STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **L**, Number 2, 2005

THE LAW OF WORD LENGTH IN A VOCABULARY

DANA AVRAM LUPSA, RADU LUPSA

Abstract. In the literature we can find linguistics laws that are then exploited by many applications. This paper presents an empirical law that describes the frequency of the words of a given length in a language's vocabulary, as well as the length of distinct words in a corpus. This is a law that applies to any language.

1. Introduction

In linguistics there are some general laws that have no imediate consequences for computational linguistics. There are also some laws that are a gold mine and are exploited by many applications on computational linguistics.

From the first category are Zipf law and Heaps law. In linguistics, Zipf law [14] states that while only a few words are used very often, many or most are seldom used. The frequency of a word ranked the n-th (notated P_n) is given by the next relation: $P_n \approx \frac{1}{n^\alpha}$, where α is almost 1. This means that a word that occurs 10 times more frequently than another word, it is ranked 10 less. Heaps law [13] is an empirical law which describes the portion of a vocabulary which is represented by an instance document (or set of instance documents) consisting of words chosen from the vocabulary V. This can be formulated as $V_R(n) = K \times n^\beta$, where V_R is the subset of the vocabulary V represented by the instance text of size n. K and β are free parameters determined empirically. With English text corpora, typically K is between 10 and 100, and β is between 0.4 and 0.6. Heaps' law means that as more instance text is gathered, there will be diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn.

In the second category is placed the distributional hypothesis introduced by Harris [5] and which is widely used in NLP applications ([1], [6], [8] and the list can continue). The basic idea is that *we should know a word by the company it keeps* ([4]).

Received by the editors: November 5, 2005.

2000 *Mathematics Subject Classification.* 68R15, 68T50.

1998 *CR Categories and Descriptors.* G12 [**Mathematics of Computing**]: NUMERICAL ANALYSIS – *Approximation – Special function approximations*; I27 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE– *Natural Language Processing – Text analysis* .

A quite complete and recent survey on linguistic principles and their applications can be found in [1], [2], [12].

This paper describes a law that we found. Section 2 presents the parametrized function that describes the frequency of the words length in a language's vocabulary. It is an empirical law which is stated as being general, in the sense that it applies for any language. To the best of our knowledge, this is the first time the distinct word length frequency in a vocabulary is stated as a law.

On the other hand, the distinct words in a corpus are an approximation of that language vocabulary (see Heap's law in [13]). In consequence, in section 3 we verify if the frequences of distinct word lengths in a corpus are described by the same law.

In this paper, our study is applied for two languages: Romanian and English. The justification of the fitting the data with the given law is done by computing the relative error.

2. The Law that Describe the Frequency of the Words of a Given Length in a Language's Vocabulary

This section presents the empirical law that describes the relation between the length of words and its absolute frequency (i.e. the number of distinct words for each possible word length) in a language vocabulary. The law takes into account the dictionary forms only, and each of them is counted once.

In the following experiments we rely of the fact that the dictionary word forms (also named basic forms) are found as entries in a dictionary and they are also the forms of the words in Wordnet synsets.

2.1. The Law.

The absolute frequency of lengths of the words in a language vocabulary states that the absolute frequency of words of a given length is approximated by the function:

$$LV(x; c, k, \theta) = c \times x^k \times e^{-\frac{x}{\theta}} \tag{1}$$

where k, θ and c are parameters determined experimentally and $LV(x; c, k, \theta)$ is the law that describe the absolute frequency of the dictionary form of the words with length x in a language vocabulary.

2.2. Experimental Data. The Vocabulary.

2.2.1. *Romanian Experimental Data.* We took the Internet version of the Romanian explanatory dictionary ([3]), 1998 edition, off-line version, that we shall call *dex98.* Its database contains 41466 entries of 39531 distinct word forms.

2.2.2. *English Experimental Data.* We extracted the English words from the Wordnet for English. We rely on the fact that the English vocabulary is formed by all the words that appears in the Wordnet synsets. We also consider that synsets contains only the base form of the words. There were 74331 distinct words.

2.3. **The Hypothesis and Romanian Vocabulary.** In this section we present the way we approximate the parameters for LV function. We also compute the relative error with which the LV function approximates the Romanian vocabulary. Considering that the computed values leave room for further improvement, we also refine our search for values of the parameters and we point out that this leads to improvement.

2.3.1. *LV Parameters Estimation.* We selected all the distinct basic word forms that are found as entries in the dictionary. We grouped them by their length. In order to verify the hypothesis, we determined the values for c, k and θ, and we computed the approximation error of the parametrized LV function. We estimated the best values for c, k and θ in the following set of possible values:

$$\begin{array}{ll} c \in \{1, 2, \ldots, 100\} & , \\ k \in \{0.1, 0.2 \ldots, 9.9, 10.0\} & , \\ \theta \in \{0.1, 0.2, 0.3, \ldots, 9.9, 10.0\} & . \end{array} \tag{2}$$

For all these possible values of c, k and θ we computed the following sum:

$$\sum_{i=1}^{n} |LV(i, c, k, \theta) - d_i|$$

where n is the number of distinct lengths of the words, and d_i is the number of the words with length i. We took as the best parameters those that minimize the above sum, that is:

$$(c, k, \theta) \leftarrow argmin_{c,k,\theta} \sum_{i=1}^{n} |LV(i, c, k, \theta) - d_i| \tag{3}$$

The best parameters for LV describing the absolute words length frequency according to the Romanian dictionary are:

$$\begin{array}{ll} c = 1 & , \\ k = 9 & , \\ \theta = 0.8 & . \end{array} \tag{4}$$

The graphical representation of the distinct word length absolute frequency, along with the function from (1) with the parameters from (4) is shown in figure 1. Experimental data computed from *dex98*are given as vertical lines. The approximation function of (1) is given as the continous curve.

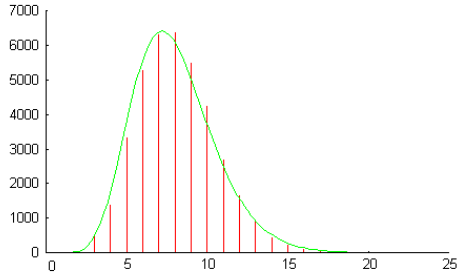

FIGURE 1. Distinct word length absolute frequency as of *dex98* approximated by LV with parameters from (4)

2.3.2. *Error of the Approximation.* Using the notations d_i for the absolute frequency of length i words and e_i the (theoretical) aproximation ($e_i = LV(i)$), we computed the absolute error (that we notate $AbsErr$):

$$AbsErr = \sum_{i=1}^{n} |d_i - e_i| = \sum_{i=1}^{21} |d_i - e_i| \approx 2848$$

The relative error (notated $RelErr$) is considered:

$$RelErr = \frac{AbsErr}{\sum_{i=1}^{n} |\max(d_i, e_i)|} \tag{5}$$

This relative error formula has the next properties:

- the minimum error value is 0
- the error is 0 if and only if $d_i = e_i$, $\forall i \in 1, 2 \ldots, n$
- the error is undefined if and only if $d_i = 0$ and $d_i = 0$ $\forall i \in 1, 2 \ldots, n$
- the maximum error value is 1
- the error is 1 if and only if
  - we are not in the next case: $d_i = 0$ and $d_i = 0$ $\forall i \in 1, 2 \ldots, n$
  - $d_i = 0$ or $e_i = 0$, $\forall i \in 1, 2 \ldots, n$

In our case, $d_i > 0$ and $e_i > 0$, $\forall i \in 1, 2 \ldots, n$, so the next relation holds:

$$0 < RelErr < 1 \ .$$

More precisely, for function from (1) and with parameter values from (4), the relative error is:

$$\begin{aligned} RelErr &= \frac{AbsErr}{\sum_{i=1}^{n} |\max(d_i, e_i)|} \\ &\approx \frac{2848}{40670} \approx 0.0700 = 7.00\% \end{aligned} \tag{6}$$

In the next section (2.3.3) we will see that the parameters can be approximated better by using smaller steps, and that, in this case, the relative error is reduced.

2.3.3. *Fine Tuning the Parameters.* The first five ranked estimated parameters (in section 2.3.1) for (c, k, θ) are presented in (7).

$$\begin{array}{llll} c = 1.00 & k = 8.50 & \theta = 0.90 & ; \\ c = 3.00 & k = 7.90 & \theta = 0.90 & ; \\ c = 1.00 & k = 9.00 & \theta = 0.80 & ; \\ c = 2.00 & k = 8.10 & \theta = 0.90 & ; \\ c = 4.00 & k = 7.80 & \theta = 0.90 & . \end{array} \tag{7}$$

Starting with the remark that the first five ranked parameters satisfy (8):

$$\begin{array}{rcll} c & \in & \{1, 2, 3, 4\} & , \\ k & \in & [7.8, 9.00] & , \\ \theta & \in & \{0.8, 0.9\} & , \end{array} \tag{8}$$

we searched the best values for c, k, and θ in the following set of possible values:

$$\begin{array}{ll} c \in \{0.10, 0.11, \ldots, 4.99, 5.00\} & , \\ k \in \{7.00, 0.11, \ldots, 10.00\} & , \\ \theta \in \{0.10, 0.11, \ldots, 3.00\} & . \end{array} \tag{9}$$

The best identified parameters for (c, k, θ) are:

$$\begin{array}{ll} c = 0.32 & , \\ k = 9.89 & , \\ \theta = 0.75 & \end{array} \tag{10}$$

In this case, the relative error (computed with formula (5)) is:

$$\begin{array}{rl} RelErr & = \frac{\sum_{i=1}^{n} |d_i - e_i|}{\sum_{i=1}^{n} |\max(d_i, e_i)|} \\ & \approx 0.0262 = 2.62\% \end{array} \tag{11}$$

The new identified parameter are better; there is a decrease of the error by 62.4%.

The graphical representation of the distinct word length absolute frequency, along with the best approximation function of the form (1) with the parameters from (10) is shown in figure 2. Experimental data computed from *dex98* are given as vertical lines. The approximation function of (1) is given as the continous curve.

2.4. **The Hypothesis and English Vocabulary.** We estimated the LV parameter values for English vocabulary (see section 2.2.2) in the same way as in sections 2.3.1 and 2.3.3 for Romanian data.

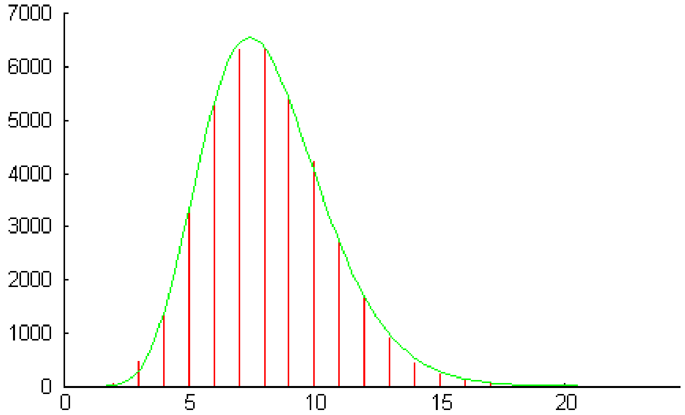

FIGURE 2. Distinct word length absolute frequency as of *dex98*approximated by LV with parameters from (10)

2.4.1. *LV Parameters Approximation.* By using the same method as in section 2.3.1 and searching possible parameter values as indicated in list (2), the best parameters we computed for (c, k, θ) are:

$$\begin{array}{rl} c = 3 & , \\ k = 7.8 & , \\ \theta = 1 & . \end{array} \tag{12}$$

2.4.2. *Error of the Approximation.* We computed the relative error as in section 2.3.2, formula (5). The relative error for parameter values from (12) is:

$$\begin{aligned} RelErr &= \frac{\sum_{i=1}^{n} |d_i - e_i|}{\sum_{i=1}^{n} |\max(d_i, e_i)|} \\ &= \frac{\sum_{i=1}^{27} |d_i - e_i|}{\sum_{i=1}^{27} |\max(d_i, e_i)|} \\ &\approx \frac{6698}{80023} \approx 0.0837 = 8.37\% \end{aligned} \tag{13}$$

2.4.3. *Fine Tuning the LV Parameters.* Best ranked five parameters are:

$$\begin{array}{llll} c = 6.00 & k = 7.10 & \theta = 1.10 & ; \\ c = 4.00 & k = 7.60 & \theta = 1.00 & ; \\ c = 2.00 & k = 8.40 & \theta = 0.90 & ; \\ c = 3.00 & k = 7.80 & \theta = 1.00 & ; \\ c = 7.00 & k = 7.00 & \theta = 1.10 & . \end{array} \tag{14}$$

Starting with the observation that the first five ranked parameters satisfy:

$$\begin{array}{llll} c & \in & \{2, 3, 4, 6, 7\} & , \\ k & \in & [7.10, 8.40] & , \\ \theta & \in & \{0.9, 1.1\} & , \end{array} \tag{15}$$

we estimated the best values for c, k, and θ in the following set of possible values:

$$\begin{array}{ll} c \in \{0.10, 0.11, \ldots, 4.99, 8.00\} & , \\ k \in \{7.00, 0.11, \ldots, 10.00\} & , \\ \theta \in \{0.10, 0.11, \ldots, 3.00\} & . \end{array} \tag{16}$$

The best values identified for parameters (c, k, θ) are:

$$\begin{array}{ll} c = 0.9 & , \\ k = 8.84 & , \\ \theta = 0.89 & . \end{array} \tag{17}$$

In this case, the relative error (see formula 5) is:

$$RelErr \approx 0.0572 = 5.72\% \tag{18}$$

There is a decrease of the error by 36.4%. This means that the new identified parameter are better.

The graphical representation of the distinct word length absolute frequency, along with the best approximation function of the form (1) is shown in figure 3. The function from (1) is represented by continuous line. The light color continuous line correspond to parameters from (12) and the dark color continuous line correspond to parameters from (17).

2.5. **Discussion.** We saw above that the distinct word length absolute frequency in a vocabulary is described by the parametrized function LV (see formula (1)). With the appropriate choice of the parameter values the absolute word length frequency in a vocabulary is aproximated by LV with precision of 97.38% for Romanian vocabulary (with parameter setting from (10)) and with precision of 94.28% for English vocabulary (with parameter setting from (17)).

We conclude that for a language there are c, k, and θ such that the frequency of the basic forms of the words of a given length is approximated by $LV(x, c, k, \theta)$, where x is the word length. This is *the law of the frequency of words Length in a Vocabulary.*

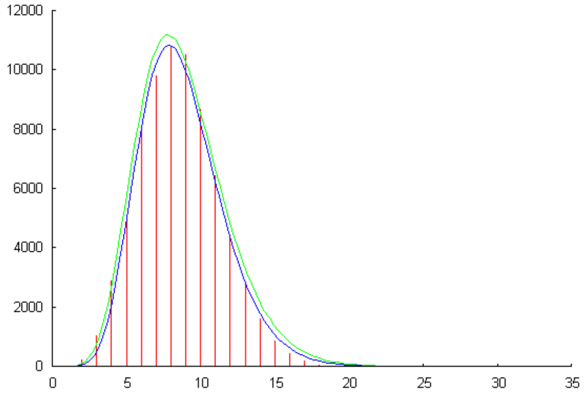

Figure 3. Distinct word length absolute frequency as of Wordnet approximated by LV with parameters from (10)

2.5.1. *A Short Comparison between Romanian and English Data.* We stated that the absolute word lengths frequency from vocabularies are approximated by the parametrised LV function (and we experimented that over Romanian and English vocabulary). If this is true, the shape of the histograms (in our case for Romanian and English) is the same.

We selected all the distinct basic word forms that are found in Romanian and English vocabulary (data presented in section 2.2) and grouped them by their length. The figure 4 represents the relative frequency of the distinct word lengths for both English and Romanian case. The length is represented on the x axis, and the relative frequency on the y axis. Note that the two histograms have the same shape.

3. The Frequency of the Distinct Word Forms of a Given Length in Texts

In this paragraph we shall see that the frequency of the distinct word lengths in texts is approximated by the same LV function.

Following the Heap's law [13], we could say that if a corpus is large enough, the distinct words from the corpus are an approximation for the words of the vocabulary of that language. That is why we expect that the frequencies of distinct words from a corpus, grouped by their length, to follow the same law. The difference between the distinct words in a corpus and the words in a vocabulary relies on the fact that the words which are entries in a dictionary (see section 2.2.1) are only the base form of the words. To one word from a dictionary usualy corresponds more than one word in a corpus; these are the flexioned forms of the word.

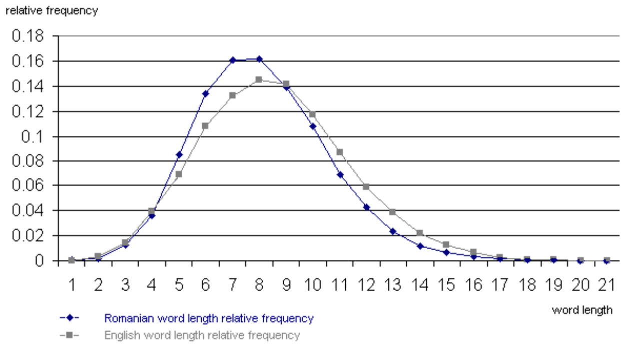

FIGURE 4. The relative frequency of distinct word lengths as of Romanian Dex and of English Wordnet.

We measured the length of each word in a given corpus, and, for each length, we counted all the distinct words of that length. We state that the number of distinct words of a given length is approximated by the LV function (see section 2.1).

We conducted the following experiment: we took all the different words from a Romanian language corpus, and we represented graphically the frequency of each word length as a function of the length.

3.1. **Romanian and English Corpora.** The Romanian corpus was automatically extracted from the Internet, by using a search by the words *limbaj* and *natural*. The corpus contains about 85000 words and 12500 among them are distinct. The English language corpus was constructed similarly to the Romanian one, by using a search by the words *natural* and *language*. The corpus contains about 50000 words and 5500 among them are distinct.

3.2. **The Approximation by LV Function Hypothesis.** The law of frequency of distinct words length in a corpus can be written as:

$$LV(x;k,\theta,c) = c \times x^k \times e^{-\frac{x}{\theta}} \tag{19}$$

where $LV(x;k,\theta,c)$ is the frequency of the distinct words of length x in the corpus (that is, the number of distinct words of length x in the corpus over the total number of distinct words in the same corpus), and k, θ, and c are experimentally determined parameters.

3.3. **LV Parameters Estimation.** Figures 5 and 6 show the approximated functions for the word length frequencies in texts. A discussion about the way the parameters are determined and the relative error of the approximation is given below.

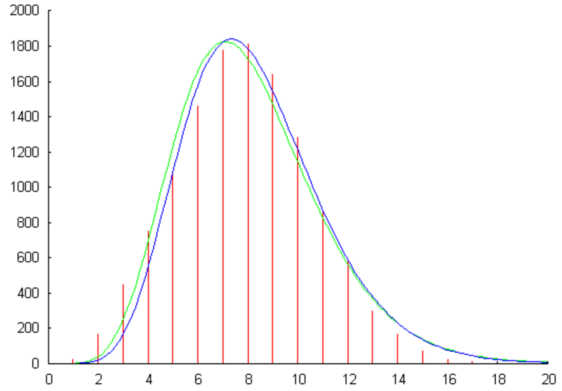

FIGURE 5. The approximation function for distinct word length absolute frequency in a Romanian corpus with two sets of parameters

We used steps from list (2) for the first approximation of LV parameters for Romanian texts. The estimated parameters were:

$$\begin{aligned} c &= 2 \quad , \\ k &= 7.10 \quad , \\ \theta &= 1.00 \quad . \end{aligned} \tag{20}$$

For those parameters, the relative error is 12.56%.

By verifing the values around the previously determined parameters by using smaller steps we determined the next parameters:

$$\begin{aligned} c &= 0.65 \quad , \\ k &= 7.99 \quad , \\ \theta &= 0.92 \quad . \end{aligned} \tag{21}$$

For those parameters we get better results; the relative error is 10.97%.

In figure 6, the light color represents LV function approximation with parameters from (20) and the dark color represents LV function approximation with parameters from (21).

We used steps from list (2) for the first approximation of LV parameters for English texts. The estimated parameters were:

$$\begin{aligned} c &= 12 \quad , \\ k &= 4.9 \quad , \\ \theta &= 1.3 \quad . \end{aligned} \tag{22}$$

For those parameters, the relative error is 7.36%.

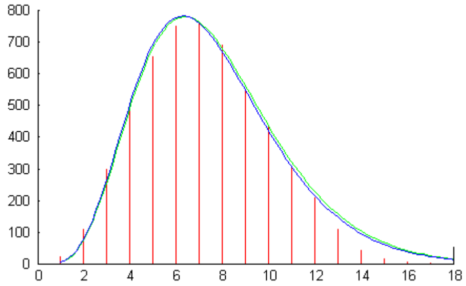

FIGURE 6. The approximation function for distinct word length absolute frequency in an English corpus with two sets of parameters

We verify the parameter values around the previously determined parameters by using smaller steps and we determined the next values:

$$\begin{array}{rl} c = 12.75 & , \\ k = 4.91 & , \\ \theta = 1.28 & . \end{array} \tag{23}$$

For those parameters we get better results; the relative error is 6.55%.

In figure 6, the light color represents LV function approximation with parameters from (22) and the dark color represents LV function approximation with parameters from (23).

4. Conclusion and Future Research

This paper presents for the first time an empirical law which we call LV. It describes the frequency of the words of a given length in the vocabulary of a given language. It also approximates the frequency of distinct words length in a corpus. This is a law stated as being general, in the sense that it applies to any language. But we have studied it for two languages: Romanian and English. We intend to extend the verification of this law over other languages and we think to use EuroWordnet in order to do that.

The experiments indicates that most frequent in a vocabulary are words with length between six and ten. It is easy to see that words that do not carry semantic information of their own (as preposition, conjunction, auxiliary verbs) are among the shortest words in a language. We intend to use this remark to try to improve contexts clustering process by introducing a new clustering parameter which is word feature length. Language independent methods of clustering similar contexts ([7], [6], [10]) on which relies a multitude of other linguistic processing, as identifying similar words ([8]), name discrimination ([9]), word sense discrimination ([11]) do not use the length as a word feature parameter.

References

[1] S. Bordag. *Algorithms extracting linguistic relations and their evaluation.* 2005.

[2] S. Bordag and G. Heyer. *A structuralist framework for quantitative linguistics.* 2005.

[3] dexonline, Romanian **EX**planatory **D**ictionary 1998 (*dex98*). http://www.dexonline.ro/ (visited 2004).

[4] J.R. Firth. Modes of Meaning. In *Papers in Linguistics*, Oxford University Press, 1957.

[5] Z. Harris. *Mathematical structures of language.* Interscience Publishers, New York, 1968.

[6] A. Kulkarni and T. Pedersen. SenseClusters: Unsupervised clustering and labeling of similar contexts. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2005.

[7] T. Pedersen. Language independent methods of clustering similar contexts. In *Eurolan 2005 Summer School. Tutorials*, 2005.

[8] T. Pedersen and A. Kulkarni. Identifying similar words and contexts in natural language with senseclusters. In *AAAI 2005*, 2005.

[9] T. Pedersen, A. Purandare and A. Kulkarni. Name discrimination by clustering similar contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, 2005.

[10] T. Pedersen, A. Purandare and A. Kulkarni. Senseclusters home page, 2005. http://senseclusters.sourceforge.net/.

[11] A. Purandare and T. Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2004.

[12] B. Rieger. Computing granular word meanings. A fuzzy linguistic approach in computational semiotics, 2005.

[13] http://en.wikipedia.org/wiki/Zipf%27s_Law

[14] http://en.wikipedia.org/wiki/Heaps%27_law

Babes-Bolyai University, Faculty of Mathematics and Computer Science, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: `davram@cs.ubbcluj.ro, rlupsa@cs.ubbcluj.ro`