Alphabet usage pattern, word lengths, and sparsity in seven Indo-European languages

Nikhil Kumar Rajput and Bhavya Ahuja
Department of Computer Science, Ramanujan College, New Delhi, India

Manoj Kumar Riyal
VCSG Uttarakhand University of Horticulture and Forestry, Tehri Garhwal, India

Abstract

An empirical study on about 1.7 million dictionary words from seven languages viz. English, French, Dutch, Spanish, Italian, Hindi, and German has been conducted. Three intriguing characteristic features have been analyzed. First, the alphabet usage pattern in a language was determined which can be used to give an idea on how alphabets have been employed. For instance, the alphabet 'e' is highly used in English, while 'q' is least used. Second, the average and range of word lengths in the languages were computed and seen to vary from 1 to 37. Average word lengths were computed in the range (6.665–11.14). For comparison, word lengths have been fitted using Gaussian distribution. Third, a new measure was derived; which we termed 'Language Sparsity'; computed as one minus ratio of number of words of a particular length already existing to the total number of possible words that can be formed. Sparsity hence gives a measure of the scope of fruition in languages. Two such measures have been defined: a weighted and a nonweighted sparsity. Nonweighted sparsity was found to be minimum (0.877) for English and maximum (0.982) for Dutch. The results obtained can play a significant role in propagating the synergy of language evolution.

Correspondence:
Bhavya Ahuja, Department of Computer Science, Ramanujan College, University of Delhi, New Delhi, India.
E-mail:
b.ahuja@ramanujan.du.ac.in

1 Introduction

There has been considerable research on human languages the world over. On the linguistic front, languages have generally been characterized building on components like phonemes, graphemes, and morphemes (Bolinger, 1948; Bobrow and Fraser, 1968; Rey *et al.*, 2000; Weingarten *et al.*, 2004).

The properties of language construct can be studied in terms of glyph, character, grapheme, and morpheme. A glyph is a particular representation of a character, while a character is the abstract concept represented by it. In other words, the same character written in a different style (say, sans serif versus calligraphic) is considered a different glyph. A grapheme is a letter or a number of letters representing a sound (phoneme) in a word. A morpheme is an indivisible unit of language that has some relevance. They possess meaning which would be lost if the morpheme is divided. It can be a root, a full word, a suffix. Morphemes are the units that morphology works with. This

Digital Scholarship in the Humanities, Vol. 35, No. 4, 2020.

doi:10.1093/llc/fqz076 Advance Access published on 26 October 2019

characterization formulates the morphology of the languages.

On the contrary, in the area of computational linguistics, statistical techniques have been deployed to study the structural attributes possessed by literary artifacts. Computer scientists and statisticians have studied languages building on features like Entropy, Zipf law, and Heaps' law to name a few (Shannon, 1951; Li, 1992; Ferrer i Canchol, 2005; Takahira *et al.*, 2016).

Some researchers have also made an attempt to study the structure of languages and its use (Finegan, 2014). Another branch called linguistic typology which is the study of structural and functional properties of languages has also attracted researchers for gaining insight in understanding the phenomenon of language evolution (Shopen, 1985; Hawkins, 2015).

A statistical study of Bhagavad Gita in four languages, Hindi, Sanskrit, English, and French, was done in Rajput *et al.* (2019) where the word frequency and length distributions were modeled. A study of phonological networks of English language in Stella and Brede (2015) showed that new words have been formed from small modifications of old words. The null model suggests constraints like avoidance of large degrees, triadic closure, and large nonpercolating clusters during word assemblies. A classification of text in the variants, Netherlandic or Flemish of Dutch, was performed using text statistics. The attribute set for classification included average word length and sentence length and syntactic features including ratios of function words and *n*-grams (Van der Lee and Van den Bosch, 2017). Through an empirical analysis of corpora of 96 languages from Wikipedia, it was found that the most frequent word forms are those that are efficiently producable and understandable to enhance communication (Mahowald *et al.*, 2018). An empirical investigation of phenomena via statistical, mathematical, and computational techniques was performed in de Araújo (2013) to understand language structures. Phoneme inventories in 451 languages of the world were also analyzed to study how languages compose their speech inventories and how an individual's linguistic variations are developed.

Words form the building block of any language which can be further considered as a combination of atomic units called an alphabet or letter. Mathematically, a language L can be defined as a set of words w. Every word in a language is constructed from an alphabet set $\sum$. The alphabets may be character literals, glyphs, or some symbols. At any point in time, the language set comprises of all the combinations of these alphabets that formulate a valid set of words (decided by an association). This set can grow further by introducing new clusters of the alphabets.

The analytical study proposed here has its foundation in the alphabet set deployed to form component words of the language. The three dimensions across which the study has been conducted can be structurally represented as:

(1) Frequency of usage υ of each alphabet in the set $\sum$ in the language L
(2) Average word lengths, $|w|$ in the language L
(3) Language sparsity

The above three parameters signify that the focus is on the identification of the present capacity that the language has utilized which can in turn point towards the direction and opportunities for further expansion. The following sections outline the proposed scheme for the analysis of words morphology in human languages along with the results derived for seven profoundly known and historic Indo-European languages: English, Hindi, German, Dutch, French, Spanish, and Italian. The results have been produced using dictionaries of the seven languages. The data source for English is raw.githubusercontent.com, Hindi is github.com, and for the others is www.gwicks.net. The analytic program was written in python and run on an Intel Core i5 CPU@2.60GHz and 4GB RAM as there were no specific platform requirements. The last section concludes the paper.

2 Alphabet usage pattern

The determination of letter usage frequencies has always been an area of immense interest. This

attribute has been functional in devising communication codes like Morse code and minimum redundancy code by Huffman (1952). It has also been the motivation behind design of QWERTY keypads David (1985). In this study, this feature has been deployed to determine the current utilization of the alphabet set of the language. This can not only be used to identify the most frequent alphabets but also those which have been underutilized.

Alphabets form the tokens that are used to construct most of the languages in the world. The number and usage pattern of the alphabets vary in different languages. Furthermore, in two different languages with intersecting alphabet sets, the relative usage of the intersecting alphabets is also not the same. This implies that an alphabet may be widely used in one language and its usage may be diminished in another.

To begin with the analysis, the frequency of utilization of this atomic unit has been computed. For each alphabet in Σ, the number of times the alphabet has been used to form valid words has been found. The values obtained are a record of the amount of usage of the alphabets in the language. The pattern of the alphabet usage may be considered related to the phonetic characteristics of the people who use the language and several other factors.

For mathematical description of the alphabet usage pattern, let us consider the set $\sum_C$ which denotes the frequency of usage of each alphabet υ_i in the language, where $i = 1, 2, \ldots, |\Sigma|$. For a language L containing a set of words constructed over an alphabet set Σ, the alphabet frequency Σ_C can be defined as a set of frequency values of the alphabet used in the words of language L. We can understand this with an example.

Example 1. Suppose a language $L = \{a, aa, ab\}$ (has only three words) is defined over an alphabet set $\Sigma = \{a, b\}$ then,
$\Sigma_C = \{4, 1\}$ Here, the alphabet 'a' is used widely whereas 'b' is mostly unused.
To exemplify, let us take the case of English, with

$$\Sigma = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z\},$$

416,267 words from the dictionary were considered. It was found that

$$\Sigma_C = \{330, 308, 71, 756, 164, 371, 124, 006, 412, 367, 43, 230, 90, 236, 103, 087, 338, 050, 6, 933, 32, 833, 216, 916, 115, 811, 275, 972, 272, 663, 120, 211, 6, 367, 269, 352, 273, 310, 247, 513, 141, 948, 378, 88, 26, 390, 11, 188, 75, 847, 16, 782\}$$

Σ_C for English points to the fact that alphabets like 'j', 'q', and 'x' are least used and the letter 'e' is most frequent. Results for the seven languages considered have been presented graphically in Fig. 1. It can be seen that in Dutch, the letter 'e' is most frequently used. Letters 'x', 'y', 'q', and the accented characters are least used. In French, letters 'k' and 'w' are least used as well as most of the accented characters, even though 'é' is used quite often. In German, letters 'j', 'q', 'x', 'y', and the accented characters are least used and the letter 'e' is mostly used. In Hindi, demarcating on the basis of vowels, consonants, and diacritics, it can be seen that the vowel, 'अ', the diacritic 'आ', the consonant 'र' are mostly used. In Italian, the letter 'a' is most frequent and the letters 'j', 'k', 'q', 'w', 'x', 'y', and the accented characters are least used. In Spanish, the letter 'a' is most used and the letters 'k', 'w', 'y', and the accented characters are least used. If the alphabets are positioned on a utility-scale ladder, the characters like accented ones, 'q', 'y' would be placed on the bottom and those like 'e' and 'a' would be on the top.

3 Average and range of word lengths in the language

In this second scheme, the analysis is advanced from alphabets to words formed by them. Languages across the globe depict a variation in terms of the length of the valid set of words they comprise. A conventional classification scheme categorizes languages into analytic and synthetic (Putnam, 1962) based on the way the words are formed from morphemes. Some language sets are composed mostly of long words and others have shorter ones. One of the

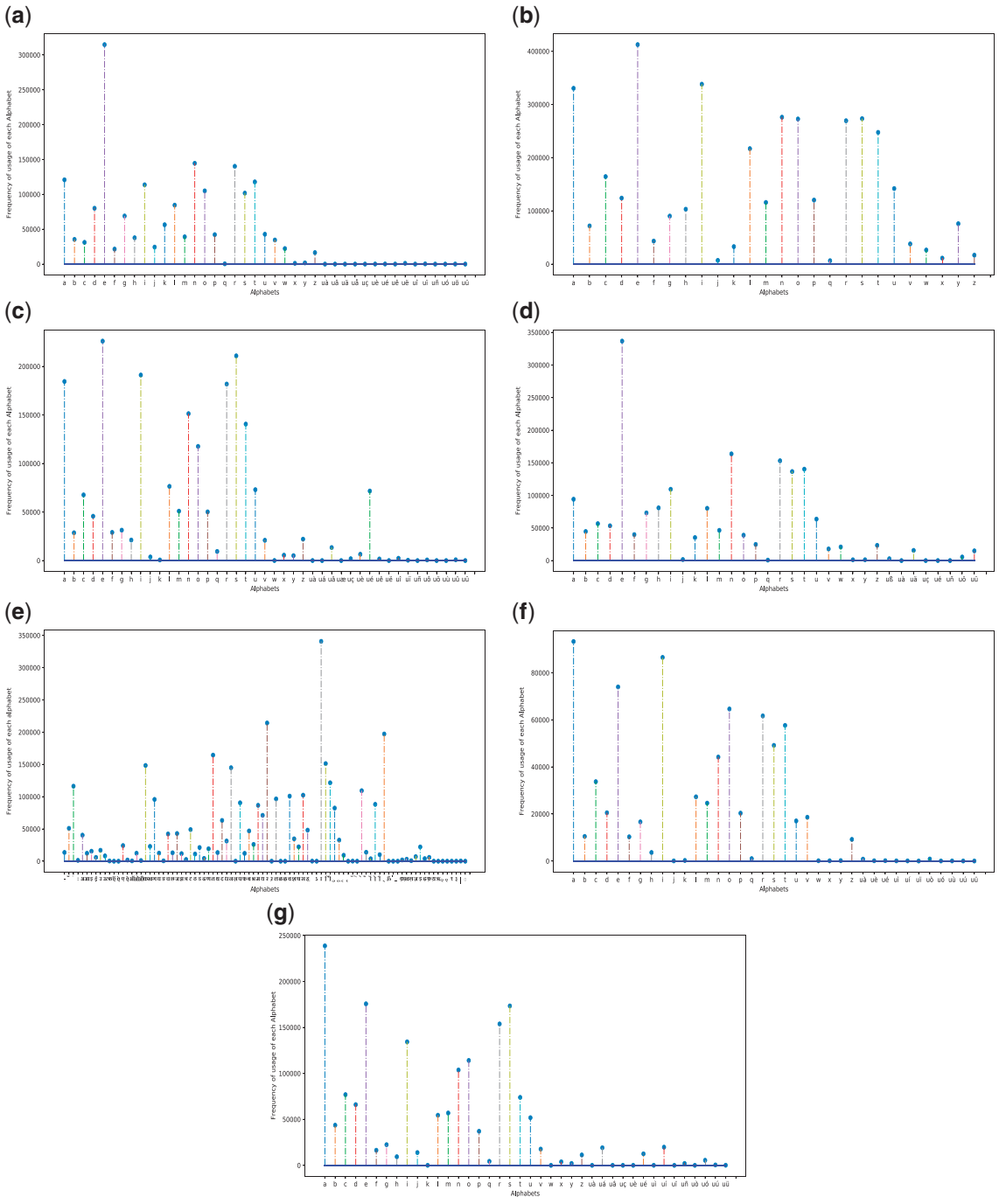

Fig. 1 Alphabet usage pattern (**a**) Dutch, (**b**) English, (**c**) French, (**d**) German, (**e**) Hindi, (**f**) Italian, and (**g**) Spanish

differentiating factors is the approach used to create words with morphemes.

Many researchers in the area of quantitative linguistics have applied statistical techniques in studying language structure. The study of word length distribution began early in the 1850s when English mathematician and logician Augustus de Morgan proposed using it as a measure of individual style and possibly for determining authorship (Eger, 2013). The study is based on three units of measurement: graphic (such as letters), phonetic (such as sounds, phonemes, syllables), and semantic (such as morphemes) (Grotjahn and Altmann, 1993). A morpheme is the atomic unit of language that is indivisible. They are the smallest components that morphology works with. Graphemes represent a set of letters that represent a sound in a word. Languages like Spanish and Italian are easy and transparent since in those languages one letter corresponds to one sound unit and languages like English portray complexity as they may have multiletter phonemes. Phonemes are the smallest unit of sound that can differentiate meaning. The results obtained have been used not only in characterizing a particular language but also for comparison with other languages. Also, these distributions have been found to follow some well-known laws. Several word length distribution models have been found namely geometric distribution (Elderton, 1949), lognormal distribution (Herdan, 1966), or compound Poisson, Ord distribution (Wimmer *et al.*, 1994), and mixed distribution models (Eger, 2013).

In Grotjahn and Altmann (1993), the authors laid precision on the fact that in most of the studies classifying languages; the data does not really fit the model estimated. They discussed six practical and theoretical problems in modeling the distribution of word lengths and tried to reconsider the model based on word length distribution. They referred to six such problems: units of measurement, population, data from a text, data from a frequency dictionary, data from a 'Normal' dictionary, and test of goodness-of-fit. Four methods were presented in Wimmer *et al.* (1994) which can be used to deduce a model based on distribution of word length. Compound Poisson and Ord distributions were found satisfactory for modeling word length distributions. German and English texts were analyzed and mixed Poisson distribution was found applicable for fifty-seven of the sixty texts analyzed (Riedemann, 1996). In Aoyama and Constable (1999), the authors have tried to find out the relationship between word length–frequency distribution in output and in the lexicon. It was found that word length frequency totals in English prose output are distributed geometrically and that the sequential distribution is random at the global level. In Constable and Aoyama (1999), the authors derived a mathematical characterization of some language features to differentiate isometrically lineated text from unlineated text. A variant of gamma distribution was derived to represent the relation between word length and frequency (Sigurd *et al.*, 2004). It was found that in Swedish and English, most words have three letters and shorter or longer length words occur less frequently. In Egghe (2006), size and rank frequency of articles and the impact factor of journals in various scientific fields were empirically analyzed and it was found that they were inversely related. Most of the size–frequency distributions were increasing and then decreasing.

In this study, word length distributions have been modeled using words from the language dictionaries. Evaluating the count of the words of a particular length is a sign of the capacity of the alphabet set utilized.

A language L was considered and the length of all the valid set of words $w \varepsilon L$; $|w|$ was computed. Post this, the count of all words of the same length was found, c_i where $i = 1,2,\ldots,N$, and i corresponds to $|w|$ and N is the maximum word length found. A set L_c was formed where $L_c = \{c_i, i = 1, 2, \ldots, N\}$.

The results for this process for Italian have been presented in Example 2.

Example 2. In Italian, 86,576 words were considered and the set L_c was obtained as: $L_c = \{5, 125, 425, 1, 483, 3, 724, 6, 425, 11, 381, 15, 159, 18, 554, 18, 951, 4, 645, 2, 755, 1, 521, 757, 395, 160, 72, 22, 12, 5\}$; $i = 1, 2, 3, 4, \ldots, 20$, where $N = 20$.

Figure 2 depicts the probability distribution for the word lengths of the seven languages. This study intends to compare the word length distributions on the basis of their Gaussian fit and a deeper analysis in this regard would be the scope of future work. It can be seen that the word length distribution in the languages replicates closely the Gaussian pattern. The results presented in Fig. 2 approximate the distribution

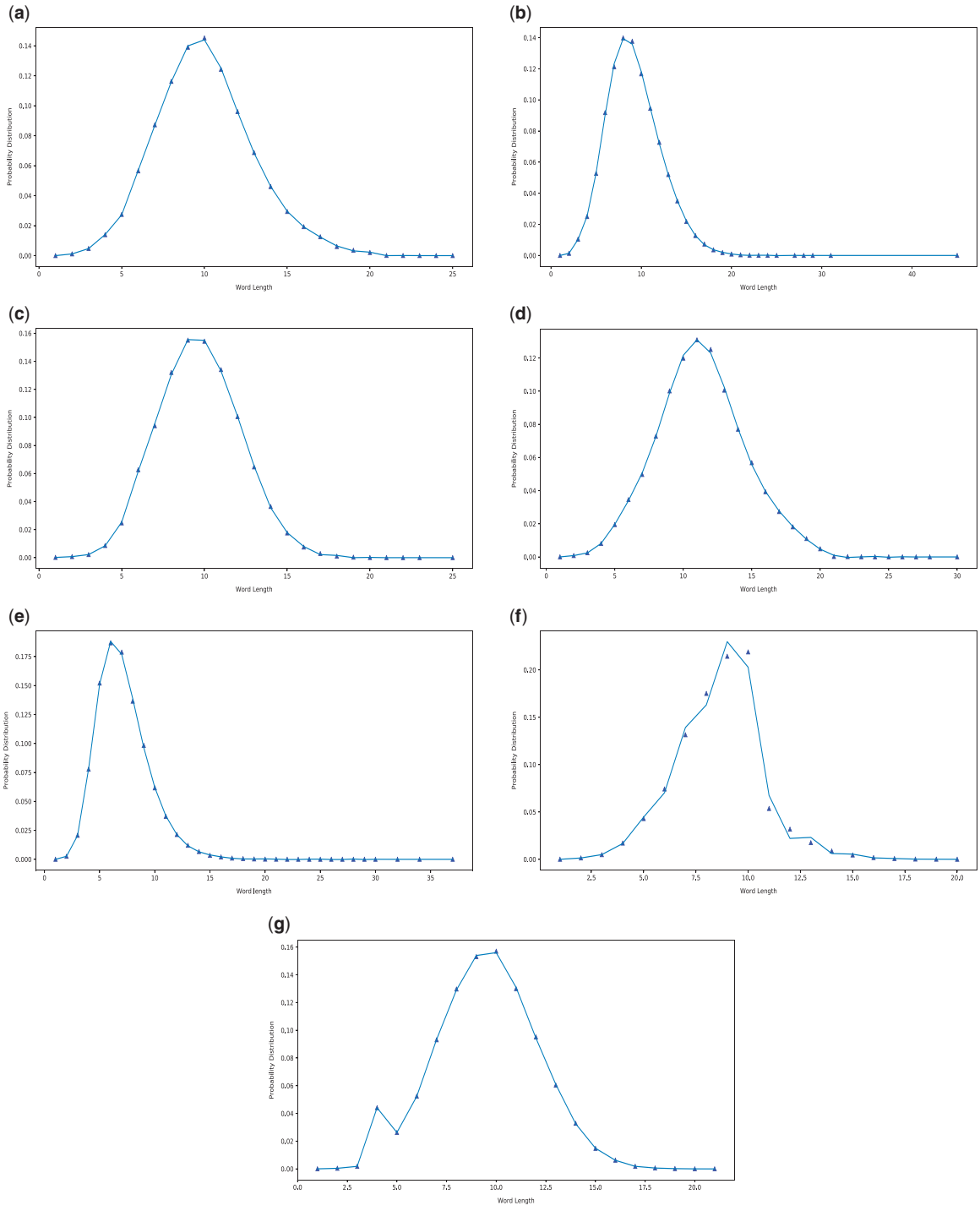

Fig. 2 Word length distributions (**a**) Dutch, (**b**) English, (**c**) French, (**d**) German, (**e**) Hindi, (**f**) Italian, and (**g**) Spanish

function and are based on the available dataset. However, even if we have the best possible dataset nearing complete list of words in a language, the deviation observed would not be significantly high. Hence, the results portray a good approximation.

Dutch, German, Italian, and French are observed to closely follow the Gaussian curve with mean values 9.761, 11.14, 8.578, and 9.59, respectively, and high variances. The curves for English (mean 8.78) and Hindi (mean 6.66) are positively skewed with English depicting an outlier. The variance reported is lesser. In Spanish, the mean value is 9.523 with high variance. An outlier too was reported in this case.

4 Language Sparsity

The third focus area is on deriving a mathematical coefficient that can give a quantification reporting the size of the yet unutilized segment of a language corresponding to alphabet combinations not yet used to create new words. Any language is composed of words formed by permutation of the valid set of alphabets. For example, with the English alphabets {t,e,a}, the set of possible permutations are {ate, aet, tea, tae, eat, eta}. Out of these the valid set of English words as of now are {ate, tea, eat} and the unutilized segment is {aet, tae, eta}.

To mathematically represent the size of this unutilized segment, a new coefficient has been determined, coined Language Sparsity. Two such measures have been computed, one a weighted mean and the other a nonweighted mean. The process for the same has been described further.

In the previous section, the average word lengths for various languages were computed and termed L_c. A language can have multiple words that can be generated by applying permutations on the alphabets. At one point in time, the valid set may comprise some number of words. The language set can be evolved further by addition of new words to the language which results from creating an unutilized combination of some alphabets from Σ. The set L_c can be used to estimate the present usage of the available alphabet set Σ in the language. Hence, it can be said that L_c denotes the present capacity of the language utilized or its maturity level.

Here, 'Language Sparsity' can be used to give a measure of the unutilized band of the language. It estimates what proportion of the alphabets has been left to be employed to form novel words in the language. It is a quantification of the possibility that the language offers in terms of vacancies still available for new plausible words to occupy. The sparsity derived can hence be used to visualize the possibility of language expansion by addition of new words. The following paragraphs explain the technique used to derive this index.

A hypothetical language L' can be created which contains words that fully utilize the alphabet set Σ. The set L' contains all possible words that can be formed from Σ_i where $i = 1,2,3, \ldots, M$. Here M is the maximum word length that we wish to keep in the language.

$$L' = \{\Sigma_1, \Sigma_2, \Sigma_3, \ldots., \Sigma_M\}$$

A set L'_c can be formed by including the word lengths possible with the alphabet set Σ.

$$L'_c = \{c_i;\ i = 1, 2, \ldots., M\}$$

The formula for constructing the set L'_c is:

$$L'_c = \{|\Sigma|^i;\ i = 1, 2, \ldots., M\}$$

This is explained with an example below:

Example 3. For alphabet set $\Sigma = \{a, b\}$ and $M = 2$, we have $L' = \{a, b, aa, bb, ab, ba\}$ and $L'_c = \{2, 4\}$, Since there are two words of length 1 and four words of length 2. A similar set L_c can be constructed which records the actual word lengths in the language L. In the above example, assuming present word set L is $L = \{a, aa, ab\}$, then $L_c = \{1, 2\}$ Now, the ratio of the current word lengths over the word lengths possible was calculated for 1 gram, bigram up to 10 grams taking $M = 10$. This value obtained is a pointer toward the amount of utilization of the alphabet set. The ratio obtained is subtracted from 1 as 1 would represent complete utilization. $S = \{1 - (L_c[i]/L'_c[i])\} = \{1 - 1/2, 1 - 2/4\} = \{1/2,\ 1/2\};\ i = 1, 2, \ldots, M$

Now, two measures have been computed which signify the language sparsity:

4.1 Nonweighted sparsity

In this case, the language sparsity has been calculated by summing up all the values obtained in S

and dividing by M. $S_{NW} = (\Sigma_i S_i)/M$; $i = 1, 2, \ldots, M$ and S_i denotes the ith ratio in the set S. The same has been exemplified below.

Example 4. In the set S above where $S = \{1/2, 1/2\}$; $S_{NW} = (1/2 + 1/2)/2 = 1/2 = 0.5$. The value obtained in this example denotes that half the capacity of the language is still left to be utilized.

4.2 Weighted sparsity

In this case, the language sparsity has been computed as a weighted mean. The significance of this measure is that it takes into consideration the most probable length of words in the language. As has been discussed in Mahowald *et al.* (2018), the words in a language are formed in accordance with convenience of communication. Some word assemblies are preferred due to their morphological and syntactic structure. Keeping this point in mind, it is imperative that higher weightage is given to the word lengths that are more likely and their neighboring lengths as opposed to very large or very small lengths. The nonweighted sparsity measure does not take account of this.

While calculating the sparsity, each word length is associated with a weight. The weights have been calculated as a function of that length in the language with the maximum number of words. This particular length, different for each language, is denoted as *mode_len*.

Say the current set $L_c = \{c_i, i = 1, 2, \ldots., N$ as defined in the previous section gives the count of the number of words of a particular length. Then *mode_len* can be defined as: $mode\_len = max(c_i), i = 1, 2, \ldots., N$.

This implies that *mode_len* is the particular length with the highest frequency of words in the language. To reinstate, M is the length of the words for which we wish to compute the sparsity. Hence, M can either be less than, greater than, or equal to *mode_len*.

Then, $S_W = \Sigma_i(wt_i \times S_i)/M$; $i = 1, 2, \ldots, M$; S_i denotes the ith ratio in the set S and wt_i denotes the ith weight. S_W is the average of all $S_i's$. Here, the weights $wt = f(mode\_len)$ are derived as follows:

$$wt_i = \begin{cases} i/mode\_len & ; 1 \le i \le mode\_len \\ (2 \times mode\_len - i)/mode\_len & ; mode\_len < i < 2 \times mode\_len \\ 1/mode\_len & ; i \ge 2 \times mode\_len \end{cases} \quad (1)$$

In this approach, the frequently occurring length *mode_len* gets the highest weight, which then keeps on decreasing as we go from *mode_len* to 1 and *mode_len* + 1 to M. Post, 2 × *mode_len*, the weights are set as 1/*mode_len* as the weightage of greater lengths should be low. The weights have been divided by *mode_len* to normalize them and bring them in the range [0–1].

This can be understood by the subsequent example.

Example 5. In the set S above where

$$S = \{1/2, 1/2\}, mode\_len = 2, M = 2;$$

$$S_w = ((1/2 \times 1/2) + (2/2 \times 1/2))/2$$

$$= (1/4 + 1/2)/2 = (3/4)/2 = 0.375$$

Table 1 shows the results of ratios for word lengths from 1 to 10 and the weighted and nonweighted sparsity obtained for the seven languages. It is noteworthy that most of the ratios are above 0.75 which shows that effectively only 25% of the total possible words actually are valid ones. Nearly 75% of the total permutations have still not been used. For word lengths 6 and above, the ratios were found to be 0.999 for all the languages considered. Two special cases are observed. First, for French, with a word length of 1, the ratio was reported to be 0.186 that signifies that quite a lot of one letter words exist in French; the plausible capacity has been well utilized. Similarly in English, for word length 2, the ratio was obtained as 0.161 which again signifies a good number of bigrams in the language. All the sparsity measures present a collaborative result of the ratios obtained. All the nonweighted sparsity measures were found to be above 0.87. This high value clearly hints at the wide unutilized segment of the language. The values obtained for the weighted sparsity also point toward the above observation. It can be seen that most of the values are approximately 0.55. The values for English, French, and Hindi are nearly 0.57. This shows that half the segment of these languages still has not been evolved.

It is stated here that the sparsity obtained is dependent on the alphabet size of the language which on the one hand increases the possible number of permutations but also can lead to enhancement of the scope of word

Table 1 Ratios for the seven languages for word lengths 1–10 and corresponding nonweighted and weighted sparsity values

Length of words/language	Dutch	English	French	German	Hindi	Italian	Spanish
1	0.951	0.885	0.186	0.853	0.955	0.865	0.825
2	0.878	0.161	0.915	0.882	0.849	0.909	0.951
3	0.988	0.752	0.994	0.989	0.986	0.992	0.995
4	0.999	0.977	0.999	0.999	0.999	0.999	0.997
5	0.999	0.998	0.999	0.999	0.999	0.999	0.999
6–10	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Nonweighted Sparsity	0.982	0.877	0.909	0.972	0.979	0.976	0.977
Weighted Sparsity	0.547	0.579	0.578	0.496	0.577	0.546	0.547

formation. Also, this value does not present a language comparison parameter as the alphabet sets vary.

5 Conclusion

Three characteristic features namely frequency of usage of the language alphabets, average word lengths in the language and language sparsity have been used to study the structural attributes of languages. Seven languages have been considered here. It was seen that the letter 'e' was frequently used in these languages and some alphabets like 'j', 'x', and the accented ones have rare occurrence. The probability distributions for the length of words have been compared on the basis of fit on Gaussian distribution and the average values of the word length for all the languages under study have been obtained. The Gaussian curve has been well replicated with the cases of English and Hindi positively skewed. The sparsity for various languages was derived and can be used a sound measure for understanding the present status of the languages in terms of evolution and highlights the scope for growing further. The understanding developed through this scheme could motivate new word formation and motivate language evolution.

References

Aoyama, H. and Constable, J. (1999). Word length frequency and distribution in English: part i. prose. *Literary and Linguistic Computing*, **14**: 339–58.

Bobrow, D. G. and Fraser, J. B. (1968). A phonological rule tester. *Communications of the ACM*, **11**: 766–72.

Bolinger, D. L. (1948). On defining the morpheme. *Word*, **4**: 18–23.

Constable, J. and Aoyama, H. (1999). Word length frequency and distribution in English: Part ii. An empirical and mathematical examination of the character and consequences of isometric lineation. *Literary and Linguistic Computing*, **14**: 507–36.

David, P. A. (1985). Clio and the economics of qwerty. *The American Economic Review*, **75**: 332–7.

de Araújo, L. C. (2013). *Statistical Analyses in Language Usage*. Ph.D. thesis, Universidade Federal de Minas Gerais.

Eger, S. (2013). A contribution to the theory of word length distribution based on a stochastic word length distribution model. *Journal of Quantitative Linguistics*, **20**, 252–65.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, **69**: 131–52.

Elderton, W. P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society. Series A (General)*, **112**: 436–45.

Ferrer i Cancho, R. (2005). The variation of Zipf's law in human language. *The European Physical Journal B-Condensed Matter and Complex Systems*, **44**: 249–57.

Finegan, E. (2014). *Language: Its Structure and Use*. USA: Cengage Learning.

Grotjahn, R. and Altmann, G. (1993). Modelling the distribution of word length: some methodological problems. In Köhler, R. and Rieger, B.B. (eds), *Contributions to Quantitative Linguistics*. Dordrecht: Springer, pp. 141–53.

Hawkins, J. A. (2015). *A Comparative Typology of English and German: Unifying the Contrasts*. London: Routledge.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*, vol. 4. Berlin: Springer.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, **40**: 1098–101.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, **38**: 1842–5.

Mahowald, K., Dautriche, I., Gibson, E., and Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, **42**: 3116–34.

Putnam, H. (1962). The analytic and the synthetic. In Feigl, H. and Maxwell, G. (eds), *Minnesota Studies in the Philosophy of Science*, vol. III, Minneapolis: University of Minnesota Press.

Rajput, N. K., Ahuja, B., and Riyal, M. K. (2019). A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita. *Pramana*, **92**: 60.

Rey, A., Ziegler, J. C., and Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition*, **75**: B1–12.

Riedemann, H. (1996). Word-length distribution in English press texts. *Journal of Quantitative Linguistics*, **3**(3): 265–71.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, **30**: 50–64.

Shopen, T. (1985). *Language Typology and Syntactic Description*, vol. 3. Cambridge: Cambridge University Press.

Sigurd, B., Eeg-Olofsson, M., and Van Weijer, J. (2004). Word length, sentence length and frequency–zipf revisited. *Studia Linguistica*, **58**: 37–52.

Stella, M. and Brede, M. (2015). Patterns in the English language: phonological networks, percolation and assembly models. *Journal of Statistical Mechanics: Theory and Experiment*, **2015**: P05006.

Takahira, R., Tanaka-Ishii, K., and Debowski, Ł. (2016). Entropy rate estimates for natural language–a new extrapolation of compressed large-scale corpora. *Entropy*, **18**: 364.

van der Lee, C. and van den Bosch, A. (2017). Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Valencia, Spain.* Association for Computational Linguistics, pp. 190–9.

Weingarten, R., Nottbusch, G., and Will, U. (2004). Morphemes, syllables, and graphemes in written word production. *Trends in Linguistics Studies and Monographs*, **157**: 529–72.

Wimmer, G., Köhler, R., Grotjahn, R., and Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, **1**: 98–106.