definitions are given comparatively concisely or brief. Unusual distribution of the number of unique tags might indicate bias or undersampling of the data which might be intended. Poor coverage characteristics might be due to less adherence in using a control vocabulary.

Furthermore, these quantitative descriptions can be used to guide the construction of computational lexicon. For example, coverage statistics can be used to determine the size of the control vocabulary in analyzing word definitions as in [6]. Apart from applications, quantitative analysis provides exploratory data for quantitative modeling, giving a more comprehensive description of the phenomenon at hand. The quantitative model of word length distribution [5] is a good example.

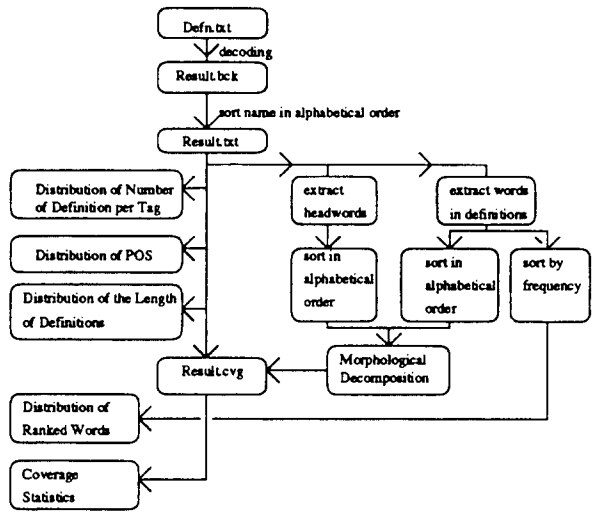

Figure 1: Schematic diagram of compiling statistics.

2 Preliminary Processing

We have obtained the Collins dictionary from the Oxford Text Archive. The dictionary is in Prolog fact base format and 172,265 word definitions have been extracted, out of 81,172 headwords. There were about 77,000 different words in the definitions, of which only about 30,000 are headwords (exact matching). Many of the non-headwords can be morphologically derived from the headwords (about 40% using a rough-and-ready affix stripping algorithm). About 3,233 non-headwords have special symbols. These so-called words are chemical/mathematical formulae, phonetic symbols, special markers and alternation short hand. There are also several words (e.g. efficacy) missing from the electronic version of the dictionary, which we have manually added. We have also checked word definitions of length less than 7 characters. A total of 880 definitions were found where corrections are made due to errors in decoding the dictionary. In addition, we have also examined alternate entries (e.g. *a* or *A* as a headword) which are separated into individual entries.

Figure 1 is the schematic diagram to compile the statistics for quantitative analysis. After decoding, the definitions are sorted according to the corresponding headwords, alphabetically. Three types of distributions were compiled from the result: the number of definitions per headword, the number of unique part-of-speech/semantic tag per headword and the length of the word definitions. Headwords and words in the definitions are also extracted from the result. For morphological decomposition, the headwords are treated as stems and the extracted words are the stems, or derived words or inflected words. The decomposition is used to estimate the coverage statistics of stems over word definitions in the dictionary. The extracted words and their corresponding frequency are used to estimate the frequency-ranked distributions of words.

Morphological decomposition is carried out by a simple affix stripping algorithm. The affixes were derived from the machine-readable Collins dictionary. The algorithm tries to determine whether the input word is a headword (including compounds like hard-working). If the exact matching with headwords fail, the algorithm converts all upper-case characters into lower-case. Otherwise, the algorithm repeatedly strip off suffixes. At each iteration, the longest matched suffix is stripped, possibly with spelling adjustment (e.g. adding *e* after *es*), and the remaining word is matched with the headwords in the dictionary. If a headword is matched with the remaining word, the iteration terminates. The number of iterations is limited at most 3. If there is still no headword matched, prefix stripping is applied to the original input word. If no headword matches, both prefix and suffix stripping are applied.

3 Frequency-Ranked Distribution

Zipf [7] has investigated that the logarithmic frequency distribution of words ranked by their occurrence frequency is linear according to the following empirical equation:

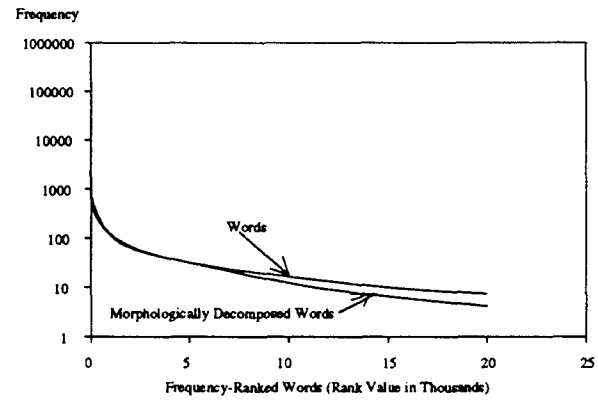

Figure 2: Frequency-ranked distribution of words and their morphologically decomposed ones (so called stems). In both cases, the curve decreases dramatically for the first 2,500 (decomposed) words and subsequently a less rapid but almost linear decline in frequency occurrence with increasing rank value.

$$\log F_i + \log R_i = C \tag{1}$$

where F_i is the occurrence frequency of the i^{th} word, R_i is the rank of the i^{th} word and C is a constant. This empirical relationship was applicable to a large volume of free text (e.g. novels and articles) where there is no explicit control over vocabulary growth. On the other hand, many dictionaries (e.g. Longman [4]) limit the amount of vocabulary used in word definitions. An interesting question is to examine whether the frequency-ranked distribution of words in definitions follow the above empirical relationship as in free text.

Figure 2 shows the frequency-ranked distributions of words in the definitions. Words are ranked with decreasing occurrence frequency. The frequency-ranked distribution of both words and stems (i.e. words after affix stripping) are not linear and both distributions are similar to each other, except with large rank values (say $> 10,000$). Both curves can be divided into two parts: non-linear and linear. The first part begins from rank 1 to about 2,500 where many of these words should belong to the control vocabulary. The second part begins from about 2,500 to the end where words are used more or less similar to words used in free text, since most of these words are not in the control vocabulary. The effect of morphological decomposition is not apparant from the figure. However, the frequency-ranked distribution of stems is initially higher than that of words and at around 1,500 the two distributions cross over. Afterwards, the distribution of stems is lower than that of words. This indicates that the frequently used words are usually morphologically derived or inflected. Although differences in frequency of frequently occurring words and stems are large, the differenes are not apparant in figure 2 because the vertical scale is logarithmic. Likewise, although there seems to be apparant differences in frequency for infrequent occurring words and stems, the actual difference in frequency is no more than 4, due to the logarithmic vertical scale.

4 Length Distribution of Definitions

Figure 3 shows the word-length distribution of definitions. The distribution is neither normal nor unimodal. The distributions for both words and stems are almost identical, indicating the reduction in length due to morphological decomposition is not concentrated at a particular length of the distribution. Two prominent peaks occur at about 10 and 20 characters. These two peaks might represent two methods of defining words. The first peak might represent the use of synonyms to substitute for the headwords. The second method might be defining words as explanations in short sentences. Very long word definitions (say over 50 characters) are usually introduction of places, people or historical events. This substantiates the claim of the dictionary that it contains entries for places, people and historical events with encyclopedic descriptions.

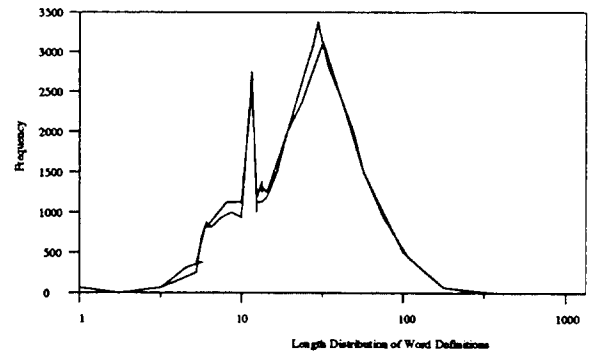

Figure 3: The distribution of word length of word definitions in terms of the number of characters. Note that the x-axis is in logarithmic scale because there is one definition which has over 1000 characters.

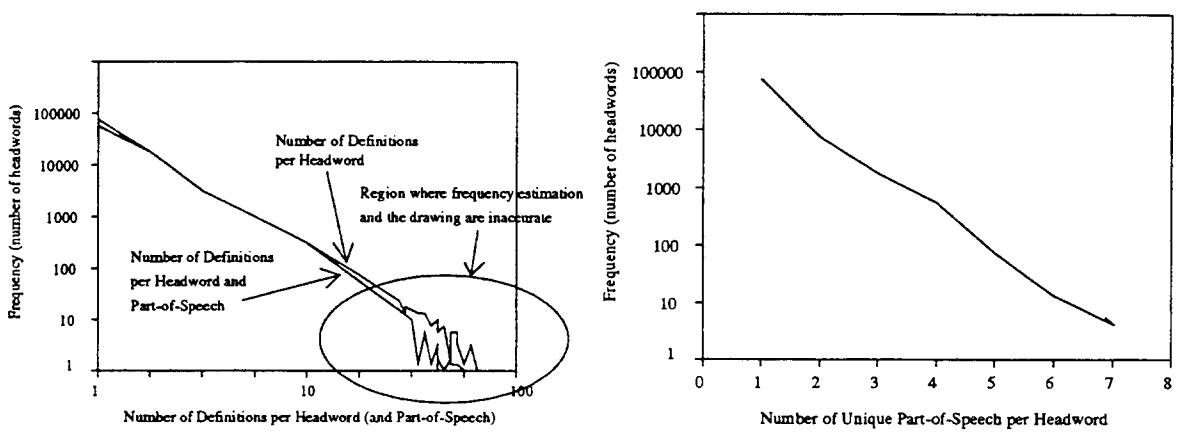

Figure 4: (Left) Frequency distribution of the number of definitions per headword and per headword and part-of-speech. At the low frequency region (i.e. on the right), the frequency estimation is not accurate and therefore the frequency distribution curves have many gitters instead of a smooth curve. Note that the x-axis is in logarithmic scale.(Right) Frequency distribution of the number of unique part-of-speech tags per headword. The curve roughly declines linearly.

5 Number-of-Tag Distribution

Each headword has a part-of-speech (POS) tag as well as a semantic tag. Figure 4 (left and Right) shows the distributions of the number of unique semantic and POS tags per headword, respectively. Both distributions roughly follows a straight line in a vertical logarithmic scale. In both cases, there are few headwords with a large number of unique tags where as many headwords have very few number of unique tags. Such a relationship is expected; otherwise, a lot of effort is needed to determine the particular meaning of every word in natural text and speech. We do not know whether the linear relationship between frequency and the number of unique tag of a headword is incidental or not. Note that the horizontal scale is logarithmic for semantic tags but linear for POS.

Since the definitions can be considered to be organized as a tree of headwords with different POS tags and then with different semantic tags, the number of unique semantic tags per headword can differ substantially from that of per headword and POS tag. Figure 4 (Left) demonstrated that the distribution per headword, and per headword and POS tag is very similar, except at large number of tags per headword and per headword and POS because estimates at low frequencies are not reliable.

We have examined whether there is any correlation between word usage frequency and the number of unique tags of the corresponding word. We plotted a scatter diagram of the first 500 most frequently occurring words in the Brown corpus [8] against the number of unique tags of those words. There are no recognizable linear relationship. The pattern we observed is the clustering of points near the origin.

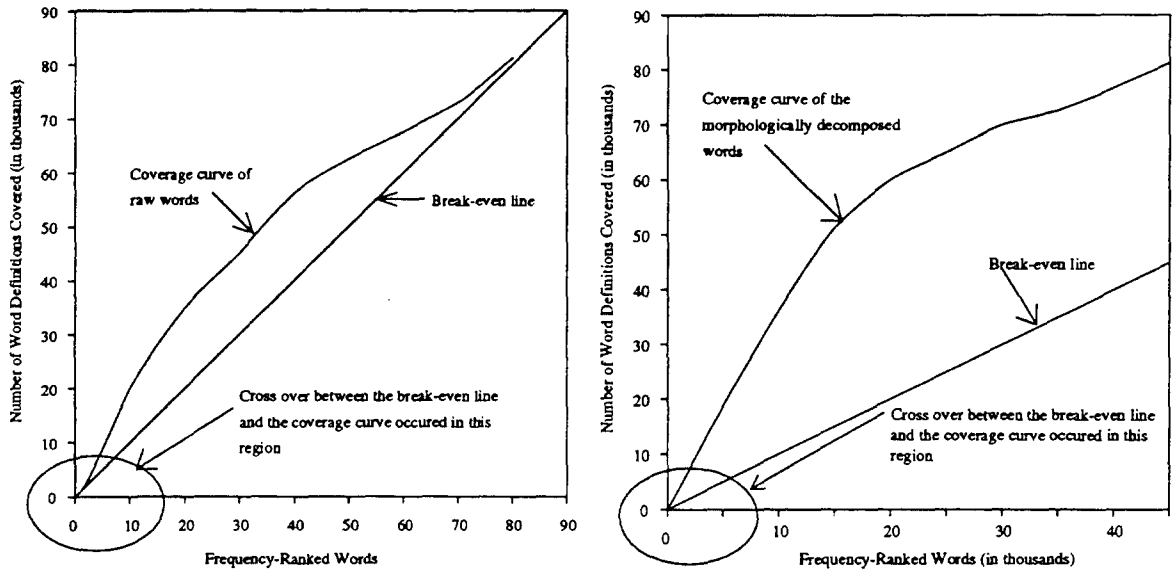

Figure 5: (Left) Coverage curve of words in the definitions. Initially, the coverage curve is below the break-even line. At $x \geq 920$, the coverage curve is above the break-even line. (Right) Coverage curve of morphologically decomposed words (so called stems) in the definitions. Again, the curve is below the break-even line until $x = 340$. Afterwards, the curve is above the break-even line.

6 FREQUENCY-RANKED COVERAGE DISTRIBUTION

The coverage of word definitions by a set of word, Σ, can be defined as the number of different words that are defined by Σ. The coverage expressed as a percentage indicates the utility of words in Σ for word definitions. The higher the percentage implies the higher the utility. Since coverage is defined in terms of a set of words, it is measured by generating subsets of words found in a defined set, Σ. The number of possible subset of words generated by Σ is $2^{|\Sigma|}$ (including the null set), which is very large (say $|\Sigma| = 70,000$) to compute. Instead, we can impose an ordering in generating subsets of words to reduce the computational load and at the same time to yield a meaningful coverage statistics based on a related ordering relationship. We propose to rank words according to a decreasing order of occurrence frequency in the word definitions (as in [6]). The first ranked word is the first subset of words generated and the percentage of coverage is determined. The next subset of words is generated by adding the word of the following rank value and the percentage of coverage is measured. This process iterates for all the words in Σ. Thus, the number of iteration (i.e. subset of words) is $|\Sigma|$ instead of $2^{|\Sigma|}$.

Figure 5 shows the amount of coverage for different words (on the left) and stems (on the right) of the definitions, respectively. The word coverage follows a tilted Mexican-hat curve and the stem coverage follows an S-curve. The break-even line represents the amount of words used for definitions that is equal to the number of different words defined. For word coverage, the rate of increase is high initially and the coverage falls off toward the break-even line with subsets of size larger then 40,000. For stem coverage, the coverage curve increases steeper and sustains a higher coverage than the break-even line than word coverage. The curve does not approach the break-even line and it holds a lead of about 40% of the word definitions over the break-even line since 40% of the words are found to have identical stems of other words. Thus, the effect of morphological decomposition raises the rate and the amount of coverage substantially where as the decomposition has little effect on rank-frequency and word-length distribution. This implies that knowing how to identify stems from words is important for using the Collins dictionary, effectively.

We have also plotted the percentage of coverage of the first 1500 ranked words. Both word and stem coverage curves were initially below the break-even line but after the first 920 and 340 words, respectively, both curves were above the break-even line. Thus, the stem coverage curve crosses over the break-even line earlier than the word coverage curve.

7 Discussion

The quantitative analysis of word definitions has opened a number of issues. First, the frequency-ranked distribution of words in the definitions do not strictly follow the Zipf empirical relationship although it is possible to consider that low frequency words behave as if they were in free text. The distribution can demonstrate the degree with which words in the control vocabulary are used. The larger the amount of non-linearity shows the more deviation from words used in free text. The second issue is that the word length distribution can be used to indicate different methods of defining words. Third, the frequency distributions of the number of unique tag were linear and it would be interesting to examine whether this linear relationship holds for other dictionaries or tagged corpora, and whether there is any underlying model of tag generation and usage (as in hidden Markov model [9]). In this case, the number of unique POS tag distribution can be used to verify whether POS tagging by hidden Markov model is consistent with the data. Fourth, the coverage statistics show that morphological decomposition is important in effectively interpreting words in the definitions where such importance was not apparent in the rank-frequency or word-length distributions. Finally, the coverage statistics can be generalized to examine the coverage characteristics of other word sets (e.g. words in school text or words in other dictionaries).

Acknowledgement

This research is supported by City University Small-Scale Research Project Grant #903249. We are grateful to the Oxford Text Archive for supporting a free access to the Collins dictionary and the Collins Publisher.

Reference

[1] COLLINS COBUILD ENGLISH LANGUAGE DCITIONARY, Collins (COBUILD), London & Glasgow, 1987.

[2] OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH, 4TH ED, Oxford University Press (OALD), London, 1989.

[3] LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH, NEW EDITION, Longman (LDOCE), Harlow, 1987.

[4] SMITH, R.N., "Conceptual primitives in the English lexicon", in W. FRAWLY AND R. STEINER (eds) *Advances in Lexicography*, pages 99-137, 1985.

[5] WIMMER, G., R. KOHLER, R. GROTJAHN AND G. ALTMANN, "Towards a theory of word length distribution", *Journal of Quantitative Linguistics*, 1:1, 98-106, 1994.

[6] WILKS, Y., D. FASS, C-M. GUO, J.E. DONALD, T. PLATE AND B.N. SLATOR "A tractable machine dictionary as a resource for computational semantics", *Technical Report MCCS-87-105*, Computing Research Laboratory, New Mexican State University, Las Cruces, 1987.

[7] ZIPF, G.K., *Human Behavior and The Principle of Least Effort*, Addison-Wesley, Reading, M.A, 1949.

[8] KUCREA, H. AND W.N. FRANCIS, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967.

[9] CHURCH, K., "A stochastic parts program and noun phrase parser for unrestricted text", *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136-143, Austin, Texas, 1988.