Truncation of Content Terms for Turkish

Hayri Sever
Department of Computer Engineering
Başkent University
06530 Bağlıca, Ankara, TR
sever@baskent.edu.tr

Yaşar Tonta
Department of Information Management
Hacettepe University
06532 Beytepe, Ankara, TR
tonta@hacettepe.edu.tr

Abstract

Stemming, truncating, suffix stripping and decompounding algorithms used in information retrieval (IR) to reduce the content terms to their respective conflated forms are well-known algorithms for their causes for improving the retrieval performance as well as providing space and processing efficiency. In this paper we investigate the statistical characteristics of the truncated terms for Turkish on a text corpus consisting of more than 50 million words and attempt to measure the vocabulary growth rates for both the whole and truncated words. Findings indicate that the truncated words in Turkish exhibit a Zipfian behavior and that the whole words can successfully be truncated to the average word length (6.2 characters) without compromising performance effectiveness. The vocabulary growth rate for truncated words is about one third of that for the whole words. The result of our study is two fold. First it surely opens the room for truncation of content terms for Turkish for which there is no publicly available stemming code equipped with morphological analysis capability. Second, use of a truncation algorithm for indexing Turkish text may yield comparable effectiveness values with that of a stemming algorithm and hence, the need for stemming may become absolote, given that morphological analyzers for Turkish is highly complex in nature.

Introduction

Exponential growth of processing and storage capacities of computers, coupled with networking technologies, made it possible to undertake large scale information retrieval (IR) applications that were heretofore inconceivable. Applications of computational linguistics and the statistical modelling of various languages by means of stemming, suffix stripping and decompounding algorithms are among them (Porter, 1980; Frakes, 1992; Harman, 1991). The effectiveness of such algorithms have been investigated for a number of languages including Spanish, German, Greek, Slovene, Malay, and Persian (Figuerola, Gomez, Rodriguez & Berrocal, 2002; Braschler & Ripplinger, 2004; Kalamboukis, 1995; Popovic & Willett, 1992; Ahmad, Yusoff & Sembok, 1996; Abu Bakar & Rahman, 2003; Tashakori, Meybodi & Oroumchian, 2002). The distributions of words and phrases in different languages have also been studied with respect to their comformance to, say, Zipf and Mandelbrot laws (Fairthorne, 2005; Ha, Sicilia-Garcia, Ming & Smith, 2002).

In parallel with these developments, a spelling checker for Turkish was introduced in early 1990s (Solak & Oflazer, 1993). The effects of various stemming algorithms, conflation and *n*-gram matching techniques on Turkish text retrieval have been studied by a number of researchers (Solak & Can, 1994; Ekmekçioğlu, Lynch & Willett, 1995, 1996; Duran & Sever, 1996; Ekmekçioğlu, Lynch, Robertson, Sembok & Willett, 1996; Ekmekçioğlu & Willett, 2000; Sever & Bitirim, 2003; Dinçer & Karaoğlan, 2003). Words appearing in various Turkish corpora have been been analyzed to find out the vocabulary growth rate and to see if they fit the "power law" distributions such as Zipf (Dalkılıç & Çebi, 2002, 2004). The application of the natural language processing (NLP) techniques to the Turkish language produced promising results (Pembe & Say, 2004). More generally, the effectiveness of Turkish search engines in terms of precision, recall, novelty and recency ratios has been tested along with the application of a metadata scheme as an information discovery and retrieval tool on the Internet (Bitirim, Tonta & Sever, 2002; Küçük, Olgun & Sever, 2000).

A member of the south-western or Oghuz group of the Turkic family of languages, Turkish is an agglutinative language with word structures formed by productive affixations of derivational and inflectional suffixes to the root words. Some 25,000-30,000 stems are actively used in Turkish. Yet, the number can go as high as a couple of millions when the inflections of the words are included. The "index of synthesis"[1] for the Turkish language is 2.86. The space efficiency and effectiveness of an IR system decreases substantially when indexing is based on content terms without stemming (Sever & Bitirim, 2003). More specifically, it is experimentally shown that the retrieval effectiveness of a typical vector-based IR system increased by about 22% in terms of normalized precision. Early Turkish search engines had had many problems: the limited coverage of the Turkish content, low precision, novelty and recency ratios; and the lack of stemming are among them (Bitirim, Tonta & Sever, 2002). Well-known search engines such as Google and Yahoo! do not have Turkish stemming capabilities, either.

Turkish is spoken natively by over 200 million speakers and a member of Turkic languages. It is also well-known fact that Turkish is one of the languages studied lesser than others, although it ranks fifth among world languages with respect to the number of native speakers. In this paper, we investigate the statistical nature of truncated terms for Turkish, The result of our study is two fold. First it surely opens the room for truncation of content terms for Turkish for which there is no publicly available stemming code equipped with morphological analysis capability. Second, use of a truncation algorithm for indexing Turkish text may yield

[1] The index of synthesis refers to the average number of affixations (morphemes) per word in a language.

comparable effectiveness values with that of a stemming algorithm and hence, the need for stemming may become absolote, given that morphological analyzers for Turkish is highly complex in nature.

In English, it was reported that indexing by truncation of each word to four or five characters yielded almost as good a discrimination between relevant and nonrelevant documents as did a system that used full terms (Salton, 1988: pg.246). We test the conjecture that truncation of Turkish words around average word length gives a similar Zipf behavior with one to another, and furthermore, we show that validity of this conjecture can be extended to whole words as well. Upon proving such a conjecture, we claim that choosing either of truncation schemes involving in taking first five, six, or seven characters for indexing is as good as choosing whole words. Validity of this claim provides us considerable space efficiency for indexing systems without loosing effectiveness.

.

In IR systems it is important to select the proper content terms for indexing. A correct practice for extracting content terms would be to use an auxiliary stop list or the frequency information. The essence of the frequency information for indexing is that any content term that is likely to be assigned as an index term should appear in relevant documents neither too frequently nor too sparsely in order to reduce the density of the document space. In this paper we also provide the constants for the growth rate to determine the approximate number of distinct words for a given text repository. More generally, we address how to determine the content terms that appear in relevant documents with medium frequencies in a Turkish text corpus, thereby providing a rapid and feasible text retrieval environment for Turkish.

Background

Zipf's Law

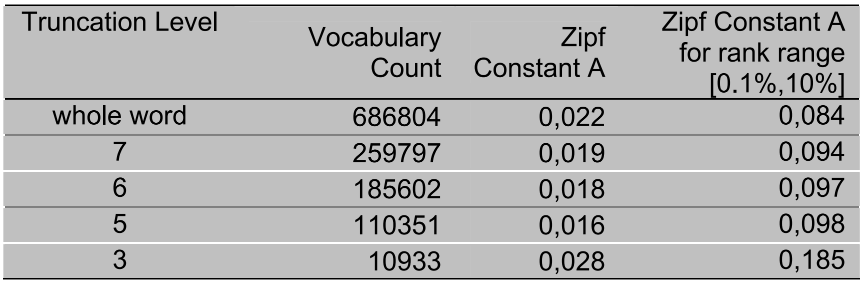
Many human-made and naturally occurring phenomena, including city sizes, incomes, word frequencies, and earthquake magnitudes, are distributed according to a *power-law* distribution. A *power-law* implies that small occurrences are extremely common, whereas large instances are extremely rare. This regularity or "law" is well known as the *Zipf Law*. The "Zipf's Law is the observation that frequency of occurrence of some event (*P*), as a function of the rank (*i*) when the rank is determined by the above frequency of occurrence, is a power-law function $P_i \sim 1/i^a$ with the exponent *a* close to unity (1) (Li, 2005). In a power-law we have $y = C x^{-a}$, which can be manipulated as $log(y) = log(C) - a\ log(x)$. So a power-law with exponent "*a"* is seen as nearly inversely linear with slope "*-a"* on a log-log plot. Zipf's Law based on *the principle of least effort* can be used to approximately model the human behavior, particularly on the use of a natural language.

Vocabulary Growth

The term "vocabulary" can be defined as the set of distinct words in a corpus. The *Heaps' Law* gives a general formula for determining the vocabulary growth for a corpus as a function of the text size. "It establishes that a text of *n* words has a vocabulary of size $V= K n^b$ where $0 < b < 1$" and the constants *K* and *b* depend on the particular text (Li, 2005). In natural language text, the *Heaps' Law* is used to predict the growth of the vocabulary size. The *Heaps' Law* law states that the vocabulary of a text of size, *V*, with *n* words (total number of words) can be approximately determined by the above formula.

Method

TurCO is a text corpus of 50,111,828 words (Dalkılıç & Çebi, 2002).. It consists of the closed-captioned talks at the Turkish Parliament (52%) and the online editions of newspapers or magazines (44%). Each file in the corpus consists of 29 lowercase Turkish characters only (consonants: b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z; Vowels: a, e, ı, i, o, ö, u, ü) and the space character. Most of the documents included in the corpus have been collected from different web sites and there are some samples of Turkish novels and stories (0.13%). In this corpus, the average word length is about 6.2 characters.



Truncation Level	Vocabulary Count	Zipf Constant A	Zipf Constant A for rank range [0.1%,10%]
whole word	686804	0,022	0,084
7	259797	0,019	0,094
6	185602	0,018	0,097
5	110351	0,016	0,098
3	10933	0,028	0,185

Table 1. Frequencies of distinct words truncated at different levels and their Zipf constants for both the whole corpus and the interval between 0.1% and10% of ranks. Note that words were partially sorted with respect to their frequencies in descending order and then, their ranks were assigned in ascending order for each row starting from top. Finally the Zipf constant was computed by averaging *A(r)=p(r)*r* values of each row, where *r* and *p(r)* indicate the rank and the probability of that rank, respectively.

Discussion

In Table 1, frequencies of truncated terms and their Zipf constants are given. We see that in regard to the value of the Zipf constant, truncated terms around the average word length can be treated equally because the maximum change factor of the Zipf constant in that group is about 18% ((0,019-0,016)/0,016). The change factor is about 4% ((0,098-0,094)/0,094) for the word frequency ranks between 300 and 4000, indicating that any truncation scheme that takes the first five, six, or seven characters is as good as one another (Figure 1). The vocabulary growth rates of truncated terms around the average word length and the whole words are shown in Figure 2.

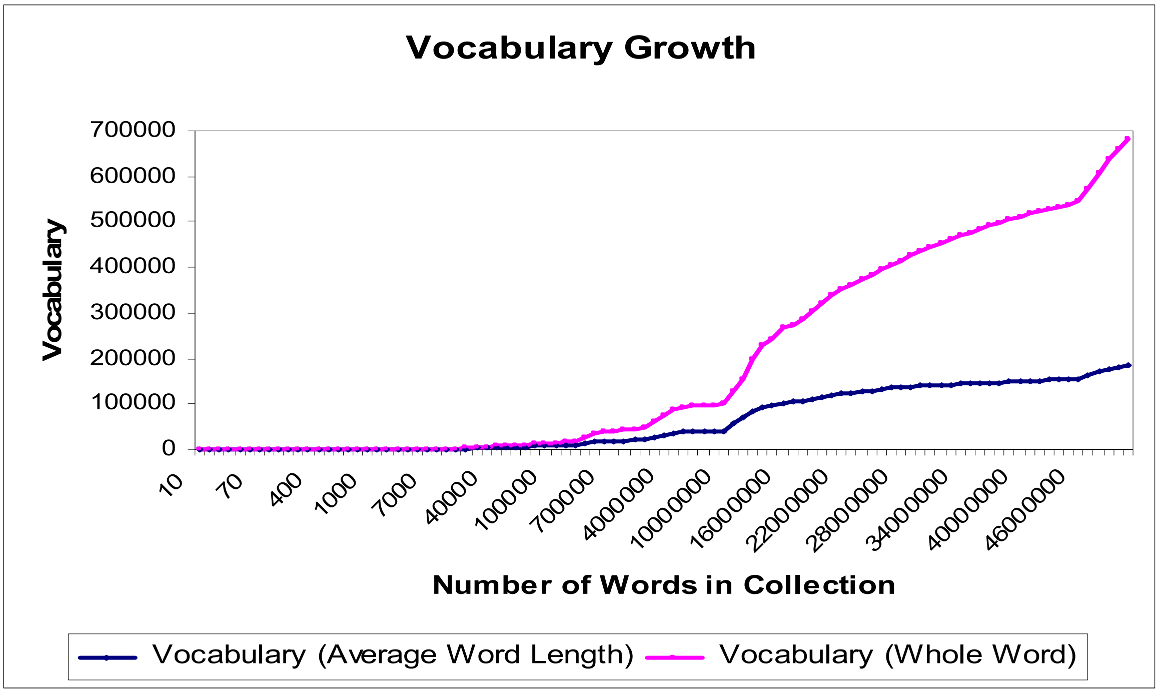


Figure 2. The vocabulary growth rate of truncated words around the average word length (6.2 characters) and of the whole words. Some files in the TurCO corpus consist of book samples whose novelty ratios can be regarded as very high, i.e., the contribution of the distinct words over the total number of words for the files in question were higher. One can easily see that the collection being made up of the whole words is very sensitive to the changes in the novelty ratios. Therefore the increase in vocabulary size for this collection is about three times higher than that of the collection being made of the truncated terms.

Conclusion

In this article we investigated the feasibility of using different truncation schemes for indexing. We showed that truncating words around the average word length yields better performance efficiency for Turkish text without compromising effectiveness. Our findings suggest that truncation can be used successfully for Turkish texts to better use the available disk space and processing capabilities of computers as well as to improve IR performance.

Acknowledgment

This work is partially funded by TUBİTAK-TİDEB under the grant number of TİDEB3040054.

References

Abu Bakar, Z. & Rahman, N.A. (2003). Evaluating the effectiveness of thesaurus and stemming methods in retrieving Malay translated Al-Quran documents. *Lecture Notes in Computer Science (LNCS)*, Springer Verlag, Vol. 2911, pp. 653-662.

Ahmad, F., Yusoff, M. & Sembok, T.M.T. (1996), Experiments with a stemming algorithm for Malay words, *Journal of the American Society for Information Science*, 47, 909-918.

Bitirim, Y., Tonta, Y. & Sever, H. (2002). Information retrieval effectiveness of Turkish search engines. In Tatyana Yakhno, ed. *Advances in Information Systems: Second International Conference, ADVIS 2002, İzmir, Turkey, October 23-25, 2002, Proceedings.* (pp. 93-103). Berlin: Springer-Verlag.

Braschler, M. & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? *Information Retrieval,* 7(3-4): 291-316.

Dalkılıç, G. & Çebi, Y. (2004). Zipf's Law and Mandelbrot's constants for Turkish language using Turkish corpus (TurCo). In Tatyana Yakhno, ed. *Advances in Information Systems: Fourth International Conference, ADVIS 2004, İzmir, Turkey.* pp. 273-282.

Dalkılıç, G. & Çebi, Y. (2002). A 300 MB Turkish Corpus and Word Analysis. In Tatyana Yakhno, ed. *Advances in Information Systems: Second International Conference, ADVIS 2002, İzmir, Turkey, October 23-25, 2002, Proceedings.* (pp. 205-212). Berlin: Springer-Verlag.

Dinçer, B.T. & Karaoğlan. B. (2003). Stemming in agglutinative languages: a probabilistic stemmer for Turkish. In: A. Yazıcı & C. Şener (eds.) *18th International Symposium on Computer and Information Sciences (ISCIS'03), Antalya, Turkey, November 3-5, 2003.* pp. 244-251.

Duran, G. & Sever, H. (1996). Turkce Govdeleme Algoritmalarinin Analizi. In Proceedings of Annual Conference of Turkish Informatic Association, Istanbul, Turkey, September 1996, pp. 235-243.

Ekmekçioğlu, F.Ç., Lynch, M. F. & Willett, P. (1995). Development and evaluation of conflation techniques for the implementation of a document retrieval system for Turkish text databases. *The New Review of Document and Text Management*, 1, 131-146.

Ekmekçioğlu, F.Ç., Lynch, M.F. & Willett, P. (1996). Stemming and N-gram matching for term conflation in Turkish texts. *Information Research*, 1(1) Retrieved July 6, 2003, from http://informationr.net/ir/2-2/paper13.html.

Ekmekçioğlu, F.Ç., Lynch, M.F., Robertson, A.M., Sembok, T.M.T. & Willett, P. (1996). Comparison of *n*-gram matching and stemming for term conflation in English, Malay, and Turkish texts. *Text Technology*, 6, 1-14.

Figuerola, C.G., Gomez, R., Rodriguez, A.F.Z. & Berrocal, J.L.A. (2002). Spanish monolingual track: The impact of stemming on retrieval. *Lecture Notes in Computer Science (LNCS)*, Springer Verlag, Vol. 2406, pp. 253-261.

Frakes, W.B. (1992), Stemming algorithms, in Frakes, W.B. and Baeza-Yates, R. (Eds), *Information Retrieval: Data Structures & Algorithms*, .(pp. 161-218). Englewood Cliffs, NJ: Prentice- Hall.

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J. & Smith, F.J. (2002). Extension of Zipf's Law to words and phrases. Retrieved July 6, 2005, from http://acl.ldc.upenn.edu/C/C02/C02-1117.pdf.

Harman, D. (1991). How effective is suffixing?, *Journal of the American Society for Information Science*, 42, 7-15.

Kalamboukis, T.Z. (1995). Suffix stripping with modern Grek. *Program*, 29, 313-321.

Küçük, M.E., Olgun, B. & Sever, H. (2000). Application of metadata concepts to discovery of Internet resources, *Lecture Notes in Computer Science (LNCS)*, Springer Verlag, Vol. 1909, pp. 304-13, 2000.

Li, W. (2005). Zipf's Law. Retrieved July 6, 2003, from http://www.nslij-genetics.org/wli/zipf/.

Pembe, F.C. & Say, A.C.C. (2004). A linguistically motivated information retrieval system for Turkish. *Lecture Notes in Computer Science*, Springer Verlag, Vol. 3280, pp. 741-750.

Popovic, M. & Willett, P. (1992). The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science,* 43, 384–90.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

Salton G.(1988). Automatic Text Processing, Addison Wesley, New York, NY

Sever, H. & Bitirim, Y. (2003). The analysis and evaluation of stemming algorithms for Turkish. *10th International Symposium on String Processing and Information Retrieval (SPIRE'03), Manaus, Brazil, October 8-10, 2003, Lecture Notes in Computer Science (LNCS)*, Springer Verlag, Vol. 2857, pp 238-51.

Solak, A., & Can, F., (1994). Effects of stemming on Turkish text retrieval. In *Proceedings of the Ninth International. Symposium on Computer and Information Sciences (ISCIS)*. (Antalya, Turkey, November 1994), pp. 49-56.

Solak, A. & Oflazer, K. (1993). Design and implementation of a spelling checker for Turkish. *Linguistic and Literary Computing,* 8, 113-130.

Tashakori, M. Meybodi, M. & Oroumchian, F. (2002). Bon: The Persian stemmer. *Lecture Notes in Computer Science (LNCS)*, Springer Verlag, Vol. 2510, pp. 487-494.

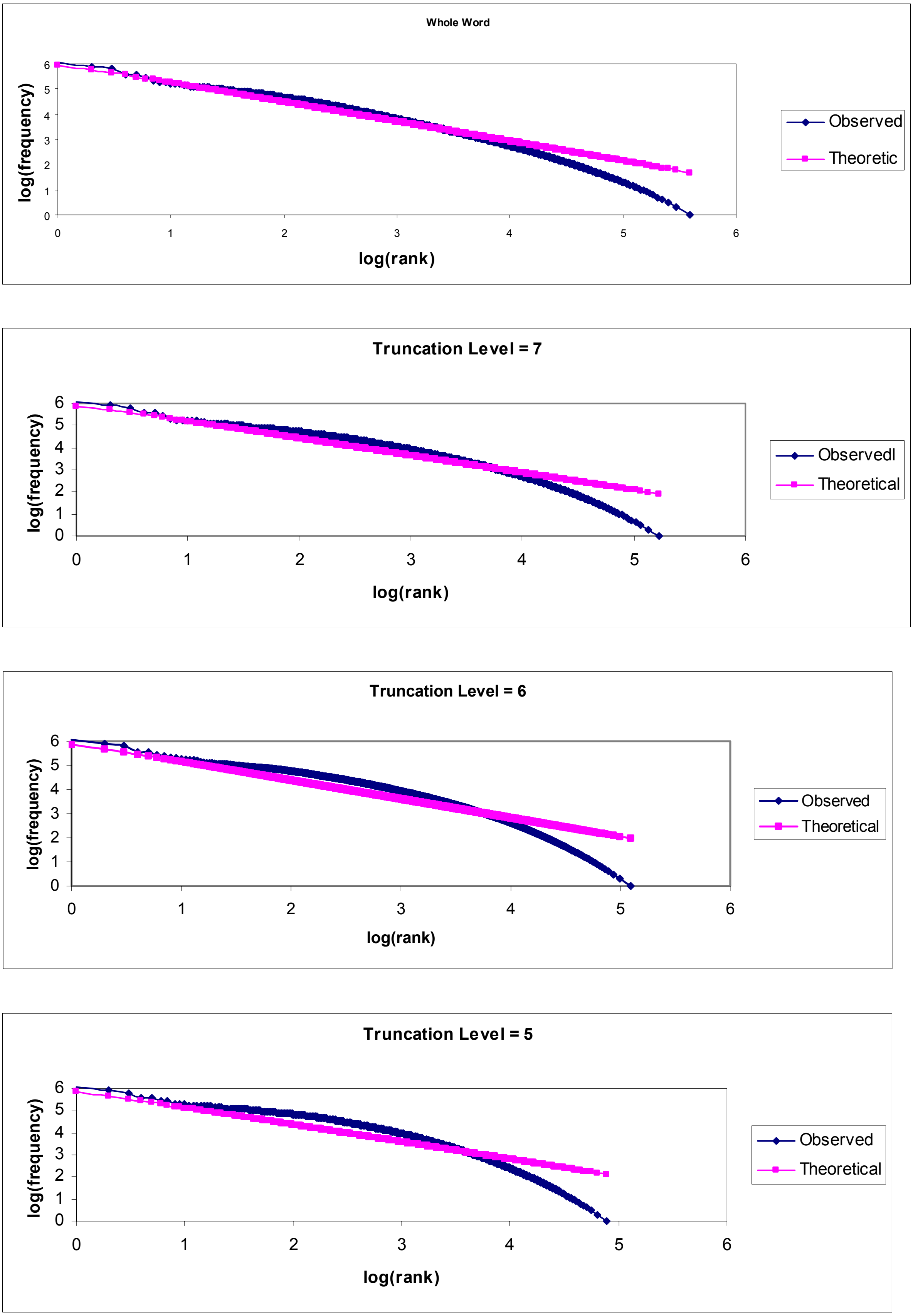


Figure 2. Charts of rank vs frequency on log base for some different truncation levels.