Menzearth's law on Word Duration

Leonardo Araujo[1], Thaïs Cristófaro-Silva2 and Hani Yehia[2]

[1]Universidade Federal de São João Del Rei, [2]Universidade Federal de Minas Gerais

leolca@ufsj.edu.br, thais,hani@ufmg.br

The psychologist and phonetician Paul Menzerath observed an inverse relation in language between the size of a construct and the size of its constituents (Menzerath, 1954). The duration of a speech sound would be shorter when it is inside a word, or syllable, that is more complex, made of many parts (speech sounds).

Since the beginning of the last century, the implications of Menzerath's law for linguistics has been a topic of debate. The length of syllables in French was studied in (Grégoire, 1899), where the difference in duration of the vowel 'a' was analyzed in the French words *patisserie, pâte* and *pâté*. The vowel was consistently shorter in longer words. Other authors also studied this compression tendency (Meyer, 1904; Roudet, 1910), but it was still an uncharted observation, until Menzerath published his work on the morphological structure of German (Menzerath, 1954).

Following the observation made by Menzerath, Gabriel Altmann made a mathematical formulation and reformulated the hypothesis in a linguistic terminology.

"Je größer ein sprachliches Konstrukt, desto kleiner seine Komponenten (Konstituenten)."[1]

(Altmann, 1980)

He proposes that *bigger* and *smaller* might not strictly refer to the size, but rather to the complexity and the number of entities used to form a construct. Such hypothesis seems plausible. Under the information transmission perspective, it is important to stablish an effective communication, using messages that are as short as possible, highly distinctive, and still keep enough redundancy to prevent loss of information in the process. As a construct gets more complex, it will increase its distinctiveness score, therefore it might be possible to reduce the constituent's size in order to create a shorter message, still maintaining a discriminable code. A construct that is made of few entities, should not squeeze them, for this process would compromise the decoding process.

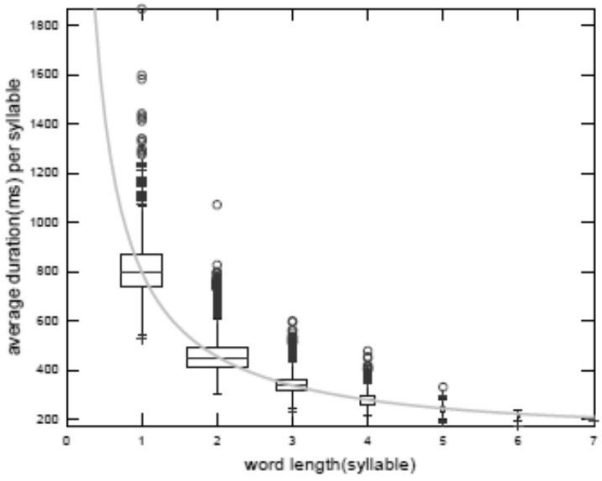
The concept of complexity of a construct is usually regarded as the number of elements used to compose it. Nonetheless, we should also consider the possible interactions between the parts (Simon, 1962) and the inner structure, nature and characteristics of the parts. Under this assumption, higher complexity could mean more than a greater number of components. The relationship described by Menzerath's law is more likely to hold true if we are dealing with direct constituents of a given construct (Altmann and Arens, 1983). For example, in the analysis of Ukrainian texts, the concept of clauses as an intermediate constituent between words and sentences was used, so that Menzerath-Altmann's law could be verified (Buk and Rovenchak, 2007). The analysis here undertaken focuses on the relation between constituents and duration, where duration might implicitly take into account, not only the number of constituents or intermediate units, but also other factors that contribute to the overall complexity in the construct.

We are going to use the general model proposed by Altmann (1980), which also considers a disturbance factor. Denoting the size of the constituent by *y* and the size of the construct by *x*, we have the following relation

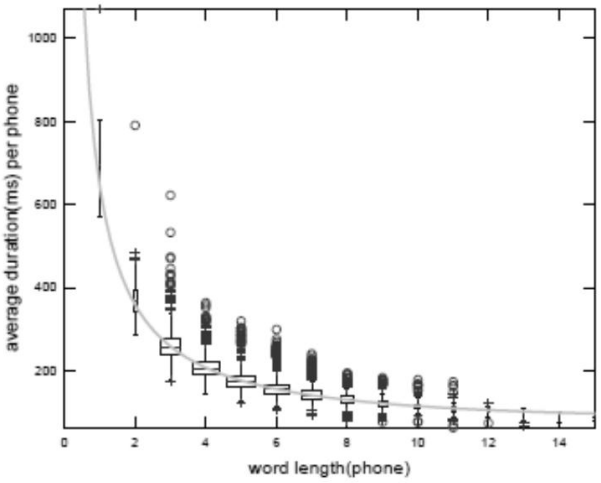
[1] The bigger a language construct, the smaller are its components (constituents).

Figure 1: Relation between the words length (number of phones or number of syllables) and the average duration of its constituent parts (phones or syllables).

(a) Average duration of syllables as the number of syllables in a word increases. The parameters found were: $a = 743$, $b = -0.916$ and $c = 0.072$. The correlation between the model and data was 0.93.



(b) Average duration of phones as the number of phones in a word increases. The parameters found were: $a = 619$, $b = -0.901$ and $c = 0.039$. The correlation between the model and data was 0.91



between constituent and construct: $y'/y = b/x + c$, where b is a proportionality constant and c the disturbance factor. The solution for this Equation is: $y = ax^b e^{cx}$.

As a quantitative law in linguistics, Menzerath's law is based on the statistical analysis of extensive corpora. We propose here to analyze the relation between word length (measured by the number of syllables or by the number of phones) and the average duration of the constituents. In order to do so, we use online dictionaries to acquire a word syllabification, phonetic transcription and utterance duration. The analysis was carried out only in English, using data collected from 4 online dictionaries. We used 10.086 words in total. The parameters of Menzerath-Altmann's law were found using a linear regression over the logarithm of the data. The results are presented in Figure 1. More details on the optimization of the Menzerath-Altmann's parameters might be seen in a publication by Andres et al. (2012).

References

Altmann, G.: 1980, *Glottometrika* 2, 1

Altmann, G. and Arens, H.: 1983, Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.

Festschrift für Peter Hartmann, Chapt. „Verborgene Ordnung" und das Menzerathsche Gesetz, pp 31--39,

Narr, Tübingen

Andres, J., Kubáček, L., Machalová, J., and Tučková, M.: 2012, *Mathematica* 51(1), 5

Buk, S. and Rovenchak, A. A.: 2007, CoRR

Grégoire, A.: 1899, *La Parole* 1, 161

Menzerath, P.: 1954, Die Architektonik des deutschen Wortschatzes, *Phonetische Studien*, F. Dümmler

Meyer, E. A.: 1904, Zur Vokaldauer im Deutschen

Roudet, L.: 1910, *Éléments de phonétique générale*, Welter Simon, H. A.: 1962, in Proceedings of the American Philosophical Society, Vol. 106, pp 467—482.

***