H. Peter Ohly
Informationszentrum Sozialwissenschaften,
Bonn, FRG

A Procedure for Comparing Documentation Language Applications: The Transformed Zipf Curve

Ohly, H.P.: A procedure for comparing documentation language applications: the transformed Zipf curve.
In: Int. Classif. 9 (1982) No. 3, p. 125–128, 3 refs.

A common procedure for describing the word concentration of texts is presented by the ZIPF curve. Interpretation difficulties arise, however, at least where documentation language applications are compared. This paper presents a transformation consisting of a percentage-taking and a compression or stretching of the values of the ZIPF curve. Using three selected documentation languages as examples it is shown that the transformed ZIPF curve permits a differentiated comparison of concentrations. (Author)

1. The ZIPF Curve as a Measure of Concentration

ZIPF developed a procedure for measuring the concentration of word frequencies (tokens) by determining the normal inequality of word frequencies, taking this normal distribution as a standard and comparing other word distributions with it. In his procedure ZIPF orders the words according to their frequency, thus elaborating hierarchical distributions and examining the relationship between frequency and hierarchical rank of the word forms (types). He notes that the product of rank and frequency remains constant in normal-language texts. When these values are plotted graphically in a logarithmic coordinate system (abscissa = ranks, ordinate = frequencies) one obtains, for normal-language texts, approximately a straight line descending at an angle of 45° from the frequency of the most common word to the frequency of the rarest word (cf. Fig. 1a).

If all words were used equally frequently (no concentration) one would obtain a straight line parallel to the abscissa axis along which the ranks of the word forms are plotted. If a single word form were to unite all frequencies (complete concentration) one would obtain a vertical line parallel to the ordinate axis along which the frequencies are plotted. In classical concentration measurement – e.g. the LORENZ curve – concentration is defined as (any) deviation from the equal occurrence of all measurement values. Here, however, the deviation from the *normal* (inequal) distribution of the word forms of a language is considered. As the normal distribution of word forms one regards the empirically found ZIPF rule specifying that frequency is reduced by one half with every advance in rank. The descending line, i.e. logarithmically uniformly declining frequencies, is explained by the fact that on the one hand there is a tendency to use words economically ("unification") and on the other hand a tendency to express things in differentiated fashion ("diversification").

Figure 1: Examples of typical forms of the ZIPF curve
Figure 1a: Form of the ZIPF curve for normal-language texts
individual words (tokens)
word forms (types)
Figure 1b: Form of the ZIPF curve for "word-narrow" texts (unification)
Figure 1c: Form of the ZIPF curve for "word-wide" texts (diversification)

An inwardly curved (as opposed to a straight) line indicates a predominance of the most common words (e.g. in the case of pathological difficulties of expression). MEIER terms such a distribution "worteng" (word-narrow) (cf. Fig. 1b). An outwardly curved line, on the other hand, shows that a large proportion of the words is used almost equally frequently (e.g. in specialized texts)[1]. This constellation is termed "wortweit" (word-wide) by MEIER (cf. Fig. 1c). For applications of documentation languages (indexing corpus) one should expect outwardly curved lines: design and application of these artifical languages will tend to reflect the contents of documents both in a sufficiently specific and a sufficiently generalizing way.

2. Transformation of the ZIPF Curve

2.1 Problems of Interpretation

It should be noted that the ZIPF curve is dependent both of the absolute number of word forms and of the absolute frequencies of the individual words. If a text of the same language is expanded the frequencies will increase more strongly than the different word forms (cf. Fig. 2).

What must already be considered in the natural-language case is all the more true for documentation languages. They represent self-contained systems in which the word forms and the indexing depth are either fixed or oriented to standards, so that the ZIPF Law "(rank)x(frequency)=constant" cannot be verified (cf. Fig. 3).

Under this aspect the ZIPF curve must be regarded as being difficult to interpret for terminologically controlled languages. Also to be noted is the fact that the extreme values are of no importance in the interpretation of a ZIPF curve. But in documentation languages particularly the extreme ranges of the word frequencies are of special importance from the point of view of retrieval efficiency.

Also, it is not the linear ascent but only a change in the straight-line relationship – i.e. (rank)x(frequency) does not remain constant – which indicates inconsistent use of a repertory of word forms. Minor deviations in the straight-line pattern can, however, be found out only with difficulty from the ZIPF curve.

Figure 2: Variation of the ZIPF curve (straight line) with text length in the natural-language case (on a logarithmic scale)

Figure 3: Variation of the ZIPF curve (straight line) with indexing depth and number of word forms (types) in the documentation-language case

I: Normal form
II: Doubling of the (applied) vocabulary with indexing depth remaining constant (conceivable e.g. by having all frequencies to be cut in half)
III: 50% reduction of indexing depth with the full vocabularly being applied

On the other hand, the absolute ratio of maximum frequency and number of word forms, which can be interpreted as external conditions and expresses itself in the rate of ascent of the straight line, furnishes no indications for unequal word use (with respect to the normal constellation).

2.2 A Curve of Percentages

If, now, the straightness of the curves is to be judged from the graphic representation, segments of equal length of the curves need to be placed alongside one another. For this purpose it is possible to take sections of equal length starting either at the center, at the peak value or from the end points. It seems to be better, however, to obtain comparable sections by percentage-taking in expressing the frequencies in percent of the highest word frequency and the rank-indicating figures in percent of the highest rank.

In such standardized curves which are logarithmically plotted, only the medium ranges show a shape similar to the diagonal.

In examples for some constellations of rank and frequency maximum values (cf. Fig. 4), different end points are obtained in the logarithmic presentation of these percentage curves. In case the original data contain more than 100 word forms, the first rank is transformed to a percentage smaller than 1% (cf. Fig. 4d). And if a 1% rank has a smaller frequency than 100 in the original data, the smallest possible difference in frequency will be larger than 1% (cf. Fig. 4b).

When we are dealing with ZIPF curves obeying the relationship (f)x(m) = constant, the curves will have identical directions. As we can see from the examples in Fig. 4, the curves represent, to different extents, the totality of all values.

In text corpora with frequencies and ranks larger than 100[3], the extreme values are cut off. In text corpora with frequencies and ranks smaller than 100, strong scattering areas are produced at the extreme values since the enlargement of the minimum difference causes steep steps

Figure 4: ZIPF curves (straight lines) transformed to percentage values (f = token frequency, m = type frequency rank)

to occur. Assessment must therefore be performed on the basis of the medium range only. In some cases it may be useful, for presentation purposes, to put one of the points of intersection with the axes on 100% (e.g. the frequency value at a rank of 1%; cf. Fig. 5, Curve IV), causing a parallel shift of the curve in a common starting point. A shifting of both axis maximum values in this fashion is not possible, since by the shifting of the maximum to the 100% value the 1% value is shifted along as well, thus causing a new value to be determined as the maximum for the vertical axis in the given case[4].

2.3 A Curve of Compressed Percentages

If one dispenses with the interpretation of the values as percentages of a fixed basis, it is possible to shift the extreme values to uniform coordinate points through simple compression or stretching[5] of the curves (cf. Fig. 5, Curve V).

Since the ZIPF curve, in the double logarithmic presentation, is a diagional, which runs in each case through the class centers of the various word forms, hence the compressed curve can be so shifted as to have its rank center above the lowest individual-word frequency lie at the 100% rank (anyway the maximum value represents usually one single word).

Here, again, all values are shifted to the curve segment between the logarithmic coordinate values 1 and 100.

of the transformed ZIPF curve. The documentation languages selected were title words, index keywords and main categories for identical research project descriptions.[6] First of all percentage-based ZIPF curve on selected[7] word forms of these three documentation languages is examined (cf. Fig. 6). In this representation the rank places of the occurrence frequency are indicated in percent of the total number of word forms and the occurrence frequency of the word forms in percent of the occurrence frequency of the word forms on the 1% rank (type IV in Fig. 5). Since the end points must show great irregularities because of the transformation, only the medium range was included in the graph, which is the most informative part of any ZIPF curve anyway.

Figure 5: Empirical example (index keywords for research projects) of various transformations of the ZIPF curve
(ordinate = number of tokens: abscissa = type tank)
Frequency
Rank

I. = ZIPF curve form the original values
II. = ZIPF curve from selected points of curve I. The numerical values in the curve indicate type ranks in percent of the last rank. (these numerical values permit following the changes in interval spacing). Only these selected points are taken into account in the further transformations and interconnected with lines for greater clarity.
III. = Curve II with token frequencies and type ranks in percent of their maximum values.
IV. = Curve II with type ranks in percent of the highest rank and token frequencies in percent of the token frequency at the 1% rank (parallel shift of curve III, such that the cutting point with the ordinate lies on 100%).
V. = Curve II with the end points being shifted to the values 1 and 100 through compression or stretching. As the lowest values do not represent 1% of the highest values (cf. curve III) the axis graduations are not valid for this curve.

This also causes the relative weight of the values to be changed with respect to the ZIPF curve through the compression or stretching. The intervals between the curve points are changed accordingly. If we are dealing with a compression the high rank values will increase in importance; if with a stretching, the lowest rank values will become dominant in the graphical representation.

3. Application Examples for the Transformed ZIPF Curves

3.1 Application of the Percented Curve

Now in the following we will try to compare three concrete documentation language applications with the aid

Figure 6: Percentage-based ZIPF curves (type IV) of the documentation languages for selected word forms
Frequency in % of the frequency at the 1% rank
Rank order in percent of the total number of word types

As we have seen, in a ZIPF curve it is not the slope of ascent but rather the arching (convexity or concaveness) which supplies information on the equal distribution of the words. A convex curve with respect to the diagonal would indicate more uniform use of words in comparison with the normal language. Our transformations according to Type IV thus all show a relatively uniform word use. Differences between the various documentation languages can be discerned only with difficulty, especially since the selection of only some points makes for distortions. On the basis of the arching in the medium range the documentation languages can be roughly classified as follows:

Word-wider[8]: index keywords (REG),
main categories (KAT)
Word-narrower: title words (TTL)

3.2 Application of the Compressed Curve

Even this very clusterd order by the percented ZIPF curve can be regarded as overly interpreted, so that we attempted a transformation according to Type V, i.e. a

compression of the entire range of the ZIPF curve with the end point being placed at the values 1 and 100. The same word forms were used as with the percentage-based ZIPF curve.

Now these compressed curves reveal much better the differences among the documentation languages as to word concentration. This is achieved by weighting the various sections of the curve differently. Particular weight is attached – independently of the size of the language corpora – to the starting and end values. These curves give rise to the following relative characterization of the documentation languages:

word-wider: main categories (KAT)
medium: index keywords (REG)
word-narrower: title words (TTL)

This order can be explained as follows: The title words of the projects have the greatest word-narrowness, as they are most similar to the normal language[9]. The main categories, representing an extremely artifical language, have the greatest word-width. The index keywords can be regarded as a middle course between conformity to the natural language and artifical abstraction[10] (see Fig. 7).

Figure 7: Compressed ZIPF curves (type V) of the documentation languages for selected word forms

Compressed frequencies — Compressed rank orders

4. On the Validity of the Compressed ZIPF Curve

These results on the basis of the compressed ZIPF curve could be confirmed by further concentration measurements. Thus, a representation of these documentation languages in the MEIER spectrum as well as a calculation of a (simplified[11]) variation coefficient likewise showed title words and main categories to lie at the corresponding extreme points. This relative agreement with the transformed ZIPF curve can be regarded as verification of the permissiblility of the compression or stretching of the curves.

Notes:

1 As to the explanation of the ZIPF relation expecially by MANDELBROT, see: HÖRMANN 1970, p. 89–93; CHERRY 1967, p. 142–150. Incidentally, there are also investigations (GUIRAUD according to HÖRMANN 1970, p. 91) which explain the various frequencies of the words on the basis of word length. For documentation languages this would mean that rare words are long and therefore potentially of composite nature, which might explain their differentiating function in the natural language but make them appear little suitable for documentation languages.

2 ZIPF examined also the connection between ambiguity (multiplicity of meanings) and frequency rank of word roots. He was able to show a linear dependence. These results correspond to the idea that an above-average frequency of word use points to homonymic application. Attempts have been made, however, to prove the existence of logarithmic relationships – such as noted by ZIPF – also in extra-linguistic fields. If such a relationship could be proven for cognitive structures as well, the ZIPF law would reflect the economy of thinking rather than of language (cf. CHERRY 1967, p. 147).

3 A text with 100 word forms (types) having a maximum frequency starting at 100 would (according to the ZIPF Law) correspond to 5050 individual words (tokens).

4 The instruction would read: Find the ideal text sequence for which the frequency of the lowest rank is 100 times the frequency of the highest rank and for which the highest rank is 100 times the lowest rank.

6 A representation of further comparison measurements as well as of further applications is found in OHLY, 1976.

7 Those word forms were selected which lay at a rank of 1%, 2%, 4%, 8%, 16%, 32%, 64% and 100% of the total number of word forms.

8 Be it recalled once more that word-wide is identical in meaning to weak concentration and word-narrow to strong concentration of the words.

9 For this transformed curve the diagonal for the normal language (with the end points at the values 1 and 10.000) no longer applies.

10 In the introduction of further intermediate steps (e.g. title words without trivial words, retrieval keywords, more refined categories) the modified ZIPF curve permitted a meaningful insertion (OHLY 1976).

11 To simplify the calculations we used in the denominator instead of the scattering of all individual values as an approximation the so-called "range" obtained as the difference between the highest and the lowest value (here: of the word rank).

References:

CHERRY, Colin: Kommunikationsforschung – eine neue Wissenschaft (Communication Research, A New Science), 2d edition, Hamburg 1967. (Title of the American original: On Human Communication, New York).

HÖRMANN, Hans: Psychologie der Sprache (Psychology of Language). Improved new impression, Berlin(-West) 1970.

OHLY, H. Peter: Möglichkeiten und Probleme der empirischen Analyse und des Vergleiches von sozialwissenschaftlichen Dokumentationssprachen (Possibilities and problems of the empirical analysis and comparison of documentation languages in the social sciences), Cologne, Seminar für Soziologie 1976 (Diploma thesis).