3-1-2022

Why Menzerath's Law?

Julio Urenda
The University of Texas at El Paso, jcurenda@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:

Technical Report: UTEP-CS-22-32

Why Menzerath's Law?

Julio Urenda and Vladik Kreinovich

Abstract In linguistics, there is a dependence between the length of the sentence and the average length of the word: the longer the sentence, the shorter the words. The corresponding empirical formula is known as the Menzerath's Law. A similar dependence can be observed in many other application areas, e.g., in the analysis of genomes. The fact that the same dependence is observed in many different application domains seems to indicate there should be a general domain-independent explanation for this law. In this paper, we show that indeed, this law can be derived from natural invariance requirements.

1 Formulation of the Problem

Menzerath's law: a brief description. It is known that in linguistics, in general, the longer the sentence, the shorter its words. There is a formula – known as the Menzerath's Law – that describes the dependence between the average length x of the word and the length y of the corresponding sentence: $y = a \cdot x^{-b} \cdot \exp(-s \cdot x)$; see, e.g., [2, 3, 5, 6, 7, 9].

To be more precise, the original formulation of this law described the dependence between the number Y of words in a sentence and the average length of the word: $Y = a \cdot x^{-B} \cdot \exp(-s \cdot x)$. However, taking into account that the average length x of the word is equal to y/Y, we thus conclude that $y = x \cdot Y = a \cdot x^{-(B-1)} \cdot \exp(-s \cdot x)$, i.e., that the relation between y and x has exactly the same form.

Julio Urenda
Deparments of Mathematical Sciences and Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA, e-mail: jcurenda@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA, e-mail: vladik@utep.edu

Menzerath's law is ubiquitous. A similar dependence was found in many other application areas. For example, the same formula describes the dependence between the length of DNA and the lengths of genes forming this DNA; see, e.g., [4, 8].

Challenge. Since this law appears in many different application areas, there must be a generic explanation. The main objective of this paper is to provide such an explanation.

2 Analysis of the Problem

There are different ways to describe length. We can describe the length of a word or a phrase by number of letters in it. We can describe this length by the number of bits or bytes needed to store this word in the computer. Alternatively, we can describe the word in an international phonetic alphabet, this usually makes its description longer. Some of these representations make the length larger, some smaller.

A good example is the possibility to describe the length in bits or in bytes. In this case, since 1 byte is 8 bits, the length of the word in bits is exactly 8 times longer than its length in bytes. In general, we can use different units for measuring length, and, on average, when you use a different unit, length y in the original unit becomes length $x' = c \cdot x$ in the new units, where c is the ratio between the two units.

For the length of the word x, there is an additional possibility. For example, in many languages, the same meaning can be described in two ways" by a prefix or a postfix or by a preposition. For example, when Julio owns a book, we can say it is a book *of Julio*, or we can say that it *Julio's* book. Similarly, we can say that a function is *not linear* or that a function is *non-linear*. There are many cases like this. In all these examples, the first case, we have two words, while in the second case, we have one longer word, a word to which a "tail" of fixed length was added. So, if we perform the transformation from the first representation to the second one, then the average length of the meaningful words will increase by a constant x_0: the average length of such an addition (and the average length of words in general will decrease). In this case, the average length of the word changes from x to $x' = x + x_0$.

The relation between x and y should not depend on how we describe length. As we have mentioned, there are several different ways to describe both the length of the phrase y and the average length of the words x. There seems to be no reasons to conclude that some ways are preferable. It is therefore reasonable to require that the dependence $y = f(x)$ should have the same form, no matter what representation we use: if we change the way we describe length x, the dependence between x and y should remain the same.

Of course, this does not mean that if we change from x to x', we should have the same formula $y = f(x')$. For example, the dependence $d = v \cdot t$ describing how the path depends on velocity v and time t does not change if we change the unit of velocity, e.g., from km/h to miles per hour. However, for the formula to remain valid, we need to also change the unit of distance, from km to miles.

In general, invariance of the relation $y = f(x)$ means that for each re-scaling $x \mapsto x'$ of the input x, there should exist an appropriate re-scaling $y \mapsto y'$ of the output y such that if we have $y = f(x)$ in the original units, then we should have the exact same relation $y' = f(x')$ in the new units.

Let us describe what this invariance requirement implies for the above two types of re-scaling.

Invariance with respect to scaling $x \mapsto c \cdot x$. For this re-scaling, invariance means that for every $c > 0$, there exists a value $C(c)$ (depending on c) for which $y = f(x)$ implies that $y' = f(x')$, where $y' = C(c) \cdot y$ and $x' = c \cdot x$. Substituting the expressions for x' and y' into the formula $y' = f(x')$, we conclude that $C(c) \cdot y = f(c \cdot x)$. Since $y = f(x)$, we conclude that $C(c) \cdot f(x) = f(c \cdot x)$. It is known (see, e.g., [1]) that every measurable function (in particular, every definable function) that satisfies this functional equation has the form $y = A \cdot x^p$ for some real numbers A and p.

Invariance with respect to shift $x \mapsto x + x_0$. For this re-scaling, invariance means that for every x_0, there exists a value $C(x_0)$ (depending on x_0) for which $y = f(x)$ implies that $y' = f(x')$, where $y' = C(x_0) \cdot y$ and $x' = x + x_0$. Substituting the expressions for x' and y' into the formula $y' = f(x')$, we conclude that $C(x_0) \cdot y = f(x + x_0)$. Since $y = f(x)$, we conclude that $C(x_0) \cdot f(x) = f(x + x_0)$. It is known (see, e.g., [1]) that every measurable function (in particular, every definable function) that satisfies this functional equation has the form $y = D \cdot \exp(q \cdot x)$ for some real numbers D and q.

How can we combine these two results? We wanted to find the dependence $y = f(x)$, but instead we found *two* different dependencies $y_1(x) = A \cdot x^p$ and $y_2(x) = D \cdot \exp(q \cdot x)$. We therefore need to combine these two dependencies, i.e., to come up with a combined dependency

$$y(x) = F(y_1(x), y_2(x)).$$

Which combination function $F(y_1, y_2)$ should we choose? Since the quantity y is determined modulo scaling, it is reasonable to select a scale-invariant combination function, i.e., a function for which, for all possible pairs of values $c_1 > 0$ and $c_2 > 0$, there exists a value $C(c_1, c_2)$ for which $y = F(y_1, y_2)$ implies that $y' = F(y'_1, y'_2)$, where $y' = C(c_1, c_2) \cdot y$, $y'_1 = c_1 \cdot y_1$, and $y'_2 = c_2 \cdot y_2$.

Substituting the expressions for y', y'_1, and y'_2 into the formula $y' = F(y'_1, y'_2)$, we conclude that $C(c_1, c_2) \cdot y = F(c_1 \cdot y_1, c_2 \cdot y_2)$. Since $y = F(y_1, y_2)$, we conclude that $C(c_1, c_2) \cdot F(y_1, y_2) = F(c_1 \cdot y_1, c_2 \cdot y_2)$. It is known (see, e.g., [1]) that every measurable function (in particular, every definable function) that satisfies this functional equation has the form $y = C \cdot y_1^{p_1} \cdot y_2^{p_2}$ for some real numbers C, p_1, and p_2.

Substituting $y_1(x) = A \cdot x^p$ and $y_2(x) = D \cdot \exp(q \cdot x)$ into this expression, we get

$$y(x) = C \cdot (A \cdot x^p)^{p_1} \cdot (D \cdot \exp(q \cdot x))^{p_2} =$$

$$C \cdot A^{p_1} \cdot x^{p \cdot p_1} \cdot D^{p_2} \cdot \exp(p_2 \cdot q \cdot x),$$

i.e., in effect, the desired expression $y = a \cdot x^{-b} \cdot \exp(-s \cdot x)$, where $a = C \cdot A^{p_1} \cdot D^{p_2}$, $b = p \cdot p_1$, and $s = p_2 \cdot q$.

Thus, we have indeed explained that the Menzerath's law can indeed be derived from natural invariance requirements.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

1. J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, 2008.
2. G. Altmann, "Prolegomena to Menzerath's law", *Glottometrika*, 1980, Vol. 2, pp. 1–10.
3. G. Altmann and M. H. Schwibbe, *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Olms, Hildesheim/Zürich/New York, 1989.
4. S. Eroglu, "Language-like behavior of protein length distribution in proteomes", *Complexity*, 2014, Vol. 20, No. 2, pp. 12–21.
5. R. Ferrer-I-Cancho and N. Forns, "The self-organization of genomes", *Complexity*, 2009, Vol. 15, No. 5, pp. 34–36.
6. L. Hřebíček. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*, Wissenschaftlicher Verlag, Trier, 1995.
7. R. Köhler, "Zur Interpretation des Menzerathschen Gesetzes", *Glottometrika*, 1984, Vol. 6, pp. 177–183.
8. W. Li, "Menzerath's law at the gene-exon level in the human genome", *Complexity*, 2012, Vol. 17, No. 4, pp. 49–53.
9. J. Milička, "Menzerath's Law: The whole is greater than the sum of its parts", *Journal of Quantitative Linguistics*, 2014, Vol. 21, No. 2, pp. 85–99.